Bridging the Gap:
Generative Machines and Inventive Minds

by

Nikhil Singh

S.M., Massachusetts Institute of Technology (2020)
B.M. Berklee College of Music (2017)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Media Arts and Sciences

at the

Massachusetts Institute of Technology

February 2025

Authored by: Nikhil Singh
Program in Media Arts and Sciences
December 19, 2024

Certified by: Tod Machover
Muriel R. Cooper Professor of Music and Media
Massachusetts Institute of Technology

Accepted by: Tod Machover
Academic Head
Program in Media Arts and Sciences

Bridging the Gap: Generative Machines and Inventive Minds

by

Nikhil Singh

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on December 19, 2024, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

Abstract

Recording technologies, from the phonograph to digital media, have profoundly reshaped the human experience by enabling the capture and reproduction of our sensory world. These technologies allow us to relive experiences through artifacts of remarkable fidelity like photographs and videos, extending the reach of our perception and memory. Of course, we didn't stop at the phonograph; we have built a rich ecosystem of tools for creating, sharing, and exploring recorded media that have had transformative effects on cognition and culture.

Recently, a new and powerful class of tools has emerged: generative models. Unlike recorded media, which reproduces external experiences, generative models can translate our *ideas* directly into artifacts. Here, ideas refer to abstract mental constructs that seed media creation, externally expressed in text prompts, sketches, vocalizations, or other intuitive representations. Just as recorded media augmented our ability to perceive and remember, generative media promises to expand our ability to imagine and invent by offering a more immediate path from cognition to high fidelity creation. Creative work often has us operating at our limits, negotiating boundaries between knowledge and novelty, skill and aspiration, from individual exploration to collective understanding. Generative models, in principle, have the potential to scaffold and accelerate how we transcend these limits by increasing the efficiency with which we discover and pursue new ideas.

In this thesis, I suggest that realizing this potential presents a complex set of challenges that span computation and design. I argue that it requires us to develop a rich stack of *precision tools for human-AI co-creation*, as we have done and continue to do for recorded media. Specifically, I present contributions across two key dimensions of this:

1. **Computational machinery** that supports creative work. I present research

on topics including visually-driven acoustic simulation, interpretable and controllable sound generation from descriptions, and audiovisual content understanding. Focusing on sound as a case study, I describe systems that effectively represent and manipulate creative knowledge across modalities and levels of abstraction.

2. **Interactive systems and studies** that investigate the integration of human and machine effort in content creation. This includes work on conceptual integration in AI-assisted story writing, author-in-the-loop description authoring for accessibility of complex scientific figures, and generative constraints for human ideation. In all, this work seeks insights for designing systems that support human creators through exploration, collaboration, and feedback, rather than aiming to replace or constrain human agency and expertise.

To conclude this thesis, I present a discussion on bridging AI and HCI to gain insights into human creative work and develop stable, generalizable design knowledge for augmenting it. I argue for the design of flexible, parametric tools that enable systematic study of creative behavior under different augmentation designs. Based on this, I propose a conceptual framework to seed the development of a more robust science of human-AI co-creation.

Thesis Supervisor: Tod Machover
Title: Muriel R. Cooper Professor of Music and Media, Massachusetts Institute of Technology

Bridging the Gap:
Generative Machines and Inventive Minds

by

Nikhil Singh

This thesis has been reviewed and approved by the following committee members

Tod Machover

Muriel R. Cooper Professor of Music and Media
Massachusetts Institute of Technology

Elena Glassman

Assistant Professor of Computer Science
Harvard University

Ramesh Raskar

Associate Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Pattie Maes

Germeshausen Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Preface

In 1843, during the immediate aftermath of the industrial revolution, Ada Lovelace was asked to translate Italian engineer Luigi Menabrea's notes on Charles Babbage's *Analytical Engine*, a design that would have been the first computer, were its construction completed. Lovelace went far beyond translation, providing detailed and highly influential notes [314] in which she articulated perhaps the first vision of the generality of computation. Lovelace's concept extended beyond calculation with numbers, envisioning a machine even capable of generating and extending music, through the symbolic manipulation of relationships between pitches, for example. This radical notion is often cited in historical discussions of computational creativity. However, Lovelace's notes are also known for making a very different (though not necessarily contradictory) argument:

> *The Analytical Engine has no pretensions whatever to originate any thing. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths.*
>
> — Lovelace, 1842

This passage, often dubbed *Lovelace's Objection* after Alan Turing [518], is frequently quoted in opposition to claims of machine intelligence. Turing himself reduced it to the question of whether computers are capable of surprise, posing a now-famous test of this. Yet, a closer reading of Lovelace's words reveals a more nuanced perspective. Lovelace continues:

> *. . . it is likely to exert an indirect and reciprocal influence on science itself in another manner. For, in so distributing and combining the truths and the formulae of analysis. . . the nature of many subjects in that science are necessarily thrown into new lights. . . This is a decidedly indirect, and a somewhat speculative, consequence of such an invention: There are in all extensions of human power, or additions to human knowledge. . .*
>
> — Lovelace, 1842

In other words, Lovelace argued that while the Analytical Engine could perhaps not *originate* in the sense of independent thought, it could act as a catalyst for

human discovery. Lovelace's formulation can be interpreted within the now-familiar paradigm of *human augmentation* [138]. Writing with a recent view of the industrial revolution, the role of the human in the face of automation was perhaps a salient subject [541].

Now, we are faced with an automation revolution that more directly affects creative work: large-scale generative AI models [57] that are increasingly capable of performing tasks traditionally associated with human creativity. These models produce media like text [61], images [407], and sounds [58], conditioned on abstract and intuitive outcome specifications like natural language descriptions. As Ludwig and Mullainathan [315] note, on the generation of scientific hypotheses:

> *The creative process is so human and idiosyncratic that it would seem to resist formalism. That may be about to change because of two developments. First, human cognition is no longer the only way to notice patterns in the world. Machine learning algorithms can also notice patterns, including patterns people might not notice themselves. These algorithms can work... with the kinds of inputs that traditionally could only be processed by the mind, like images or text. Second, ... data on human behavior is exploding... The kind of information researchers once relied on for inspiration is now machine readable: what was once solely mental data is increasingly becoming actual data.* — Ludwig and Mullainathan, 2024

The history of computing has not equipped us well to conceptualize this shift. Even futuristic visions like Vannevar Bush's *Memex* concept [69], which foresaw the importance of information retrieval and has inspired many technological developments, assumed that computers were capable only of repetitive, not creative, functions. As such, generative AI models have sparked both excitement and apprehension. Will these tools devalue human creative work, or can they unlock unprecedented opportunities for human creators? The end-to-end design of such models hints at potential automation of human creative work.

Still, the arguments for augmentation are compelling. Creativity occurs in response to rich, large-scale, essentially human contexts [153]. Humans learn, generalize, and innovate in remarkably robust and data-efficient ways [501]. Machines lack the human propensity for causal explanations, and may be brittle in novel circumstances

due to factors like hidden biases [465]. Augmentation may also lead to greater prosperity [64]. In light of these factors, I argue that the most promising approach lies in developing tools and systems that facilitate effective creative augmentation, rather than aiming for complete automation. In focusing on augmentation, we can harness the strengths of both human creators and generative models, enabling new forms of creative work while ensuring that the process remains grounded in human values, intentions, and expertise.

The history of recorded media technologies, on the other hand, offers a compelling analogical lens through which to view the potential of generative AI. Take the phonograph, for instance. Initially conceived as a tool for simple sound capture and playback, it has grown into a rich ecosystem of devices, algorithms, interfaces, and even social platforms that empower us to record, edit, discover, and share representations of our world. The camera, once a cumbersome apparatus for static image capture, has similarly evolved into a complex array of tools and platforms supporting visual storytelling. By enabling precise manipulation and transformation, these tools have even extended recorded media beyond the bounds of the physical world, allowing us to craft artifacts that tell compelling stories, convey complex information, and evoke deep emotions.

I propose that we adopt a related approach for generative AI, developing a comprehensive stack of *precision tools* that span tasks and modalities. I use this term as a guiding principle to refer to computational and interactive systems that can augment and refine human creative processes. It encompasses both tools that directly manipulate creative outputs and those that operate at a lower level, for example facilitating extraction of meaningful patterns and structures from data that can be operationalized in creative work. With this serving as context, I propose and discuss a number of projects that contribute to this effort in both computational and interactive ways. Building on this, I propose a path to building more generalizable design knowledge, aiming to better understand how we can design most effectively for expanding the human creative toolkit leveraging the powerful generative capabilities of modern machines.

Acknowledgements

There's a peculiar asymmetry to the structure of my PhD that only became apparent in retrospect. I began it in a silo, in some ways. In the fall of 2020, my research environment was largely confined to pixels on a screen. My body was, in turn, largely confined to a small and temperamental Back Bay studio. I had met almost none of my future collaborators, and had scarcely begun to conceptualize any of the work that populates this document.

The asymmetry is this: even as I write these words by myself, I'm concluding this PhD anything but alone. My research environment has grown from a text editor to a Zoom call to an office to a shared lab space to a building, and ultimately a network that extends far beyond it. Many of those pixels have become people: I am fortunate to have collaborators, mentors, and supporters to thank for contributions both modest and miraculous. In other words, this thesis is largely a product of this ever-expanding asymmetric trajectory toward community.

First, to my advisor, Tod Machover: Thank you for welcoming me into the lab and trusting me to explore a wide (and sometimes bewildering) array of topics. Your encouragement to ask unusual questions and your patience with my many detours were invaluable.

To my thesis committee—Elena Glassman, Ramesh Raskar, and Pattie Maes—Thank you for sharing your deep expertise and for asking questions that brought greater cohesion and purpose to this thesis. Elena, your kindness, generosity, and belief in me have been a privilege, both within and beyond our collaboration. Ramesh and Pattie, your questions have guided me toward clarity and opened new paths for future explorations. Thanks also to Pawan Sinha for serving on my generals committee; your insightful comments on all matters multisensory integration helped me see familiar machine perception problems through an exceptionally rich, informative new lens.

A special acknowledgment to Manuel Cherep, my co-conspirator in much completed and ongoing work. You are a great collaborator in both the chess and the jazz of research, and an even more amazing friend. I'm grateful for the (many) hours we spend brainstorming, debugging, writing, rebutting, and laughing, and our collabo-

ration has grown into one of the best parts of my academic life. To all other readers: as Yanai and Lercher suggest, "It takes two to think." [563] I encourage you to seek out such partnerships—they are invaluable to both the progress of the work and the joy of the journey.

To Mahdi Kalayeh, and additionally to Chih-Wei Wu and Iroro Orife, for your mentorship during my internship at Netflix. Thank you for helping me learn how to scale machine learning models in practical contexts, and for the many discussions on sound, multimodality, and computation over the years. Special thanks to Mahdi for also serving on my generals committee. To Jonathan Bragg and Lucy Lu Wang, my mentors at Ai2. Your advice on methodology and impact in human-centered research continues to shape how I think as a researcher. I admire the depth and conceptual rigor with which you approach open-ended problems and hope to live up to these values in all my future work.

Mentorship comes in many less formal flavors. Thanks to Rébecca Kleinberger: you have been an amazing mentor, friend, and role model. Thank you for teaching me how to navigate the complexities of academic life while staying grounded. To Matt Groh, your intellectual generosity and emotional intelligence set a very high standard for academic citizenship. You are a veritable institution, of the best kind. Iddo Drori, thank you for introducing me to a variety of machine learning research areas early on, which has helped me develop a richer mental model of the intellectual landscape we all share. Ishwarya Ananthabhotla, your expertise and encouragement have been instrumental to my growth in audio and otherwise. I admire your clarity, and am grateful for your patient, wise counsel.

I want to say thanks to the many other (100+!) folks I've had a chance to collaborate and coauthor with. In all, they include PIs, students, postdocs, and alumni from over 10 groups at the Media Lab, 4 departments across MIT, several other institutions, levels of seniority ranging from first-year undergraduates to emeritus faculty, and a wide range of disciplinary perspectives, from CS to Music to Math to Anthropology and other areas. The perspective I have gained from these interactions and experiences is enviable, and I would readily wish it on anyone else with interdisciplinary ambitions.

To the *Opera of the Future* group—my home base for this journey—thank you for your creativity, humor, and friendship. Kimy, your kindness has been a lifeline. Your

strength, thoughtfulness, and willingness to strategically “acquire” burritos for others are further testament to a one-of-a-kind character that I can only hope is contagious. The Jessicas, of which I was briefly one; Jessie, for your excitement, warmth, and generally great vibes. I love how you think about people, and can’t wait for all the ways you will share that. Jess, for your humor, musicality, and entropy. I never know what you’re going to say, but I know I’m going to say “yes, and…” Ana, for your constant commitment to audio, humor, crosswords, and most significantly to selflessly supporting us all. Manaswi, who has brought much positivity to our talks and travels. Alexandra, for the kindness, support, and many extraordinary facts I have learned from you. Clémence, for all the support while putting up with my ridiculous sense of humor. Sam Chin, the honorary member without whom Opera doesn’t seem complete. We’ve had many an adventure, and extraordinarily wide-ranging discussions about everything from spatial perception to labor economics. Thanks for being such an interesting intellectual companion, and even more for being such a great friend.

Thanks also to Opera folks past who I’ve had the privilege of overlapping with. Max, whose energy, glee, and warmth are unmatched. Karsten, whose extraordinary facility in the physical world is matched only by his humility. Ben, thank you for always assuming I’m more capable than I am, and for sharing your encyclopedic knowledge on all things sound and production with generosity. Charles, for improving me as an engineer almost by osmosis, and being a partner in many early explorations. David, Hane, and Nicole, thank you for being my first friends at the lab and for your continued friendship. Aarón and Hannah for lots of fun conversations about everything from synthesizers to screaming traffic cones. Akito, for many connections with the music tech world I wouldn’t have otherwise. Priscilla, for being at times guide, at times bestie, at times voice of reason. Alaa, for being a helpful mentor, a fantastic collaborator, and a great support system.

I’m also grateful to various other faculty. At the lab, Joe, Deb, Zach, Roz, and Mitch have all given me helpful advice. Taking courses with Jacob Andreas, Pulkit Agrawal, and Ted Adelson has expanded and informed my perspective in very useful ways, despite not concretely doing research in NLP, RL, or haptics. Mentors past: Dr. B, without whom I might have wound up on a very different, and no doubt much less interesting, path. Marti, Neil, Gabriele, and other faculty at Berklee for opening my mind to so much about music and creativity. Thanks are also due to the MAS team for helping me navigate the somewhat labyrinthine configuration of this

academic program: Sarra, Mahy, Meghan; thank you. As it turns out, I'm also quite the investment; funding a PhD is no joke. Thanks to LEGO for generously supporting my first two years through a Papert fellowship.

Friends and colleagues around the lab have been a big part of this journey. There are those who I started alongside: Patrick, Will, Mike, Nic, Praneeth, Joanne, Noah, Alex, Abhi, Erik, Manuj, Maggie, Belén, Pat. Those who were around before me and helped me feel welcome to begin with: Guillermo, Brian, Neil, Zivvy, Spencer, and others. Of course, those I met along the way: Kush, the Samanthas, Rob, Cathy, Hang, Aruna, Lancelot, Shayne, Jad, Kayla, Suyash, Cedric, Leticia, Chelsi, Tony. Folks beyond the lab: Ash, Joanna, Nikos, Darius, Anat, Sizi, Matt, Katy, Alec, Jeff, Shruti, Indus crew, and so many others. UROPs and other mentees I've learned much from working with; in particular Luke and Quinn, whose work contributes to this thesis, Ninon and Manvi, with whom I've had many instructive conversations.

My family: my mom, my dad, Suresh; thank you for all your belief in me (even when not entirely warranted). Tara, thank you for always being there, and for being a north star of kindness and compassion. Clint, your warmth and patience are a wonderful addition to the family. Heather and Rick; thank you for your encouragement, advice, and kindness through the years, and for engaging deeply with my work but always valuing my person more. Arden, Patrick, Paul, Emily; thank you for truly welcoming me into your family. To all the tinies; thank you for reminding me about the things that matter most.

Rose. Your love has been both compass and catalyst through all of this. You have a way of rewriting what seems possible, in work and in life. I deeply admire the way you make excellence look like exploration, and precision look like play. My normally robust cynicism dissolves in the presence of your care, kindness, and singular sense of humor. There's so much more I wish I knew how to express, but this paragraph will have to serve as an acute reminder of the poverty of natural language. Thank you for nurturing my curiosity, for making the best parts of this journey ecstatic and the worst parts bearable, and for helping me understand the reason I love research: the unrelenting opportunity to learn.

If putting this thesis together illuminated the social asymmetry in my PhD, writing this section has in turn brought out a subtle (if ironic) symmetry. I began this journey wanting to understand how to amplify human creativity, and I conclude it having been myself amplified by the creativity of all these remarkable humans.

Contents

List of Figures

List of Tables

Published Material

Significant portions of this thesis have been previously published in peer-reviewed venues and revised in their appearance here (including minor editorial edits).

* indicates co-first authors (equal contribution).

- Chapter 2 [469]: Published as *Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis. IEEE/CVF International Conference on Computer Vision (ICCV), 2021.*

- Chapter 3 [474]: Published as *Nikhil Singh, Chih-Wei Wu, Iroro Orife, and Mahdi Kalayeh. Looking Similar, Sounding Different: Leveraging Counterfactual Cross-Modal Pairs for Audiovisual Representation Learning. IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), 2024.*

- Chapter 4 [89]: Published as *Manuel Cherep* and Nikhil Singh*. Synthax: A Fast Modular Synthesizer in JAX. Audio Engineering Society (AES), 2023.*

- Chapter 5 [90]: Published as *Manuel Cherep*, Nikhil Singh*, and Jessica Shand. Creative text-to-audio generation via synthesizer programming. International Conference on Machine Learning (ICML), 2024.*

- Chapter 6 [90]: Released as a preprint, *Manuel Cherep* and Nikhil Singh*. Contrastive Learning from Synthetic Audio Doppelgängers. arXiv preprint arXiv:2406.05923, 2024.*

- Chapter 7 [349]: Published as *Luke Mo*, Manuel Cherep*, Nikhil Singh*, Quinn Langford, and Pattie Maes. Articulatory synthesis of speech and diverse vocal sounds via optimization. NeurIPS Audio Imagination Workshop, 2024.*

- Chapter 8 [470]: Published as *Nikhil Singh*, Guillermo Bernal*, Daria Savchenko*, and Elena L Glassman. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. ACM Transactions on Computer-Human Interaction (TOCHI), 2022.*

- Chapter 9 [473]: Published as *Nikhil Singh, Lucy Lu Wang, and Jonathan Bragg. FigurA11y: AI Assistance for Writing Scientific Alt Text. ACM Conference on Intelligent User Interfaces (IUI), 2024.*

- Chapter 10 [472]: Published as *Nikhil Singh, Manaswi Mishra, and Tod Machover. AI for Musical Discovery. An MIT Exploration of Generative AI, 2024.*

1

Introduction

When we marvel at human creativity, we often fixate on the final artifact: a painting that captures a special moment, or an algorithm that elegantly and efficiently solves a complex problem. Yet the deepest understanding of how these works come to be often emerges from examining the creative *process*. The primacy of process over product in understanding—and supporting—creative work is not only a theoretical stance. Rather, it follows from a few key observations and has some important consequences. First, creative artifacts represent only the successful endpoints of often lengthy, complex exploratory trajectories, obscuring important decision points and constraints that shaped their development and could have shaped it differently. Focusing on the product hides the many ways in which computational augmentation can impact the creative output. Second, the same output can in principle arise from radically different processes, suggesting the importance of supporting diverse pathways. Third, as we increasingly augment human creativity with computational systems, we must discover how the different technical and design components embedded in them impact creative work.

As such, our interventions must often examine and decompose this process into modular parts, to better understand how systems we build might offer a helping hand. In the process, we find that some parts are better supported through designing new computational methods. Others may require prototyping new interactions. Still others may rest on empirical knowledge obtained from studying process, that can then be translated into abstract design principles.

In this thesis, I present a concerted effort across several such parts towards this broader goal. To motivate these parts, consider the following scenarios:

- An architectural acoustician wants to estimate what a space might sound like, but only has access to visual renderings in the early stages of a design. How

can they reason about the sonic properties of spaces that exist only as visual abstractions? Though a rich 3D model could drive a physics-based simulation of how sound might propagate in the space, the connection between a visual abstraction and its acoustic correlate eludes precise formal description. This is also true of spaces in the real world: looking at an image of the international space station's interior might offer a sense of what it might sound like, but capturing an imprint with which to simulate it necessitates a costly round-trip.

- A sound designer wants to craft a sound that evokes a sense of birdsong, and turns to a synthesizer. Though they have an intuition of what the target might sound like, and can express its high-level semantics (birdsong), they may twist tens of knobs for hours in frustration trying to bridge the gap between these intuitions and technical execution. Even with deep expertise, the process is encumbered by a laborious search for a semantic manifold in a very high-dimensional parameter space.

- Two writers are working (separately) on fictional stories, and both find themselves stuck. One would benefit from a new idea that makes a dramatic twist, potentially requiring a page one rewrite of core aspects of their story. The other just needs a nudge as to what comes next. Their individual and situational needs diverge, but these needs are sometimes hard to know (let alone express clearly). Should an assistant (human or machine) ignore this important context?

These scenarios are some of the kind considered in this thesis's contributions. In particular, they correspond to the three overarching parts that separate the chapters to come, with an additional fourth part focusing on a synthesis of individual findings and proposal of a new methodological framework for studying human-AI creative interaction.

Part I investigates how we can model relationships between visual and auditory elements of scenes. In particular, we look at situations where our goal is to capture *possible* relationships, rather than only observed ones.

- Chapter 2 presents Image2Reverb, a system that learns to generate acoustic impulse responses directly from images of spaces, enabling rapid acoustic simulation from only visual observations (or renderings) without costly measurement or complex 3D modeling.

- Chapter 3 examines how visually similar scenes can have substantially different acoustic characteristics, particularly through language variation, and demonstrates how leveraging this through the proxy of second-language dubbed audio tracks (which serve as counterfactual-like augmentations) can improve audiovisual representation learning.

Part II explores sound synthesis at the intersection of humans and machines. In all, this work models the synthesizer, and in particular its parameter space, as a synergistic playground for both humans and machines.

- Chapter 4 introduces SynthAX, a fast modular synthesizer implementation enabling rapid sound generation.
- Chapter 5 demonstrates how this can be used to create abstract sound interpretations from text descriptions through automated synthesizer programming, bridging semantic and technical gaps.
- Chapter 6 shows how synthetic audio pairs generated through controlled parameter variation can improve learned audio representations.
- Chapter 7 explores a very different kind of synthesizer: one modeled on the human vocal tract, and considers how we might use this to reconstruct recordings of the human voice by optimization. In principle, this allows the absorption of the human voice into the synthesis framework that we explore in the previous chapters.

Part III examines how AI systems can augment creative work while maintaining human agency.

- Chapter 8 studies how writers integrate multimodal AI suggestions into their creative process, identifying patterns in how they make "integrative leaps" from suggestions to story development.
- Chapter 9 investigates how AI can help authors write better alt text descriptions for scientific figures, improving accessibility while preserving author knowledge and discretion.
- Chapter 10 zooms out to gain perspective on what the role of AI in music should be—arguing that it should be to support musical *discovery*—and presents a vision of what this might practically look like.

Finally, **Part IV** aims towards a synthesis of these investigations.

- Chapter 11 considers what themes recur across the diverse projects presented in this thesis.
- Chapter 12 considers the knowledge integration problem: how do we build generalizable design knowledge from such investigations? I argue that we must move past the current status quo of fragmented theories and empirical results, and propose a framework for systematically studying human-AI creative interaction through "meta-prototypes": flexible systems that enable controlled experimentation. I lay out a plan for how these might be conceptualized, built, and studied.

This organization proceeds from computational foundations in cross-modal modeling and synthesis to investigations of human-AI creative interaction, and culminates in the design of a broader conceptual framework for understanding and designing AI creative augmentation systems that embrace the primacy of process over product.

Part I

Modeling the Audiovisual Scene

2

Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis

The relationship between visual and acoustic properties of spaces has long captivated us. For example, architects and acousticians often have sophisticated intuitions about how spaces might sound based on visual inspection alone. Yet, these intuitions are trapped in their minds, difficult to externalize into real simulations that could be valuable for various design, engineering, and artistic tasks. While physics-based acoustic simulation is possible given detailed 3D models of an environment, the barrier between visual observation and acoustic experience remains high, especially in early design stages or for spaces that are inaccessible or no longer exist.

This chapter examines whether neural networks can bridge this gap by learning to simulate acoustic properties directly from images. Rather than pursuing perfect acoustic recreation, we demonstrate that useful approximations are possible: approximations that support rapid prototyping and creative exploration while leveraging coarse human intuition in the form of visual specification. In a way, the goal is to translate an implicit human understanding of visual-acoustic relationships into an explicit computational model that can support the creation of novel artifacts. The system accepts diverse visual inputs ranging from photographs to renderings to paintings, enabling acoustic simulation for both real and imagined spaces.

Abstract

Measuring the acoustic characteristics of a space is often done by capturing its impulse response (IR), a representation of how a full-range stimulus sound excites it. This work generates an IR from a single image, which can then be applied to other

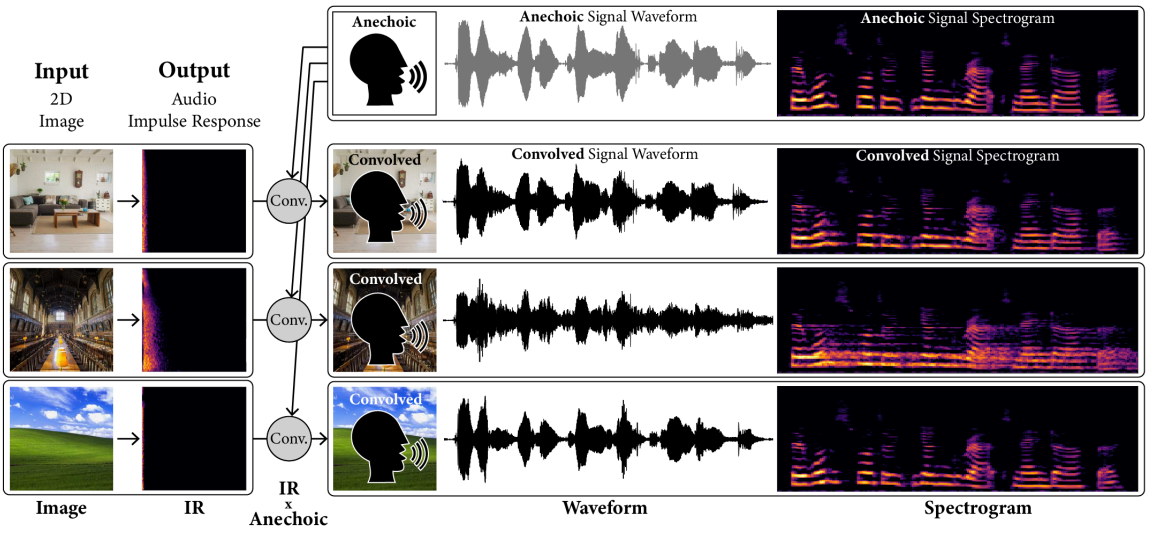

Figure 2-1: Generating audio impulse responses from images. Left: given an image of an acoustic environment as input, our model generates the corresponding audio impulse response as output. Right: generated impulse responses are convolved with an anechoic (free from echo) audio recording making that recording sound as if it were in the corresponding space. Waveforms and spectrograms are shown of the source anechoic signal and the same signal after convolution with the corresponding synthesized IR. All spectrograms are presented on a mel scale. Image2Reverb is the first system demonstrating end-to-end synthesis of realistic IRs from single images.

signals using convolution, simulating the reverberant characteristics of the space shown in the image. Recording these IRs is both time-intensive and expensive, and often infeasible for inaccessible locations. We use an end-to-end neural network architecture to generate plausible audio impulse responses from single images of acoustic environments. We evaluate our method both by comparisons to ground truth data and by human expert evaluation. We demonstrate our approach by generating plausible impulse responses from diverse settings and formats including well known places, musical halls, rooms in paintings, images from animations and computer games, synthetic environments generated from text, panoramic images, and video conference backgrounds.

2.1 Introduction

An effective and widely used method of simulating acoustic spaces relies on audio impulse responses (IRs) and convolution [427, 523]. Audio IRs are recorded mea-

surements of how an environment responds to an acoustic stimulus. IRs can be measured by recording a space during a burst of white noise like a clap, a balloon pop, or a sinusoid swept across the range of human hearing [413]. Accurately capturing these room impulse responses requires time, specialized equipment, knowledge, and planning. Directly recording these measurements may be entirely infeasible in continuously inhabited or inaccessible spaces of interest. End-to-end IR estimation has far ranging applications relevant to fields including music production, speech processing, and generating immersive extended reality environments. Our Image2Reverb system directly synthesizes IRs from images of acoustic environments. This approach removes the barriers to entry, namely cost and time, opening the door for a broad range of applications.

In this work we model IR generation as a cross-modal paired-example domain adaptation problem and apply a conditional GAN [192, 197, 346] to synthesize plausible audio impulse responses conditioned on images of spaces. Next, we will describe related work that informs our approach.

2.2 Related Work

Artificial reverberation. Historically, recording studios built reverberant chambers with speakers and microphones to apply reverb to pre-recorded audio directly within a physical space [418]. Reverberation circuits, first proposed in the 1960s, use a network of filters and delay lines to mimic a reverberant space [452]. Later, digital algorithmic approaches applied numerical methods to simulate similar effects. Conversely, convolution reverb relies on audio recordings of a space's response to a broadband stimulus, typically a noise burst or sine sweep. This results in a digital replica of a space's reverberant characteristics, which can then be applied to any audio signal [11].

Convolutional neural networks have been used for estimating late-reverberation statistics from images [264, 265], though not to model the full audio impulse response from an image. This work is based on the finding that experienced acoustic engineers readily estimate a space's IR or reverberant characteristics from an image [263]. Room geometry has also been estimated from 360-degree images of four specific rooms [415], and used to create virtual acoustic environments which are compared with ground-truth recordings, though again IRs are not directly synthesized from the

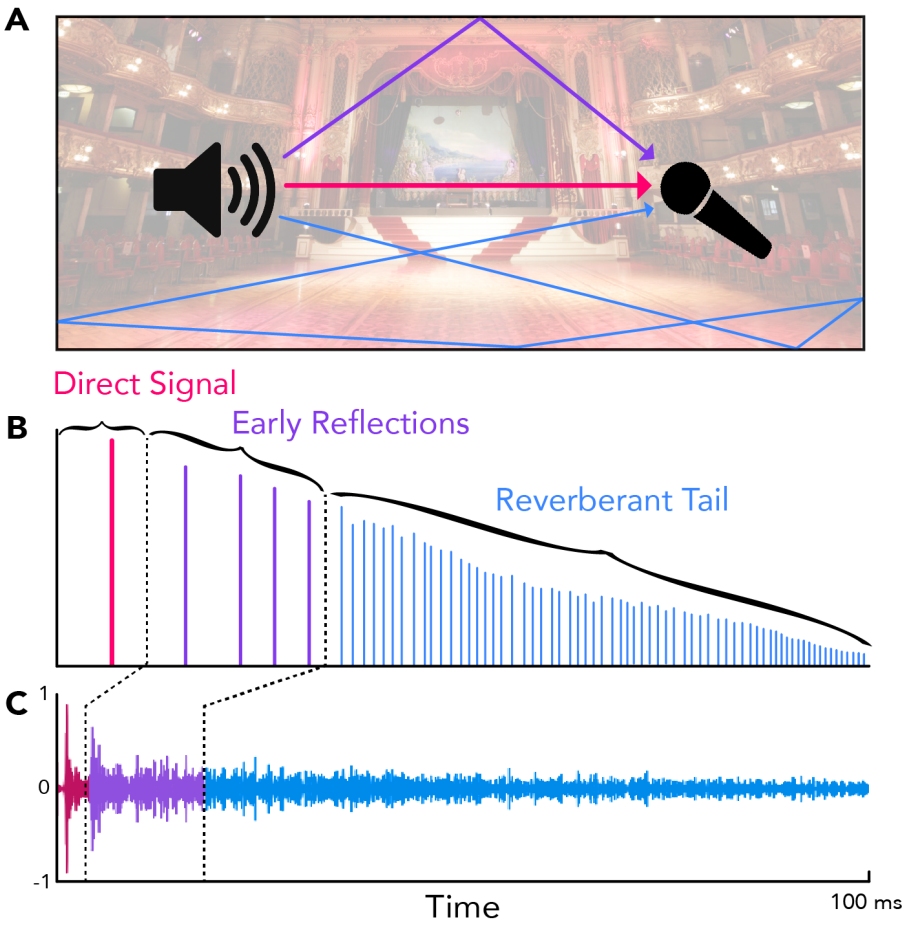

Figure 2-2: Impulse response overview. **(A)** Sound waves propagate across multiple paths as they interact with and reflect off their environment. These paths include the direct path from source to listener, early reflections including 1st and higher order reflections (after reflecting off 1 or more surfaces) and a more diffuse tail as they trail off and become more densely packed in time. These reflections make up the impulse response of the environment illustrated **(B)** schematically and **(C)** as a waveform.

images. A related line of work synthesizes spatial audio based on visual information [168, 256, 294]. Prior work exists on synthesis of IRs using RNNs [442], autoencoders [483], and GANs: IR-GAN [410] uses parameters from real world IRs to generate new synthetic IRs; whereas our work synthesizes an audio impulse response directly from an image.

Generative models for audio. Recent work has shown that GANs are amenable to audio generation and can result in more globally coherent outputs [127]. GANSynth [135] generates an audio sequence in parallel via a progressive GAN architecture allowing faster than real-time synthesis and higher efficiency than the autoregressive WaveNet [525] architecture. Unlike WaveNet which uses a time-distributed latent coding, GANSynth synthesizes an entire audio segment from a single latent vector. Given our need for global structure, we create a fixed-length representation of our input and adapt our generator model from this approach.

Measured IRs have been approximated with shaped noise [63, 288]. While room IRs exhibit statistical regularities [511] that can be modeled stochastically, the domain of this modeling is time and frequency limited [26], and may not reflect all characteristics of real-world recorded IRs. Simulating reverb with ray tracing is possible but prohibitively expensive for typical applications [448]. By directly approximating measured audio IRs at the spectrogram level, our outputs are immediately applicable to tasks such as convolution reverb, which applies the reverberant characteristics of the IR to another audio signal.

Cross-modal translation. Between visual and auditory domains, conditional GANs have been used for translating between images and audio samples of people playing instruments [84]. Our work builds on this by applying state-of-the-art architectural approaches for scene analysis and high quality audio synthesis, tuned for our purposes.

2.3 Methods

Here we describe the dataset, model, and algorithm.

2.3.1 Dataset

Data aggregation. We curated a dataset of 265 different spaces totalling 1169 images and 738 IRs. From these, we produced a total of 11234 paired examples with

a train-validation-test split of 9743-154-1957. These are assembled from sources including the OpenAIR dataset [354], other libraries available online, and web scraping. Many examples amount to weak supervision, due to the low availability of data: for example, we may have a “kitchen" impulse response without an image of the kitchen in which it was recorded. In this case, we augmented with plausible kitchen scenes, judged by the researchers, gathered via web scraping and manual filtering. Although this dataset contains high variability in several reverberant parameters, e.g. early reflections and source-microphone distance, it allows us to learn characteristics of late-field reverberation.

Data preprocessing. Images needed to be filtered manually to remove duplicates, mismatches such as external pictures of an indoor space, examples with significant occlusive “clutter" or excessive foreground activity, and intrusive watermarks. We then normalized, center-cropped at the max width or height possible, and downsampled to 224x224 pixels. We converted the audio IR files to monaural signals; in the case of Ambisonic B-Format sources we extracted the W (omnidirectional) channel, and for stereo sources we computed the arithmetic mean of channels. In some cases, 360-degree images were available and in these instances we extract rectilinear projections, bringing them in line with the standard 2D images in our dataset.

Audio representation. Our audio representation is a log magnitude spectrogram. We first resampled the audio files to 22.050kHz and truncate them to 5.94s in duration. This is sufficient to capture general structure and estimate reverberant characteristics for most examples. We then apply a short-time Fourier transform with window size ($M = 1024$) and hop size ($R = 256$), before trimming the Nyquist bin, resulting in square 512x512 spectrograms. Finally, we take $\log(|X|)$ where $|X|$ represents the magnitude spectrogram; audio IRs typically contain uncorrelated phase, which does not offer structure we can replicate based on the magnitude.

2.3.2 Model

Components. Our model employs a conditional GAN with an image encoder that takes images as input and produces spectrograms. This overall design, with an encoder, generator, and conditional discriminator, is similar to that which Mentzer et al. [343] applied to obtain state-of-the-art results on image compression, among many other applications. The generator and discriminator are deep convolutional networks

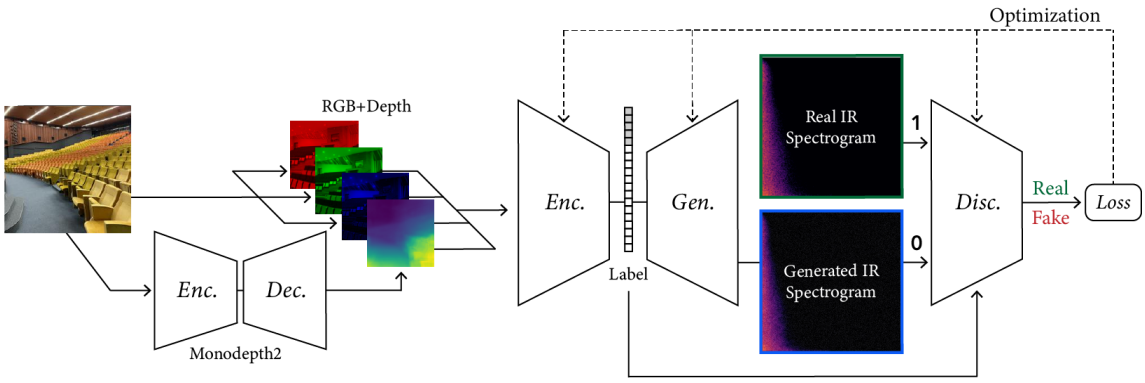

Figure 2-3: System architecture. Our system consists of autoencoder and GAN networks. Left: An input image is converted into 4 channels: red, green, blue and depth. The depth map is estimated by Monodepth2, a pre-trained encoder-decoder network. Right: Our model employs a conditional GAN. An image feature encoder is given the RGB and depth images and produces part of the Generator's latent vector which is then concatenated with noise. The Discriminator applies the image latent vector label at an intermediate stage via concatenation to make a conditional real/fake prediction, calculating loss and optimizing the Encoder, Generator, and Discriminator.

based on the GANSynth [135] model (non-progressive variant), with modifications to suit our dataset, dimensions, and training procedure.

The encoder module combines image feature extraction with depth estimation to produce latent vectors from two-dimensional images of scenes. For depth estimation, we use the pretrained Monodepth2 network [189], a monocular depth-estimation encoder-decoder network which produces a one-channel depth map corresponding to our input image. The main feature extractor is a ResNet50 [207] pretrained on Places365 [579] which takes a four-channel representation of our scene including the depth channel (4x224x224). We add randomly initialized weights to accommodate the additional input channel for the depth map. Since we are fine-tuning the entire network, albeit at a low learning rate, we expect it will learn the relevant features during optimization. Our architecture's components are shown in Figure 2-3.

Objectives. We use the least-squares GAN formulation (LSGAN) [327]. For the discriminator:

Algorithm 1 Forward and backward passes through the Image2Reverb model. Notation is explained in Table 2.1.

Input:
Monodepth2: $x \sim X$; Encoder: $\tilde{x} \sim \tilde{X}$; Generator: $z = E(\tilde{x}) \oplus u$; Discriminator: $(G(z), E(\tilde{x}))$ OR $(y, E(\tilde{x}))$
Parameters: (weight variables)
Output:
Monodepth2: $\mathbf{x_d}$; Encoder: $E(\tilde{x})$; Generator: $G(z)$; Discriminator: $D(G(z), E(\tilde{x}))$ OR $D(y, E(\tilde{x}))$
for number of epochs **do**
  Sample B training images
  Get depth $x_d = M(x)$
  Append depth features to RGB channels: $y \oplus y_d$
  Encode image to feature-vector: $E(\tilde{x})$
  Append noise: $z = E(\tilde{x}) \oplus u$
  Generate spectrogram: $G(z)$
  Forward pass through discriminator with either fake or real spectrogram: $D(G(z) \mid E(\tilde{x}))$ OR $D(y \mid E(\tilde{x}))$
  Backward pass: update parameters for discriminator (W_D), generator (W_G), and encoder (W_E)
end for

Notation	Definition
x	input image
x_d	estimated depth map
$\oplus$	concatenation operator
$\tilde{x}$	image with depth map ($x \oplus x_d$)
y	Real spectrogram
E, G, D	Encoder, Generator, Discriminator
M	Monodepth2 Encoder-Decoder
W_*	weights for a model
u	Noise, $u \sim \mathcal{N}(0, 1)$
z	Latent vector, encoder output and noise ($E(\tilde{x}) \oplus u$)

Table 2.1: Notation and definitions for variables indicated in different parts of this chapter.

$$\min_D V(D) = \mathbb{E}_{\mathbf{y}\sim p_{data}(\mathbf{y})}[(1 - D(y \mid E(\tilde{x}))^2] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[(D(G(z) \mid E(\tilde{x}))^2] \tag{2.1}$$

For the generator, we add an $\ell 1$ reconstruction term, scaled by a hyperparameter ($\lambda_a = 100$ in our case). This is a common approach in image and audio settings. In all:

$$\min_G V(G) = \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}\left[\,(1 - D(G(z) \mid E(\tilde{x}))^2 + \lambda_a \left\|G(z) - y\right\|_1 \right. \tag{2.2}$$

Training. We train our model on 8 NVIDIA 1080 Ti GPUs. Three Adam optimizers for each of the Generator, Discriminator, and Encoder were used to optimize the networks' parameter weights. Hyperparameters are noted in Table 2.2. We make our model and code publicly available [1].

Parameter	Value
η_G	4e-4
η_D	2e-4
η_E	1e-5
β	(0.0, 0.99)
ϵ	1e-8

Table 2.2: Hyperparameters for the Generator, Discriminator, and Encoder initial learning rates, the optimizer beta (β), and epsilon (ϵ) for the Adam optimizers we use (one each for D, G, E)

2.4 Results

Using Image2Reverb we are able to generate perceptually plausible impulse responses for a diverse set of environments. In this section, we provide input-output examples to demonstrate the capabilities and applications of our model and also review results of a multi-stage evaluation integrating domain-specific quantitative metrics and

[1] Model and code: https://github.com/nikhilsinghmus/image2reverb

expert ratings. Our goal is to examine output quality and conditional consistency, generally considered important for conditional GANs [123] and most relevant for our application.

2.4.1 Examples

We present several collections consisting of diverse examples, with inputs curated to illustrate a range of settings of interest including famous spaces, musical environments, and entirely virtual spaces. All examples are made available as audiovisual collections[2] and were generated with a model trained in around 12 hours, with 200 epochs on a virtual machine. Figure 2-4 shows examples from our test set that were used in our expert evaluation (4 of 8, one from each category of: Small, Medium, Large, and Outdoor). We convolve a spoken word anechoic signal with the generated IRs for the reader to hear. Figure 2-5 takes images of diverse scenes (art, animation, historical/recognizable places) as inputs. Figure 2-6 demonstrates how sections of 360-degree equirectangular images are cropped, projected, and passed through our model to generate IRs of spaces for immersive VR environments.

We strongly encourage the reader to explore these examples on the accompanying web page. We include examples of musical performance spaces, artistic depictions (drawings, paintings), 3D animation scenes, synthetic images from OpenAI's DALL•E, as well as real-world settings that present challenges (e.g. illusions painted on walls, reflections, etc.). These are largely created with real-world environments for which we may not have ground truth IRs, demonstrating how familiar and unusual scenes can be transformed in this way.

2.4.2 Ablation Study

To understand the contribution of key architectural components and decisions, we perform a study to characterize how removing each affects test set T_{60} estimation after 50 training epochs. The components are the depth maps and the pretrained Places365 weights for the ResNet50 encoder. Table 2.3 reports descriptive statistics of the T_{60} error distributions over the test set for each of these model variants.

Our model reflects better mean error (closer to 0%) and less dispersion (a lower standard deviation) than the other variants. The former is well within the just noticeable difference (JND) bounds for T_{60}, often estimated as being around 25-30%

[2]Audiovisual samples: https://web.media.mit.edu/~nsingh1/image2reverb/

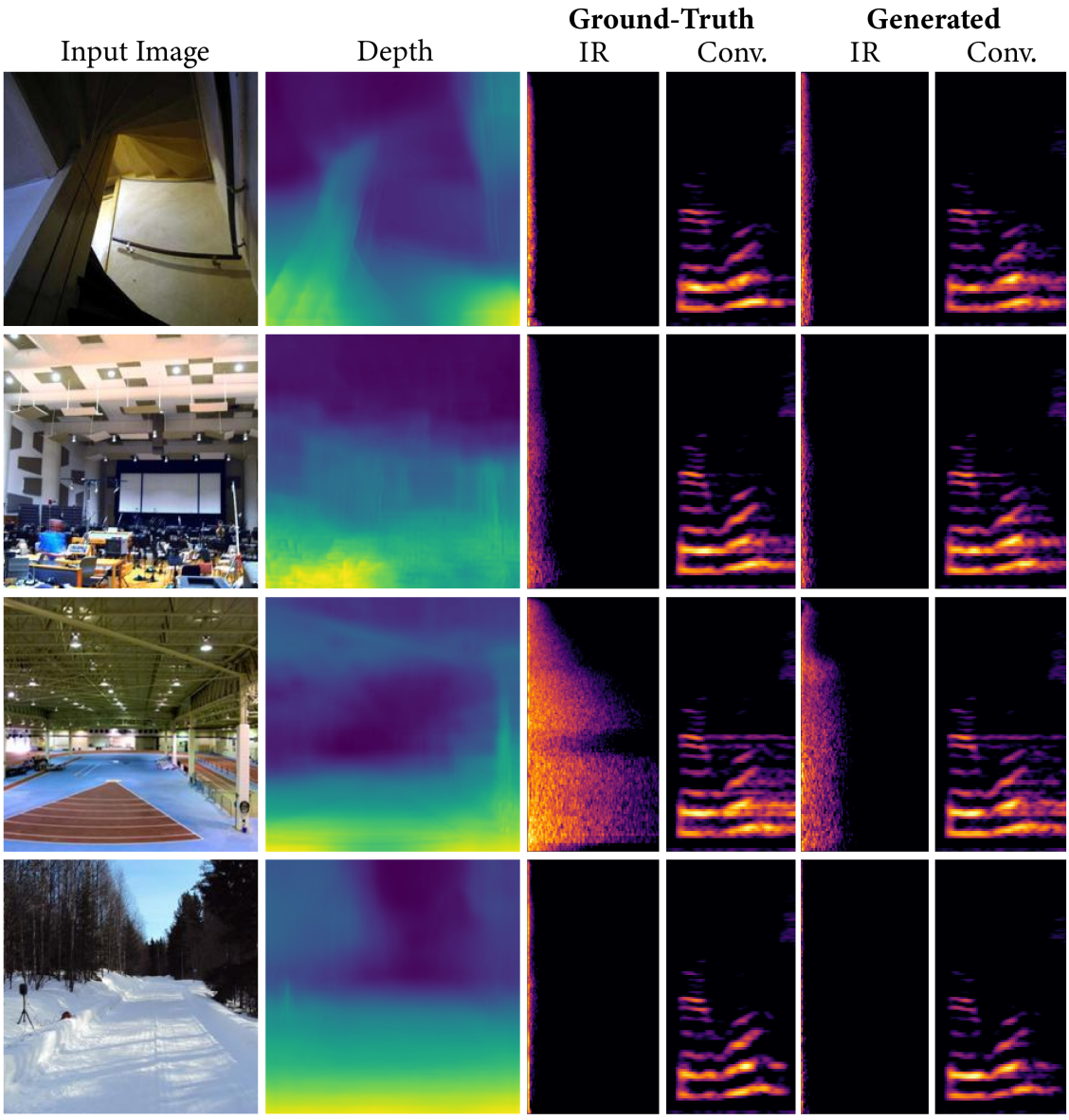

Figure 2-4: Ground-truth measured IRs vs generated IRs. Columns show input images, depth maps, measured IRs with corresponding convolved speech, and generated IRs with corresponding convolved speech. Larger indoor spaces here tend to exhibit greater T_{60} times with longer measured impulse responses. The outdoor scene has a very short measured IR and corresponding generated IR. Input images are all examples that were used in the expert survey and were drawn from the test set.

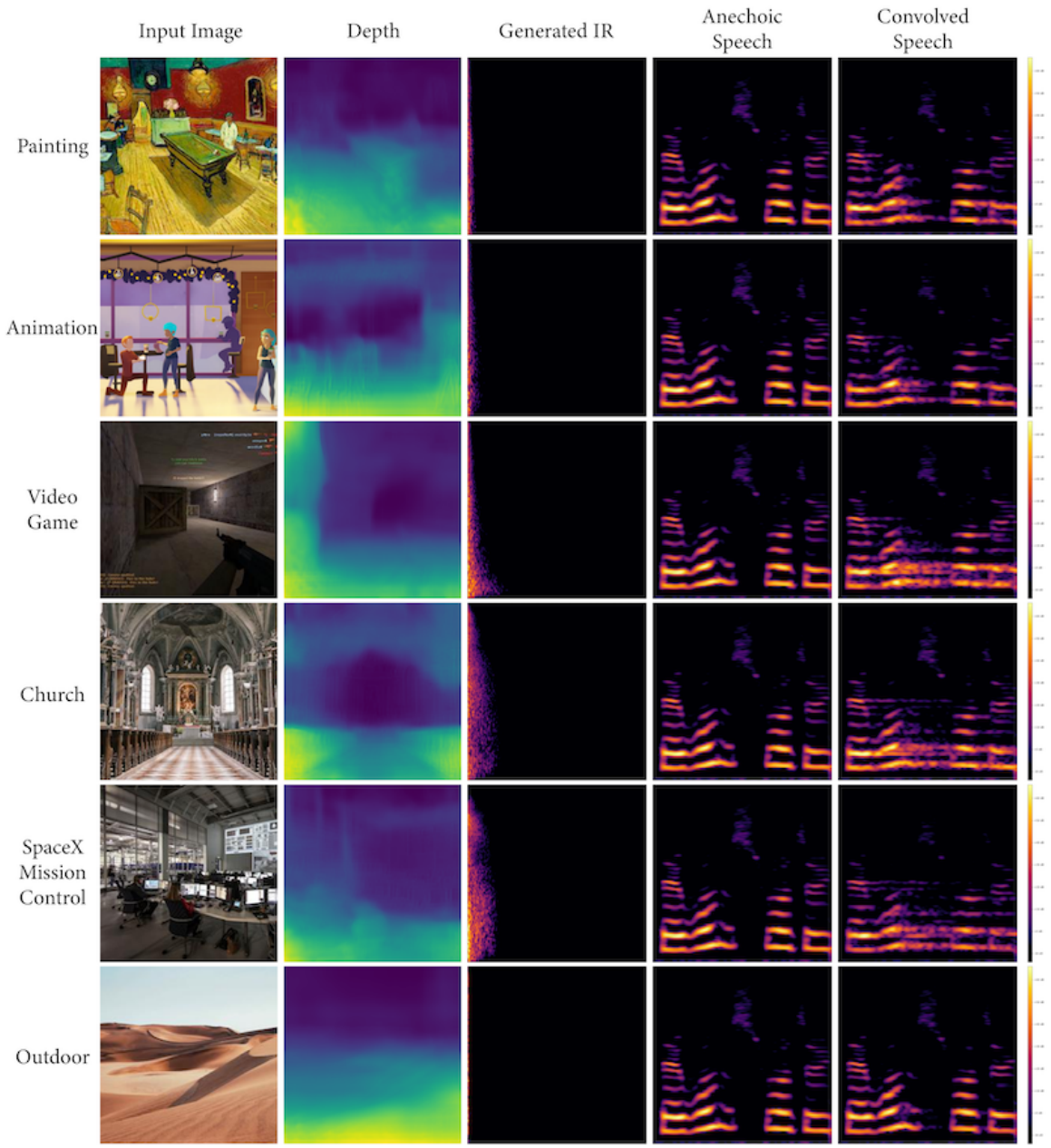

Figure 2-5: Generated IR examples. Columns show input images, depth maps, generated IRs, and a dry anechoic speech signal before and after the generated IR was applied via convolution. Input images come from a variety of spaces which illustrate possible applications of our model. Some images are synthetic, including: an oil painting, a 3D animation still, and a video game screenshot. Others come from real-world scenes like a church (where music is often heard), a famous yet inaccessible space (SpaceX), and an outdoor desert scene. Larger indoor spaces tend to exhibit longer impulse responses as seen here.

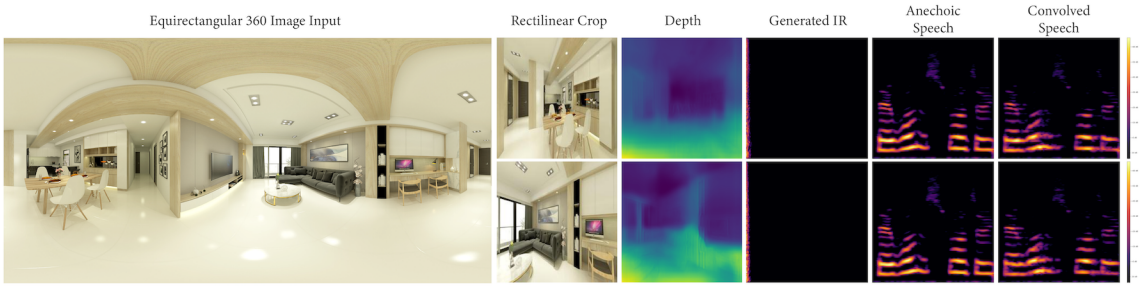

Figure 2-6: VR. Impulse responses generated from an equirectangular 360-degree image by sampling points on a sphere, cropping and applying a rectilinear projection to the resulting image, and feeding them into our model. This demonstrates how our model directly generates realistic impulse responses of panoramic virtual reality compatible images. Future work may allow generation of impulse responses using an entire 360-degree image, though at present there is a lack of paired data available for training.

		Main	-Depth	-P365	NN
T_{60} Err (%)	μ	**-6.03**	-9.17	43.15	149
	σ	**78.8**	83.1	144.3	491.02

Table 2.3: T_{60} estimation error (%) statistics from each model version. “Main" is our architecture as described earlier, “-Depth" omits depth maps and “-P365" does not use the pretrained Places365 weights for the ResNet50 encoder. “NN" indicates a nearest-neighbor approach with Places365-ResNet50 embeddings for images. For mean and median, values closer to 0 reflect better performance. For the standard deviation, lower values reflect better performance.

for a musical signal [342]. Additionally, this is an upper bound on authenticity: a more rigorous goal than perceptual plausibility [386]. The lower standard deviation indicates generally more consistent performance from this model across different examples, even in the presence of some that cause relatively large estimation errors due to incorrect interpretation of relevant qualities in the image, or inaccurate/noisy synthesis or estimation.

2.4.3 Expert Evaluation

Following the finding that experienced acoustic engineers readily estimate a space’s reverberant characteristics from an image [263], we designed an experiment to evaluate our results. We note that this experiment is designed to estimate comparative perceptual plausibility, rather than (physical) authenticity (e.g. by side-by-side

comparison to assess whether any difference can be heard). These goals have been differentiated in prior work [386]. We selected two arbitrary examples from each of the four scene categories and recruited a panel of 31 experts, defined as those with significant audio experience, to participate in a within-subjects study. For each of these examples, we convolved an arbitrary anechoic signal with the output IR, as well as the ground truth IR. These 16 samples were presented in randomized order and participants were instructed to rate each on a scale from 1 to 5 based on 1) reverberation quality, and 2) realism or "match" between their expected reverb based on the image and the presented signal with reverb applied. Participants answered one reverb-related screening question to demonstrate eligibility, and two attention check questions at the end of the survey. The four scene categories are: Large, Medium, Outdoor, and Small. These demonstrate diversity in visual-reverb relationships. The dependent variables are quality and match ratings, and the independent variables are IR source (real vs. fake) and scene category (the four options listed previously).

A two-way repeated-measures ANOVA revealed a statistically significant interaction between IR source and scene category for both quality ratings, $F(3, 90) = 7.04$, $p \leq .001$, and match ratings, $F(3, 90) = 3.73$, $p = .02$ (reported p-values are adjusted with the Greenhouse-Geisser correction [195]). This indicates that statistically significant differences between ratings for real and fake IR reverbs depend on the scene category. Per-participant ratings and rating changes, overall and by scene, are shown in Figure 2-7.

Subsequent tests for simple main effects with paired two one-sided tests indicate that real vs. fake ratings are statistically equivalent ($p < .05$) for large and small quality ratings, and large, medium, and small match ratings. These tests are carried out with an ϵ of 1 (testing for whether the means of the two populations differ by at least 1). Results are shown in Table 2.4. Notably, outdoor scenes appear to contribute to the rating differences between real and fake IRs. We conjecture this is due to outdoor scenes being too different a regime from the vast majority of our data, which are indoor, to model effectively. Additionally, medium-sized scenes appear to contribute to differences in quality.

2.4.4 Model Behavior and Interpretation

Effect of varying depth. We compare the full estimated depth map with constant depth maps filled with either 0 or 0.5 (chosen based on the approximate lower and

Rating	Scene	DoF	p
Quality	Large	56	$< .001$
Quality	Medium	56	$.28$
Quality	Outdoor	56	$.62$
Quality	Small	56	$< .001$
Match	Large	56	$< .001$
Match	Medium	56	$.006$
Match	Outdoor	56	$.29$
Match	Small	56	$< .05$

Table 2.4: Simple main effect tests for equivalence between real and generated IRs across different categories of scenes. We use paired two one-sided tests with bounds (ϵ) of 1 and Bonferroni-adjusted p-values. These results suggest that real vs. fake ratings are statistically equivalent within one rating unit (the resolution of the rating scale) for large and small quality ratings, and large, medium, and small match ratings. Notably, outdoor scenes contribute to the difference between real and fake IRs and medium-sized scenes contribute to differences in quality.

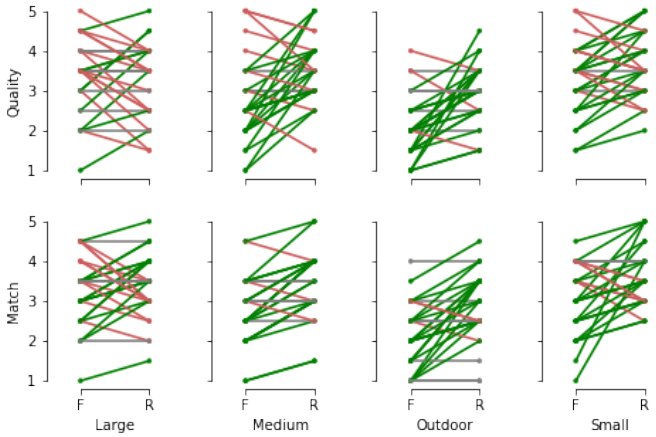

Figure 2-7: Expert evaluation results. Paired plots showing per-participant quality and match differences in rating for each scene category. Green lines indicate higher rating for real IRs, red lines for generated IRs, and grey lines equivalent ratings.

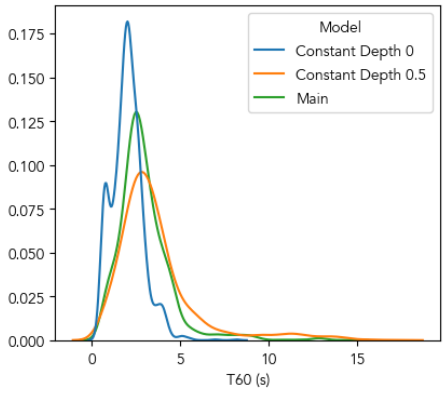

Figure 2-8: Effect of Depth on T_{60}. Distributions of estimated T_{60} values for the model with estimated depth maps, plus constant depth maps set to either 0 (low) or 0.5 (high). Manipulating the depth value allows us to "suggest" smaller or larger scenes, i.e. bias the output of the model. Table 2.5 shows corresponding descriptive statistics. These results indicate a level of "steerability" for the model's behavior in human-in-the-loop settings.

		Main	Depth 0	Depth 0.5
T_{60} (s)	μ	2.07	2.01	3.62
	σ	1.54	0.87	2.36
	$Mdn.$	2.69	2.00	3.07

Table 2.5: Descriptive statistics for the model with estimated depth maps, as well as constant depth maps set to either 0 or 0.5. The full depth map's results are between that of the 0 and 0.5 depth maps. Figure 2-8 visualizes the corresponding distributions.

upper bounds of our data). We survey the distributions of generated IRs' T_{60} values over our test set, the results of which are shown in Figure 2-8. Table 2.5 reports descriptive statistics for these distributions, showing that the main model's output IRs' decay times are biased lower by the 0-depth input and higher by the 0.5-depth input respectively. These may indicate some potential for steering the model in interactive settings. We do note, however, that behavior with constant depth values greater than 0.5 is less predictable. This may be due to the presence of outdoor scenes, for which the scene's depth may not be correlated with IR duration.

Effect of transfer learning. To understand which visual features are important to our encoder, we use Gradient-weighted Class Activation Mapping (Grad-CAM) [455]. Grad-CAM is a popularly applied strategy for visually interpreting convolutional

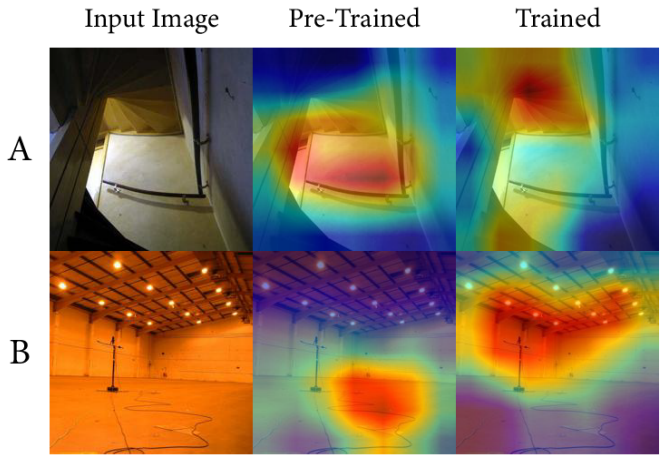

Figure 2-9: Grad-CAMs for images passed through the pre-trained Places365 ResNet50 encoder vs. our fine-tuned encoder, showing movement towards significant reflective areas for **(A)** a small, and **(B)** a large environment. The fine-tuned model's activations highlight larger reflective surfaces: depth of staircase for **(A)** vs. railing that may be more optimal for scene identification, and wall-to-ceiling corner plus surrounding areas for (B).

neural networks by localizing important regions contributing to a given target feature (or class in a classification setting). We produce such maps for our test images with both the ResNet50 pre-trained on Places365 dataset, as well as the final encoder model. All resulting pairs exhibit noticeable differences; we check for this with the structural similarity index (SSIM) metric [539], which is below 0.98 for all examples.

We qualitatively survey these and identify two broad change regimes, which are illustrated with particular examples. First, we observe that the greatest-valued feature is often associated with activations of visual regions corresponding to large reflective surfaces. Examples are shown in Figure 2-9. Often, these are walls, ceilings, windows, and other surfaces in reflective environments. Second, we find that textured areas are highlighted in less reflective environments. Examples of these are shown in Figure 2-10. These may correspond to sparser reflections and diffusion.

Limitations and future work. Many images of spaces may offer inaccurate portrayals of the relevant properties (size, shape, materials, etc.), or may be misleading (examples in supplementary material), leading to erroneous estimations. Our dataset also contains much variation in other relevant parameters (e.g. DRR and EDT) in a

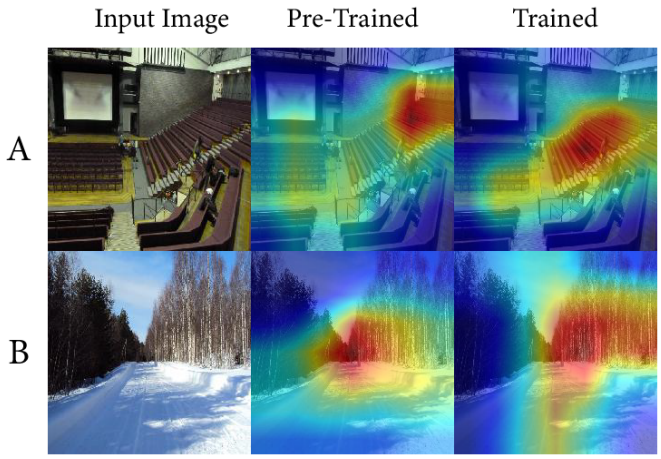

Figure 2-10: Grad-CAMs for images passed through both the pre-trained Places365 ResNet50 encoder and our fine-tuned encoder, showing movement towards more textured areas for **(A)** an indoor, and **(B)** an outdoor environment. The former seems to contain significant absorption and the latter has few reflective surfaces. In both cases, textured areas are highlighted. These may be associated with absorption, diffusion, and more sparse reflections depending on the scene.

way we cannot semantically connect to paired images, given the sources of our data. New audio IR datasets collected with strongly corresponding photos may allow us to effectively model these characteristics precisely.

2.5 Conclusion

We introduced Image2Reverb, a system that is able to directly synthesize audio impulse responses from single images. These are directly applied in downstream convolution reverb settings to simulate depicted environments, with applications to XR, music production, television and film post-production, video games, videoconferencing, and other media. Our quantitative and human-expert evaluation shows significant strengths, and we discuss the method's limitations. We demonstrate that end-to-end image-based synthesis of plausible audio impulse responses is feasible, given such diverse applications. We hope our results provide a helpful benchmark for the community and future work and inspire creative applications.

3

Looking Similar, Sounding Different: Leveraging Counterfactual Cross-Modal Pairs for Audiovisual Representation Learning

---

Just as humans intuitively grasp visual-acoustic relationships in physical spaces, we understand how visual scenes relate to their soundtracks in audiovisual media. This capacity extends beyond simple correspondences: we readily accept that the same visual scene can host many different sonic environments (e.g. a busy vs. empty restaurant), and the same sonic environment can host many different conversations, for example. Yet, most computational approaches to audiovisual learning focus on discovering consistent relationships between sight and sound, potentially missing the rich space of valid variations that we so naturally comprehend. This chapter investigates whether embracing such variations, rather than minimizing them, could actually improve machine audiovisual understanding. We leverage a unique data source that naturally encodes this variation: dubbed films, where visually identical scenes are paired with systematically varying audio tracks. These "counterfactual" audiovisual pairs provide a controlled way to study how scenes might sound different while looking similar.

This work challenges conventional wisdom about audiovisual learning in two ways. First, it suggests that carefully constructed training distributions that deviate from naturalistic assumptions can sometimes yield more generalizable models. Second, it shows that a source of apparent inconsistency in human-created media (i.e. different dubs for the same visual track) can actually encode valuable information about

the space of possible audiovisual relationships. Like Chapter 2, this suggests that computational models can benefit from embracing the flexibility and ambiguity inherent in human sensory understanding. By building models that better represent possible audiovisual relationships, rather than just the most probable ones, we can also enable systems that better support human creative work: work that often involves exploring unlikely but meaningful combinations of sight and sound, guided by our robust perceptual capabilities.

Abstract

Audiovisual representation learning typically relies on the correspondence between sight and sound. However, there are often multiple audio tracks that can correspond with a visual scene. Consider, for example, different conversations on the same crowded street. The effect of such counterfactual pairs on audiovisual representation learning has not been previously explored. To investigate this, we use dubbed versions of movies and television shows to augment cross-modal contrastive learning. Our approach learns to represent alternate audio tracks, differing only in speech, similarly to the same video. Our results, from a comprehensive set of experiments investigating different training strategies, show this general approach improves performance on a range of downstream auditory and audiovisual tasks, without majorly affecting linguistic task performance overall. These findings highlight the importance of considering speech variation when learning scene-level audiovisual correspondences and suggest that dubbed audio can be a useful augmentation technique for training audiovisual models toward more robust performance on diverse downstream tasks.

3.1 Introduction

Can two videos look similar while sounding different? Consider the two scenes on the left in Fig. 3-1. These come from different sources, but share elements like a violinist in the background, other tables further away, and a couple's voices in an upscale restaurant environment; but what are they saying? This can vary considerably between the two scenes, even without changing other aspects. General-purpose self-supervised audiovisual representations are often focused on non-speech applications, evidenced by both existing training datasets and common downstream evaluation tasks. In audio alone, there is a myriad of applications beyond semantic speech

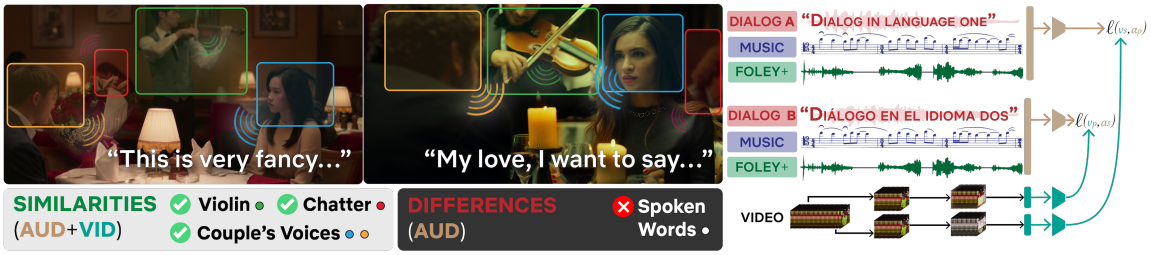

Figure 3-1: **(Left)** Audiovisual scenes can be perceptually similar even as the words spoken in them differ, which may be a challenge for self-supervised audiovisual representation learning. **(Right)** We propose to leverage movie dubs during training and show that it improves the quality of learned representations on a wide range of tasks.

processing, leading to recent benchmarks which evaluate generalization across and trade-offs between types of tasks [517, 536]. How then can we focus on learning robust representations from audiovisual content with speech mixed into it? Importantly, there are many non-semantic, or *paralinguistic*, speech processing tasks of interest, as speech is much more than audible text. These too require discovering other similarities beyond words.

Imagine a movie discussion scene, as in Fig. 3-2. Many audiovisual elements are present: background chatter, glasses clinking, music, footsteps, and characters' voices, but *a priori* this scene could contain many different dialogs without changing the fundamental scene attributes, beyond local features such as lip movements, and this indicates an explicitly counterfactual structure. Note that there are also other counterfactual cross-modal structures which relate to different problems, such as multiple videos of dancing to the same music. Differences in spoken words are one specific case of this which we explore.

In this work, we hypothesize that this *looking similar, while sounding different* problem, as it can occur in real-world audiovisual data distributions, may inhibit the performance of self-supervised audiovisual representation learners. Established approaches, such as cross-modal contrastive learning, where models learn to discriminate true audiovisual pairs from false ones, could be affected; linguistically different but otherwise similar audio-video pairs could act as confounders in this case. However, counterfactual versions of exactly the same scene with only different dialog are generally not available, even if the distribution of real-world audiovisual scenes exhibits this overall trend.

We propose to leverage a data source which naturally resembles this counterfactual-like structure as a proxy: *dubs*. Dubs are alternate versions of movie audio tracks where the speech is replaced with a second-language adaptation, and the rest of the sounds are generally unchanged. Recent works have shown how training on movie scenes can yield strong performance [85, 242], since they contain diverse audiovisual mixtures, compared with popular audiovisual datasets which are curated to focus on specific objects or actions. Although this distribution may help in learning representations focused on overall scene attributes rather than the dialog's semantics, which is our goal, contrastive training on aligned audio and video from movies does not explicitly account for scenes that look similar and sound different due to linguistic variation. We improve upon this strategy by leveraging multilingual dubbed versions of movies[1]. Specifically, we create a dataset of movies and television shows, each with up to seven audio tracks: English (**EN**), Spanish (**ES**), French (**FR**), Japanese (**JA**), German (**DE**), Italian (**IT**) and Korean (**KO**). We plug our training strategy into a well-established self-supervised contrastive learning formulation, *i.e.* SimCLR [86], and we show that this can improve performance in both multimodal and unimodal setups. Overall, this work contributes:

- An approach to improving self-supervised audiovisual representation learning using *dubs*, secondary audio language versions of movies.
- Extensive experiments showing that this approach not only improves performance on a range of auditory and audiovisual tasks but also yields new state-of-the-art on multiple benchmarks.
- Additional experiments to investigate potential trade-offs. These show that we can get an improvement without majorly affecting the performance on language identification, and semantic speech tasks.
- An example pipeline for producing counterfactual pairs in various languages; we apply the workflow to the LVU [551] dataset and demonstrate the possibility of creating alternate audio tracks that potentially empower the research community to further investigate the impact of spoken words in audiovisual representation learning.

[1]The pretraining data also includes episodes of television shows. To avoid clutter, we refer to all long-form content as movies unless it is necessary to specify.

Figure 3-2: Consider the pictured scene. Which of these dialog examples is more likely? Both are plausible within the scene, yet their phonetic-acoustic characteristics would create differences in the soundtrack.

3.2 Related Work

Self-supervised and Multimodal Learning Self-supervised learning relies on pretext tasks with engineered supervision based on data structure, rather than human labels, to learn useful representations [24, 126, 185, 202, 208, 348, 368, 573]. We focus on *contrastive learning*, which has shown strong performance by maximizing mutual information between views of the same instance [24, 86, 169, 208, 505, 505, 514]. These can then be adapted to novel tasks by fine-tuning, or by appending simple (often linear) models, both with smaller-scale task-specific supervision requirements. *Cross-modal* contrastive learning specifically leverages multimodal data like image and text [402], or, as in our case, video and audio [7, 14, 22, 150, 219, 266, 318, 356, 357, 376, 378, 384, 535, 560, 565].

Audiovisual Learning Audiovisual learning harnesses cross-modal correspondences for tasks like action [248, 266] and speaker [98, 355] recognition, source separation [78, 424, 522], media synthesis [166, 198, 377, 488], audio spatialization [172, 351, 561], acoustic simulation [81, 325, 469], and more. Much work takes a contrastive approach, recognizing that audio and video can be treated as two complementary sensory views of a single underlying phenomenon, and focuses on learning *coordinated* [29] representations. Prior work has found that cross-modal training can lead to better results than within-modal training [352], so we use this cross-modal setup as the basis for our framework. In this work, we rely on multilingual audio dubs and videos from long-form content, e.g. movies and television shows. Movies contain rich audiovisual correspondences mimicking real-world experiences,

and are more diverse and novel than user-generated videos while being abundant and scalable [85, 225, 500].

General-purpose Audio Representation Learning and Evaluation Sound is heterogeneous, with speech, music, and environmental sounds having very different characteristics. Even within speech, for example, tasks like speech recognition [93, 299] and speech emotion recognition [463] differ dramatically. This has motivated developing general-purpose audio representations [363, 439] and benchmarks like HARES [536] and HEAR [517]. We focus our audio evaluation on HEAR [517] since it provides a consistent API. The central hypothesis is that if dub-augmented training in the cross-modal setting improves the generality of the representations, performance on various tasks should increase while avoiding a significant trade-off on language-related tasks.

Multilingual Audio Multilingual speech processing has enabled progress in areas like speech recognition [70] through pretraining on diverse data [104, 184]. Recently, speech-to-speech translation has been possible as well [287]. Speech translation in audiovisual media is often referred to as *dubbing*. This is a type of audiovisual translation [80] in which speech content from a media artifact (*e.g.* a movie) is re-recorded in another language. Dubs predominate over subtitles in many cultures [79]. This provides naturalistic multilingual data at scale, and offers a specific case for our hypothesis about audio-visual consistency: a dub's soundtrack differs from the original only in spoken language. We seek to leverage dubs' parallel primary and secondary audio, differing only in speech, to learn more robust audiovisual representations. We also produce a synthetic pipeline for creating counterfactual pairs, to demonstrate the concept of counterfactual cross-modal pairs, while enabling future exploration and validation from the research community.

3.3 Pretraining Dataset

Our dataset consists of ∼20K movies and ∼33K television episodes, which constitutes ∼59K video-hours in total. We have paid extra attention to the diversity of titles used in our pretraining dataset in order to minimize the potential implicit biases in our learned representations, and limited ourselves to only a small part of the catalog to investigate this question. Fig. 3-3 provides details on the distribution

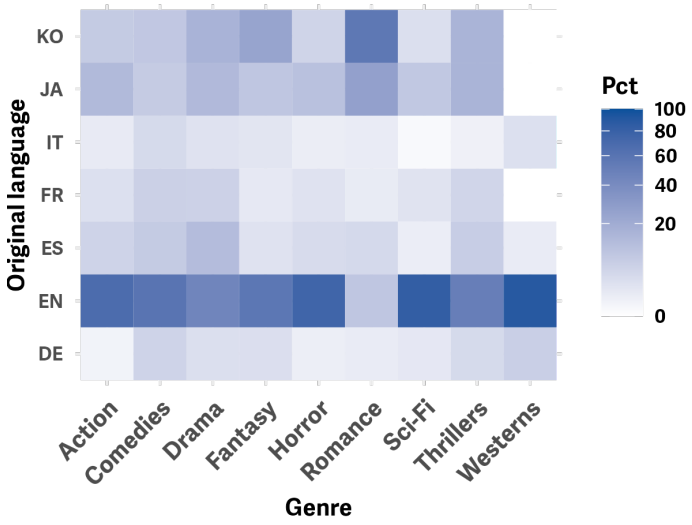

Figure 3-3: Movies and television episodes included in our pretraining dataset are chosen from a diverse set of original languages and genres. Our goal is to minimize potential content and story biases that could potentially impact our self-supervised models. Note that beyond curating the dataset, we do not use this metadata for representation learning. We normalize per column for visualization.

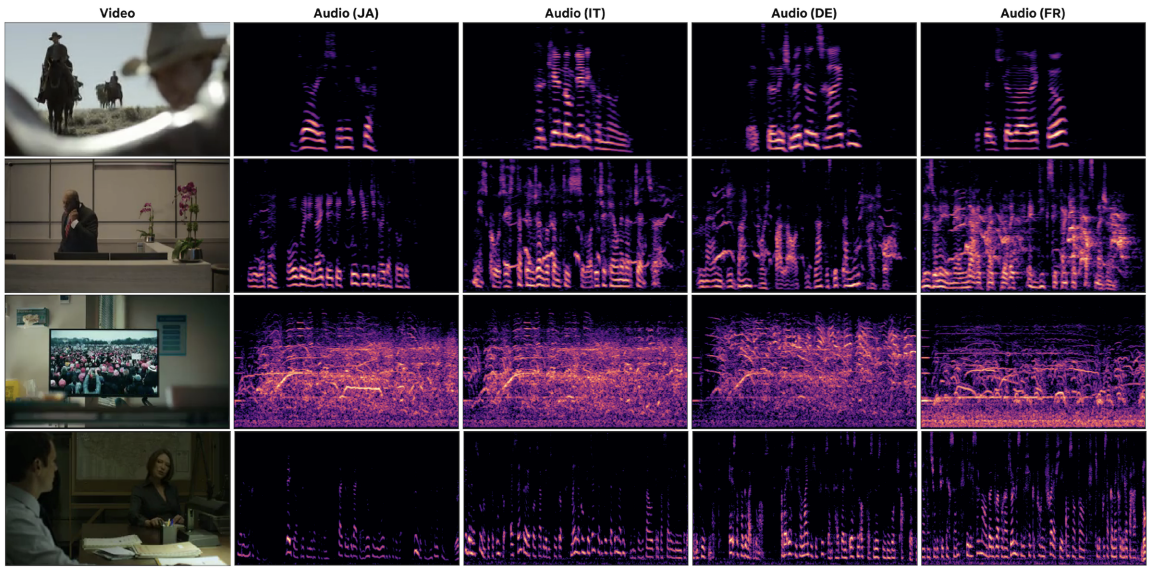

Figure 3-4: Example clips from our pretraining dataset, showing video stills and mel spectrograms for each of the audio tracks.

of genre, and original language of the titles included in our dataset[2]. Each title contains a video track, as well as up to seven audio tracks: English (**EN**), Spanish (**ES**), French (**FR**), Japanese (**JA**), German (**DE**), Italian (**IT**) and Korean (**KO**). Most titles have only a single audio track, which is almost always their original language while about a quarter of the dataset is multilingual where on average 2.8 audio tracks are available for each title. Such a dataset allows us to explore the impact of spoken words in audio for self-supervised audiovisual representation learning. Having multiple dub options enables us to investigate trade-offs between secondary languages, and whether "multilingual" models might further strengthen downstream performance.

We recognize that this kind of data has the potential to significantly benefit research. We are actively investigating the necessary legal steps to potentially release a variant of it for non-commercial use. Fig. 3-4 illustrates a few samples from our dataset but readers are encouraged to check out our project page for more examples[3].

3.4 Methodology

3.4.1 Approach

Our pretraining dataset is denoted by $\mathcal{X} = \{\mathcal{X}_n | n \in [1 \cdots N]\}$, where $\mathcal{X}_n = \{x_{n,m} | m \in [1 \cdots M_n]\}$ contains M_n non-overlapping snippets which are temporally segmented from the duration of the n^{th} title in the dataset. $\mathcal{Q}$ is a function class which we use to create quadruplet training instances $(v_p, a_p, v_s, a_s) \sim \mathcal{Q}(x_{n,m})$[4] where v_p and v_s are obtained through spatio-temporal augmentation of video modality in $x_{n,m}$. Similarly are a_p and a_s for the audio modality, yet, unlike video, we do have the opportunity to further add dub-augmentation to audio instances. When more than one language is available this would ensure that a_p and a_s are similar except in their spoken language.

Randomly sampling negatives, the traditional approach in metric and contrastive learning, has been observed to be suboptimal [318, 446]. A number of recent works develop methods for so-called *hard negative mining*, where the goal is to populate the negative set with challenging examples [370, 426]. In our case, the data is hierarchical; snippets are naturally nested within source long-form titles, and those

[2]Further details are given in the appendix.
[3]nikhilsinghmus.github.io/lssd
[4]subscripts stand for primary and secondary

from the same title share several common attributes including characters, places, objects, voices, and aesthetics. Hence, following prior work [242], to create a minibatch $\mathcal{B} = \{x_i | i \in [1 \cdots B]\}$, we first uniformly sample a title, $n \sim \mathbb{U}(1, N)$, and then draw multiple distinct snippets from $\mathcal{X}_n$. This ensures that for each instance in $\mathcal{B}$, there are always a sufficient number of samples from the same title to act as hard negatives. This is important since $B \ll N$, hence for $n \sim \mathbb{U}(1, N)$ and $m \neq m'$, $\mathsf{P}(x_{n,m} \in \mathcal{B} \wedge x_{n,m'} \in \mathcal{B}) \rightarrow 0$. In other words, the naive random sampling policy of $x_i \sim \bigcup_{n=1}^{N} \mathcal{X}_n$ would mainly lead to easy cross-title negatives.

We can now formulate the training objective. Considering a cross-modal setup, $\mathcal{B} = \{(v_i, a_i) | i \in [1 \cdots B]\}$ represents a minibatch of size B, where video and audio modalities of the i^{th} instance are denoted by v_i and a_i. We use z_v^i and z_a^i to represent their respective embeddings. For the i^{th} element in the minibatch, (z_v^i, z_a^i) serves as the positive pair, while assuming negative pairs for both modalities, $\mathcal{N}_i = \{(z_v^i, z_a^j), (z_v^j, z_a^i) | j \in [1 \cdots B], i \neq j\}$ constitutes the set of negative pairs. With that, Equation 3.1 shows the cross-modal normalized temperature-scaled cross-entropy objective [86] associated with the i^{th} instance. Since $(v, a) \in \{(v_p, a_s), (v_s, a_p)\}$, in practice we optimize Equation 3.2 which aggregates over all available instances.

$$\ell_i(v, a) = -\log\left(\frac{e^{((z_v^i)^\intercal(z_a^i))/\tau}}{e^{((z_v^i)^\intercal(z_a^i))/\tau} + \sum\limits_{(z_v', z_a') \in \mathcal{N}_i} e^{((z_v')^\intercal(z_a'))/\tau}}\right) \tag{3.1}$$

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{B}\Big(\ell_i(v_p, a_s) + \ell_i(v_s, a_p)\Big) \tag{3.2}$$

$$\mathcal{L}_v = \sum_{i=1}^{B} \ell_i(v_p, v_s), \quad \mathcal{L}_a = \sum_{i=1}^{B} \ell_i(a_p, a_s) \tag{3.3}$$

Equation 3.3 shows the within-modal variants of the loss function for video and audio modalities. Unless explicitly mentioned otherwise, we train our models from scratch and *cross-modally*, *i.e.* we compute the contrastive loss between modalities as shown in Eq. 3.2. We do this based on the observation in our early experiments that, when training from scratch without tuning additional scaling parameters, the within-modal contrastive task is too easy comparatively and results in early convergence on

the corresponding terms. This approach is also supported by prior literature [352]. Despite not directly optimizing for within-modal terms, we track $\mathcal{L}_v$ and $\mathcal{L}_a$ during self-supervised pretraining and observe that they diminish as a byproduct of minimizing $\mathcal{L}$. There are variants in our modeling where $\mathcal{L}_v$ and $\mathcal{L}_a$ are included in total loss function (*e.g* $\mathcal{L} + \lambda_v\mathcal{L}_v + \lambda_a\mathcal{L}_a$) which we'll discuss later in Sec. 3.5.2.

3.4.2 Architecture

As we seek to validate the effect of our data and training approach, we rely on standard backbone architectures. Our video model is a multi-scale vision transformer [144], specifically MViT-S, and our audio model follows a similar architecture except a slight modification to allow processing audio spectrograms as input. Note that we train all our models from scratch on our pretraining dataset detailed in Sec. 3.3. We use a single (weight sharing) audio backbone which processes all audio spectrograms, regardless of language. As is common in contrastive learning, we use multi-layer perceptron (MLP) projection heads, one for each modality, to further reduce the dimensionality of representations during training, prior to computing the contrastive loss. These additional layers are discarded after pretraining.

3.5 Experiments

3.5.1 Downstream Tasks

Audio Tasks and Benchmarks We evaluate on a diverse set of auditory tasks to probe the quality of our learned representations, taken from the HEAR [517] challenge benchmark. We subselect tasks relevant to our hypotheses, and focus on those which use pooled (rather than temporally dense) representations.

Sound and Scene Classification: These tasks are firmly non-linguistic, and we hypothesize performance on them should benefit from de-emphasizing language in training. We include ESC-50 [392], FSD50K [157], and Vocal Imitations (VI) [255]. VI is a query-by-vocalization (QBV) task, however since it is based on AudioSet [179] ontology sound events, we place it in this category.

Non-Semantic Speech: Many non-semantic or *paralinguistic* attributes of speech or vocal signals may be shared between languages, and such signals are important for a range of tasks. We include here CREMA-D [75] for emotion recognition, GTZAN [520]

for music/speech discrimination, and LibriCount [486] for speaker count estimation. We hypothesize performance should improve, if our scheme increases focus on non-linguistic speech attributes.

Semantic Speech: To probe a potential trade-off, we evaluate on semantic speech tasks. We consider keyword understanding as a proxy for speech recognition that uses pooled representations. To do so, we employ the *full* version of Speech Commands [544] implemented in HEAR [517].

Language: Another way to measure a possible trade-off is by evaluating how models perform on an audio-based language identification task, to see if features useful for this are preserved in learned representations. We include VoxLingua107 Top10 [524] for this reason.

Visual and Audiovisual Tasks We also evaluate the visual representations independently, and coordinated with, the auditory representations. Following recent work on representation learning from long-form content [85], we include the LVU [551] benchmark covering various aspects of long-form video understanding to our evaluation suite. LVU [551] contains small-scale tasks covering a wide range of aspects of long-form videos, including content understanding (*relationship*, *speaking style*, *scene/place*), and movie metadata prediction (*director*, *genre*, *writer*, *movie release year*). Among the LVU tasks, we explore benefits and potential trade-offs using both visual and auditory representations. In general, we expect improvement except for *speaking style*, where it is not *a priori* clear whether de-emphasizing spoken words during pretraining is harmful for such a downstream task.

Evaluation Once the self-supervised pretraining is over, we discard the projection heads and use the backbone architectures to extract features from audio and video assets. Unless mentioned otherwise, we do spatio-temporal mean pooling on the output tensors in order to obtain a d-dimensional vector embedding for each data instance in the downstream tasks. We then train either an MLP or linear probe on these representations following the prescribed approaches in the relevant benchmarks. More implementation details can be found in the appendix.

3.5.2 Models

In total, we train **11** model variants, detailed in Table 3.1, and evaluate them on **15** different tasks across audio and video modalities.

	# data	init.	(λ_v, λ_a)	original language	avg. # dubs	dub augment
A.1	4.6M	rand.	(0,0)	**ESF**	2.8	✗
A.2	4.6M	rand.	(0,0)	**ESF**	2.8	✓
A.3	4.6M	A.1	(0,0.2)	**ESF**	2.8	✗
A.4	4.6M	A.2	(0,0.2)	**ESF**	2.8	✓
B.1	11.8M	rand.	(0,0)	**EN**	1.0	✗
B.2	9.8M	rand.	(0,0)	$\mathbb{U}\backslash$**EN**	0.2	✗
B.3	19.4M	rand.	(0,0)	$\mathbb{U}$	0.6	✗
B.4	5.1M	B.3	(0,0)	$\mathbb{U}$	2.8	✓
B.5	5.1M	B.3	(0.2,0.2)	$\mathbb{U}$	2.8	✓
C.1	19.4M	rand.	(0,0)	$\mathbb{U}$	0.6	✓
C.2	5.1M	C.1	(0.1,0.1)	$\mathbb{U}$	2.8	✓

Table 3.1: Details of different pretraining model variants. Here, **ESF** $\coloneqq$ {**EN**, **ES**, **FR**} is denoting the union of three languages. $\mathbb{U}$ represents the universal set including all the seven languages.

	HEAR							LVU						
	ESC	LibCnt	CREMA	VI	FSD	Speech	VoxLng	Director	Genre	Relation	Scene	Speak	Writer	Year
A.1	77.20	67.29	59.52	10.37	44.52	74.83	27.16	44.86	54.42	36.59	**45.12**	42.86	**38.10**	41.84
A.2	75.95	67.94	59.76	11.14	44.23	73.80	23.87	47.66	56.63	36.59	41.46	40.74	33.33	41.84
A.3	82.00	67.87	**62.69**	11.39	48.90	**79.47**	**28.70**	**49.53**	57.65	43.90	39.02	43.92	33.93	46.10
A.4	**83.05**	**68.65**	61.95	**12.57**	**49.42**	74.38	26.55	44.86	**59.01**	**46.34**	**45.12**	**48.15**	29.17	**47.52**
B.1	84.15	67.12	61.00	**13.05**	50.29	82.31	24.69	47.66	57.14	51.22	41.46	42.33	32.14	**45.39**
B.2	82.00	67.10	61.98	11.86	49.07	82.90	28.09	42.99	55.95	48.78	42.68	47.62	30.36	44.68
B.3	**85.60**	66.31	62.79	11.55	**53.69**	**83.82**	**30.35**	50.47	**60.20**	46.34	42.68	48.68	37.50	**45.39**
B.4	83.75	68.88	63.18	10.82	51.61	77.12	28.19	**51.40**	59.69	**56.10**	46.34	**49.21**	**38.10**	44.68
B.5	85.25	**69.16**	**63.27**	11.38	52.48	76.99	27.98	**51.40**	58.33	51.22	**52.44**	48.68	36.31	**45.39**
C.1	84.10	67.57	63.70	**12.12**	51.96	**81.88**	29.42	42.99	**58.84**	48.78	46.34	41.27	38.69	41.13
C.2	**85.50**	**68.90**	**64.28**	11.90	**52.55**	77.14	**29.94**	**48.60**	57.65	48.78	**51.22**	**50.79**	**39.88**	**49.65**

Table 3.2: **Ablation results with audio.** All metrics are top-1 accuracy, except for FSD50K [157] and Vocal Imitation [255] (Mean Average Precision). We have followed the prescribed evaluation strategy from HEAR [517] benchmark; training an MLP on frozen embeddings of the downstream tasks. For LVU [551], we use the official data splits and train a linear probe. Results are shown on the test split where the best epoch to report is chosen based on the same metric on the validation set. All model variants obtained 100.0 top-1 accuracy on GTZAN, hence we did not include that task here. We denote the top performance(s) within each ablation group with **bold**. The HEAR [517] tasks from left to right are ESC-50, LibriCount, CREMA-D, Vocal Imitation, FSD-50k, SpeechCommands (Full), and VoxLingua107 Top10.

	ESC	LibCnt	CREMA	VI	FSD	Speech	VoxLng	GTZAN
Bench [517]	96.65	78.53	75.21	22.69	65.48	97.79	72.02	99.23
Bench (SSL)	80.50	78.53	75.21	18.48	50.88	96.87	71.40	96.86
GURA [558]	74.35	68.34	75.21	18.48	41.32	94.68	71.40	93.59
PaSST [267]	94.75	66.01	61.04	18.20	64.09	63.87	25.93	97.69
CLAP [133]	96.70	77.83	64.36	–	58.59	96.83	–	100.0
Ours								
B.3 (A)	85.60	66.31	62.79	11.55	53.69	83.82	30.35	100.0
B.4 (A)	83.75	68.88	63.18	10.82	51.61	77.12	28.19	100.0
B.5 (A)	85.25	69.16	63.27	11.38	52.48	76.99	27.98	100.0

	Director	Genre	Relation	Scene	Speak	Writer	Year
Obj Tr [551]	58.90	56.10	54.70	60.00	40.30	35.10	40.60
M2S [85]	70.90	55.90	71.20	68.20	42.20	53.70	57.80
ViS4mer [231]	62.61	54.71	57.14	67.44	40.79	48.80	44.75
SCALE [444]	49.09	58.97	76.47	74.02	42.27	62.76	39.23
STCA [124]	66.70	56.62	59.25	69.15	41.62	52.93	53.30
Ours							
B.3 (V)	69.16	60.88	60.98	63.41	46.03	48.81	52.48
B.4 (V)	67.29	61.73	60.98	65.85	47.62	41.67	55.32
B.5 (V)	69.16	64.29	58.54	64.63	46.03	41.07	52.48

Table 3.3: State-of-the-art results across HEAR [517] (adding GTZAN Music/Speech) and LVU [551] tasks we evaluate on. On HEAR, we compare to (1) the best result on each task, on the HEAR leaderboard, (2) same as (1) but considering only self-supervised models, (3) GURA Fuse HuBERT [558], the best performer on average, (4) CP-JKU PaSST 2lvl+mel [267], the strongest average performer after the GURA models, (5) the recent CLAP model [133]. On LVU, we compare to the Object Transformer from the original LVU paper [551], along with recent advances: ViS4mer [231], the SVT SCALE model [444], STCA [124], and Movies2Scenes [85]. Movies2Scenes uses movie metadata, which introduces task-specific supervision. When reporting our results, (A) indicates audio representations only, and (V) means video representations only.

First (A) group of model variants demonstrates a small-scale multilingual pretraining regime, as a first study of the impact of dub-augmentation. We sample English (**EN**), Spanish (**ES**), or French (**FR**) titles which have at least one dub available, so we can systematically study the effect of dub-augmentation. For each title, we sample dubs from *all* seven total languages. **A.3** and **A.4** variants incorporate an explicit within-modal term, *i.e* $\mathcal{L}_a$. We hypothesize that, with dub-augmentation, $\lambda_a > 0$ may yield a broader gap on linguistic and language identification tasks. This is because the optimization explicitly maximizes the similarity of audio embeddings that are only different in their spoken language, rather than just implicitly through $\mathcal{L}$. Importantly, the total number of pretraining steps is the same for **A.3** and **A.4** , similarly when one compares **A.1** and **A.2** .

	Director	Genre	Relation	Scene	Speak	Writer	Year
A.1	53.27	54.59	43.90	52.44	34.39	36.90	42.55
A.2	53.27	55.44	41.46	50.00	**41.27**	35.12	42.55
A.3	57.01	**57.48**	**46.34**	**57.32**	39.68	**38.69**	46.10
A.4	**63.55**	**57.48**	36.59	53.66	36.51	33.93	**47.52**
B.1	60.75	55.78	**48.78**	53.66	38.10	35.71	42.55
B.2	54.21	57.65	46.34	51.22	37.04	**38.69**	44.68
B.3	**65.42**	57.48	41.46	53.66	39.68	38.10	45.39
B.4	62.62	**58.50**	36.59	**59.76**	**43.39**	35.12	46.81
B.5	62.62	58.16	43.90	**59.76**	39.15	37.50	**49.65**
C.1	**63.55**	55.10	43.90	57.32	**40.74**	**39.29**	**45.39**
C.2	61.68	**56.63**	**46.34**	**60.98**	40.21	36.90	43.97

Table 3.4: **Ablation results with video**. All metrics are top-1 accuracy. We have followed prescribed data split from LVU benchmark and trained a linear probe on frozen **video** embeddings of the downstream tasks. Results are shown on the test split where the best epoch to report is chosen based on the validation set. We denote the top performance within each ablation group with **bold**.

Second (B) group of model variants aims at understanding the impact of data scale and language diversity. We approximately double the number of pretraining instances compared to experiments in group **A** and study whether this leads to higher quality representations. This is important since self-supervised pretraining is computationally expensive and it is not clear *a priori* if bigger and more diverse pretraining data necessarily leads to better models. **B.3** is trained on all pretraining instances including all languages to test the limit of multilingual pretraining *without* dub-augmentation. By comparing **B.4** and **B.5** , we hope to shed light on the behavior of the within-modal objective function which the latter uses.

Third (C) group of experiments explore the impact of deeper architectures, namely MViT-B [144] (vs MViT-S [144] as our default). We keep the data scale and diversity the same as in the **B.3** , **B.4** and **B.5** variants. Similarly to these, here we initially train on the entire data, then fine-tune from the final checkpoint of **C.1** only on a subset of titles which have more than one audio tracks. This ensures that dub-augmentation is present in every optimization step of **C.2** .

We are now set to comprehensively study various aspects of multilingual and multi-modal representation learning, thanks to a wide variety of pretrained models and downstream tasks across audio and video modalities.

3.5.3 Ablation Study

Does dub-augmented pretraining help? To address this, we start by looking at the **first (A)** group of model variants in Table 3.2. We've hypothesized that dub-augmentation should improve the performance on sound/scene classification and non-semantic speech tasks. On the HEAR [517] benchmark, with the exception of CREMA-D [75], our quantitative results confirm this. LVU [551] tasks are also considered non-linguistic and Table 3.2 shows that, in most of them, dub-augmented variants lead to large performance gains over their baseline counterparts. Our second hypothesis was that dub-augmentation should impact linguistic and language identification tasks as it aims at diminishing the influence of spoken words in audio representations. Indeed, we can see **A.4** which utilizes dub-augmentation is underperforming **A.3** on Speech Commands and VoxLingua. Table 3.2 also suggests that dub-augmentation benefits from within-modal objective *i.e.* $\mathcal{L}_a$, and for this approach to be effective, we actually need as expected, sufficient number of instances with alternative audio tracks during pretraining.

Can dub-augmented models still recognize language and conduct linguistic tasks? Results shown in Table 3.2 on VoxLingua demonstrate that enforcing dub-augmentation in both small (**A** variants) and large-scale (**B** variants) regimes clearly affects language identification performance. We measure this by comparing **A.2** vs. **A.1** , or **B.4** vs. **B.3** . We observe similar behavior for Speech Commands [544], our proxy for linguistic performance implemented as keyword spotting. However, in both cases, the degradation is not large enough to prevent dub-augmented models from recognizing language or conducting linguistic tasks. We hypothesized this modeling trade-off, *i.e.* that while performance might reduce, the significance of this would be limited.

Is the quality of video representations impacted? To answer this, we look at Table 3.4 where LVU [551] tasks are evaluated via a linear probe on frozen video embeddings. In the small-scale pretraining regime, we observe a mixed pattern where dub-augmented variants, *i.e.* **A.2** and **A.4** , outperform their counterparts in 3 tasks ("Director", "Speaking Way", and "Year") while being either worse or on par on the rest. In the large-scale pretraining regime, we see a more clear trend where **B.4** and **B.5** show improvements over **B.3** in 5 out of 7 LVU tasks demonstrating that on a diverse evaluation set, dub-augmented pretraining is overall helpful to even video-only tasks.

How does language diversity influence pretraining? Properly addressing this research question demands a closer look at **B.1** , **B.2** , and **B.3** . It is worth reiterating that despite a different number of pretraining instances (see Table 3.1), we have trained all three of these model variants with approximately the same number of gradient optimization steps to establish a fair comparison. In general, across both audio (ref. Table 3.2) and video (ref. Table 3.4) we observe performance gains when we maximize language variation (ref. **B.3**). However, the inclusion of English (**EN**) language titles, as our most dominant original language (see Fig 3-3), during pretraining seems to be crucial. Table 3.2 illustrates a clear pattern for VoxLingua [524] and Speech Commands [544], where greater language diversity during pretraining leads to significant gains *e.g.* absolute 5.6% on VoxLingua [524].

Is a deeper architecture better? For each task in Tables 3.2 and 3.4, we can compare the strongest **B** model variants against **C** variants. With a few exceptions, our quantitative results do not indicate that using MViT-B [144] with ∼40% more parameters provides a meaningful boost over the smaller MViT-S [144] to justify the significant additional computation during pretraining. We acknowledge that this conclusion might not have held if downstream tasks where evaluated by fine-tuning (instead of linear/MLP probing), especially for large-scale tasks in HEAR [517].

Additional Experiments In the appendix, we provide additional results on a small dubbed audiovisual dataset with matched smaller backbone architectures, where we have exact parity between four languages (over 700 **EN** titles with all of **ES**, **FR**, and **JA** available). We also compare to a speech-removal strategy, where we source-separate the full dataset and remove the speech part as an alternate strategy for de-emphasizing the speech. Since we have language parity, we also evaluate "bilingual" models with specific dub-augmentation pairs (e.g. **EN**+**ES**). These results show systematically that dub-augmented training is beneficial even in this smaller-scale setup, that it outperforms the speech removal strategy, and that multilingual models (with multiple dubs, randomly sampled as in our main results here) can add further robustness.

3.5.4 Comparison with State-of-the-Art

HEAR Table 3.3 compares our results to several strong results on HEAR [517] tasks. On ESC-50, FSD50K, and GTZAN Music/Speech, our results beat the top self-supervised result on the HEAR Leaderboard and at least one more result. On

most tasks (except Vocal Imitation), we beat at least one of the models, showing robustness across these different tasks.

LVU Also in Table 3.3, we compare our strongest models with state-of-the-art results on 7 LVU [551] tasks. Our models achieve new state-of-the-art performance on the *Genre* and *Speak* tasks, showing substantial improvements over prior results. Without considering Movies2Scenes [85], which uses movie metadata, we also get state-of-the-art results on *Director* and *Year* (4/7 total). On the remaining tasks, our results are highly competitive. This demonstrates that models pretrained on our dataset with dub-augmentation can match or exceed the performance of the best available models on a diverse range of video understanding benchmarks. Overall, these results highlight the effectiveness of our approach.

3.6 Synthetic Counterfactual Pairs

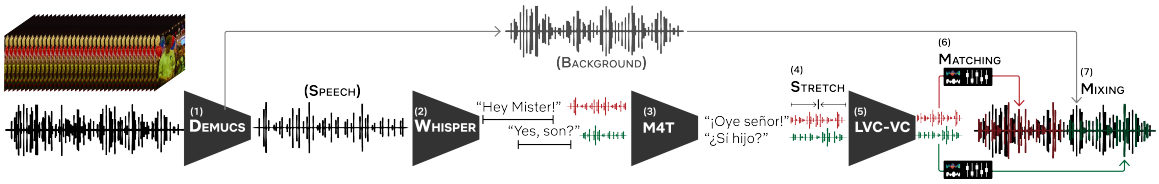

Figure 3-5: Pipeline to produce the synthetic counterfactual pairs.

To encourage the study of counterfactual pairs in audiovisual representation learning, we propose a modular pipeline, shown in Fig. 3-5, for simulating dub-like counterfactual pairs that are similar to the one-to-many audio-visual distribution from our pretraining data on arbitrary target clips. The proposed pipeline, while being limited in terms of the synthetic quality, serves as a simple tool to alleviate the data constraint for the research community when conducting a similar study.

The steps are (1) Isolate speech from background sounds using Demucs [120], (2) Transcribe and segment the speech using Whisper [403], producing timestamped segments (3) Translate speech (or, optionally, text) into the target language(s) with SeamlessM4T [35] (4) Align translations to original segments using stretching (5) Convert voices to match original actors' using LVC-VC [244] (6) Loudness-normalize and EQ-match the output with the original using Pyloudnorm [484] and matchering[5] (7) re-place segments into their original locations, remix with background audio, and

[5] https://github.com/sergree/matchering

mux with original videos. The pipeline also implements other intermediate steps, such as resampling, to bridge between the main steps.

As a proof-of-concept resource for the community, we use this pipeline to produce a multilingual version of LVU [551]. LVU-M demonstrates the feasibility of generating counterfactual data at scale. We will open-source the pipeline to enable creating such *"looking similar, sounding different*" datasets. We also hope that future advancements can improve the quality and enable deeper research of such data structure.

3.7 Conclusion

In this work, we introduced the *looking similar, while sounding different* problem, wherein perceptually similar scenes can have different speech content. We showed we can leverage a similarly structured counterfactual data source, dubbed movies, to improve audiovisual representation learning in a well-established cross-modal contrastive learning scheme. Our experiments with a large pretraining dataset of movies and television shows demonstrated that this improves performance across a range of auditory and audiovisual tasks. Dub-augmented training is, as such, a scalable and effective approach for learning more robust audiovisual representations without supervision.

Part II

Synthesizer Programming by Humans and Machines

4

SynthAX: A Fast Modular Synthesizer in JAX

The history of electronic music is, in many ways, a history of human-machine interaction: of artists and engineers developing new interfaces for controlling and shaping sound. Modern sound synthesis often prioritizes real-time interaction, enabling the tight feedback loops essential to creative exploration. What new possibilities might emerge if we could dramatically accelerate synthesis beyond real-time speeds? In particular, this might hold promise for human-AI interactions; interactions in which the ability to rapidly tweak and generate sounds could be useful for machines to iterate in concert with human use. This chapter introduces a high-performance software modular synthesizer implementation that serves as technical infrastructure for several subsequent investigations. The system's design emphasizes interpretability and controllability, but seeks the computational efficiency needed for interactive applications in intelligent audio production.

Abstract

Modern audio production relies heavily on realtime audio synthesis. However, accelerating audio synthesis far beyond realtime speeds has a significant role to play in advancing intelligent audio production techniques. Fast synthesis methods have been used to generate useful datasets, implement audio matching procedures for automatic sound design, and infer synthesis parameters for real-world sounds. In this chapter, we present SynthAX, a fast virtual modular synthesizer written in JAX. At its peak, SynthAX generates audio over 80,000 times faster than realtime, and significantly faster than the state-of-the-art in accelerated sound synthesis. We present SynthAX as an open source[1]and easily extensible API to stimulate and support applications of

fast sound synthesis at scale.

4.1 Introduction

Realtime sound synthesis is a cornerstone of modern audio production. It affords producers the ability to tweak sounds and hear them change; a loop of perception and action that results in diverse auditory creations to support music, film, and other media. Modern audio technologies increasingly employ techniques that benefit from *automatically* tweaking synthesizers, such as optimization and machine learning. In these scenarios, the ability to rapidly tweak sounds and compute with them at scale offers a vast space of opportunities for designing and developing powerful new audio technologies. As such, fast sound synthesis can be an essential tool. We define faster-than-realtime as generating more than one second of audio per second of processing time. In particular, we deal with cases where the processing is a lot faster than this (i.e. $>$1000x).

In this chapter, we introduce SynthAX, a fast modular synthesizer written using the JAX [59] framework for accelerated and differentiable computing. By offering synthesis at speeds that peak at over 80,000$\times$ realtime, SynthAX provides a high-performance, flexible virtual modular synthesizer in the form of an expanding and easily extensible open source Python library. Additionally, we implement an API based on *torchsynth* [516], a recent high-performing synthesizer written in PyTorch, to allow for an easy substitution for end-users. Our results in this chapter show considerable speedups over *torchsynth*, ranging up to just under 9$\times$ depending on the hardware configuration and batch size.

4.2 Related Work

4.2.1 Programmatic Synthesis

One important element of SynthAX is allowing programmatic control of a synthesizer. Indeed, many software synthesizers are ultimately written to be controllable by other software, such as VST plugins by DAW automation. However, not many synthesizers are designed to be fully specifiable and controllable in code written by end-users.

[1] https://github.com/PapayaResearch/synthax

Some well-known options include *Surge XT*[2] and *torchsynth* [516]. The former is written as a plugin that offers an API, and the latter is written as a library for non-realtime synthesis. We implement ours following the latter example, which means that there is not a direct application of our method to realtime synthesis. However, since JAX [59] compiles code to XLA, it is likely possible to implement SynthAX in a realtime synthesizer plugin to have it bridge these two different approaches to programmatic synthesis.

4.2.2 *torchsynth*

Developed for audio synthesis, *torchsynth* [516] serves as a modular synthesizer that is capable of generating audio on a single GPU at $\geq$16200$\times$ faster than realtime. It consists of a variety of audio and control modules. The default synthesizer in *torchsynth* is *Voice*, which the authors used to generate a dataset containing a billion audio clips. As we detail later, we base our API and implementation on *torchsynth* as it provides an existing and familiar reference point. We also compare to *torchsynth* in our experiments studying the performance of SynthAX.

4.3 System Design

The design of the API is inspired by the inherent modularity of hardware synthesizers. SynthAX leverages the power of JAX [59] to build on *torchsynth* [516], which is a state-of-the-art high-throughput synthesizer implemented in PyTorch to take advantage of its accelerated computational routines. Maintaining a similar API makes the transition for end-users seamless without any major rewriting or learning curve.

Each module serves a different function but can be connected together to create a synthesizer. SynthAX modules mimic their counterparts in analog and digital synthesizers, consisting of amplifiers, envelopes, filters, keyboards, low-frequency oscillators (LFOs), mixers, and voltage-controlled oscillators (VCOs). The output from these modules can represent audio signals or control voltages, depending on the module's intended function. Audio modules, such as VCOs, produce audio signals. Control modules, such as LFOs, produce "control voltages" that modulate the parameters of other modules. The keyboard outputs parameters that are used as input for other modules. All modules follow the Flax [209] module system known as

[2] https://surge-synthesizer.github.io/

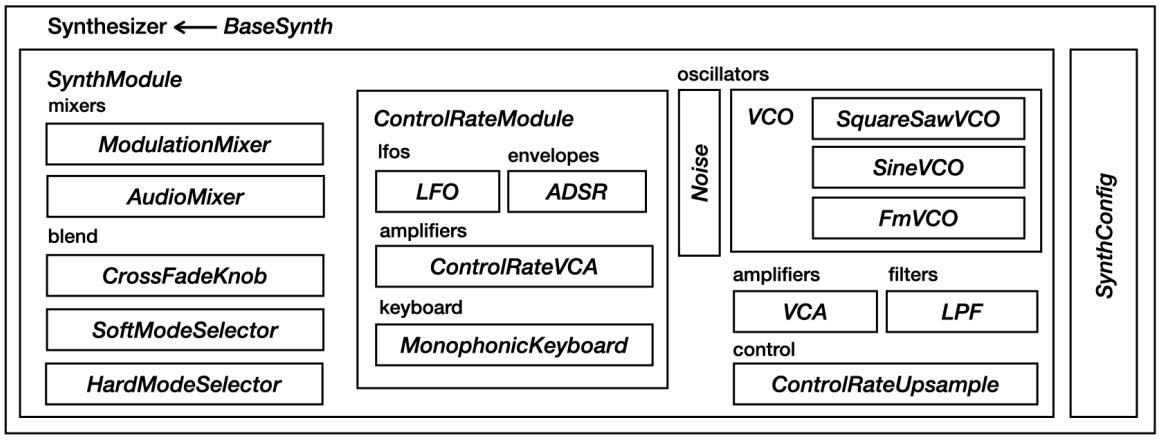

Figure 4-1: Structure of the API. We separate the synthesis modules into Python modules which group related elements. These modules are shown in lower-case letters above the relevant classes. The class inheritance structure, which mirrors *torchsynth* [516], is indicated by the *TitleCase* names. Inner boxes are subclasses of the larger boxes they are embedded in.

Linen to organize the modules into independent components. Figure 4-1 shows the structure of the API, where a synthesizer consists of modules and a configuration.

In our implementation, we aim for allowing users flexibility in how they specify synthesizers. Modules with parameters can be initialized in a few different ways. If only initial values are given, they are expected to be in human-readable (i.e. unnormalized, e.g. frequency in Hz) range within the default ranges of the parameters. Alternatively, the modules also accept range objects, which specify only a range within which parameter values are initialized uniformly randomly. Finally, users can also provide the initial values and ranges together as an object. In all cases, the parameters store the values in the (normalized) interval [0, 1].

In addition to the differences between SYNTHAX and *torchsynth* that arise from JAX features such as easy and flexible vectorization, parallelization, and just-in-time (JIT) compilation, we introduce these additional features: a filter module, currently containing a simple low-pass filter that can be shaped by control modules; a parametric definition of a synthesizer to easily explore different synthesizer topologies; functions to write and load a synthesizer including its hyperparameters and parameters, in the human- and machine-readable YAML format. This also means that synth specifications can, in principle, be directly composed in YAML and loaded to a synth with a matching parameter architecture.

Listing 1 Code snippet for generating audio with a *ParametricSynth*. This synthesizer supports a user-configured architecture, in contrast to the *Voice* synthesizer which encodes a fixed topology design (78 parameters). This allows control of the degrees of freedom available to manipulate the sound synthesis.

```
import jax
from synthax.config import SynthConfig
from synthax.synth import ParametricSynth

# Generate PRNG key
config = SynthConfig(
    batch_size=16,
    sample_rate=44100,
    buffer_size_seconds=4.0
)
# Instantiate synthesizer
synth = ParametricSynth(
    config=config,
    sine=1,
    square_saw=1,
    fm_sine=1,
    fm_square_saw=0
)
# Initialize and run
key = jax.random.PRNGKey(42)
params = synth.init(key)
audio = jax.jit(synth.apply)(params)
```

We adhere to JAX's explicit randomness handling in our design. JAX uses a pseudo-random number generator (PRNG), an algorithm that produces sequences of numbers that approximate true randomness given an initial key (i.e. value). Therefore, users need to provide such random keys to their synthesizers. Though this adds an extra consideration, it also ensures better reproducibility. Listing 1 shows how to define a configuration, instantiate a parametric synthesizer and, finally, synthesize audio.

JAX supports a wide variety of hardware and leverages powerful function transformations such as just-in-time compilation (JIT), auto-vectorization, and hardware parallelism. We can vectorize (jax.vmap) and parallelize (jax.pmap) in a single line of code. It also conforms to the *Single-Program, Multiple-Data* (SPMD) model, which means that the same computation for different input data runs in parallel on multiple devices. In order to maximize performance and throughput when using JAX,

SynthAX renders audio in batches.

Extending SynthAX can be done seamlessly due to its modularity, since the API is designed to easily integrate other synthesizers or modules. SynthAX joins the JAX ecosystem and can be easily integrated with other well-known libraries such as Optax [23], evosax [281], EvoJAX [499], and QDax [296].

4.4 Results

4.4.1 Performance Evaluation

First, we characterized the speed and memory performance of SynthAX. We used *torchsynth* [516] as a strong baseline to compare against, since it is the *de facto* state-of-the-art fast synthesizer and can take advantage of similar hardware acceleration capabilities (e.g. GPUs). For both synthesis libraries, we use the *Voice* synthesizer with 78 parameters. In our setup, we computed the time needed to synthesize 100 batches of sounds at different batch sizes (powers of 2 from 2 to 1024). We randomized the synthesis parameters for each batch. As *torchsynth* does, we also report the speed as compared with realtime synthesis. We calculated this as

$$\frac{\text{Num. Batches} \times \text{Batch Size} \times \text{Sound Duration}}{t}$$

where t denotes the time taken for one loop of 100 batches. Finally, we also report memory usage in GB after each 100-batch loop.

We report averages over 10 100-batch loops for all three quantities (time, speed $\times$ realtime, and memory), to account for variance. Additionally, we computed a full set of results for a GPU and a CPU, although we expect GPUs to be the primary usage platform. To account for the effect of JAX's [59] JIT compilation, we produced one batch of sounds (for both SynthAX and the *torchsynth* baseline) at the very beginning, outside the evaluation loop. This is so that we measure the typical performance, as the JIT compilation only needs to occur once.

These results are given in Figure 4-2. We do not show error bars as the results are generally stable, resulting in very small variance. Overall, we see that SynthAX substantially outperforms *torchsynth* on time-based metrics for both CPU and GPU.

At peak performance within this evaluation, SynthAX shows more than 80,000× realtime synthesis speed. SynthAX shows a comparable memory utilization profile to *torchsynth*, especially lower at higher batch sizes on GPU and CPU. We disabled JAX's memory preallocation for our experiments to measure the real memory footprint.

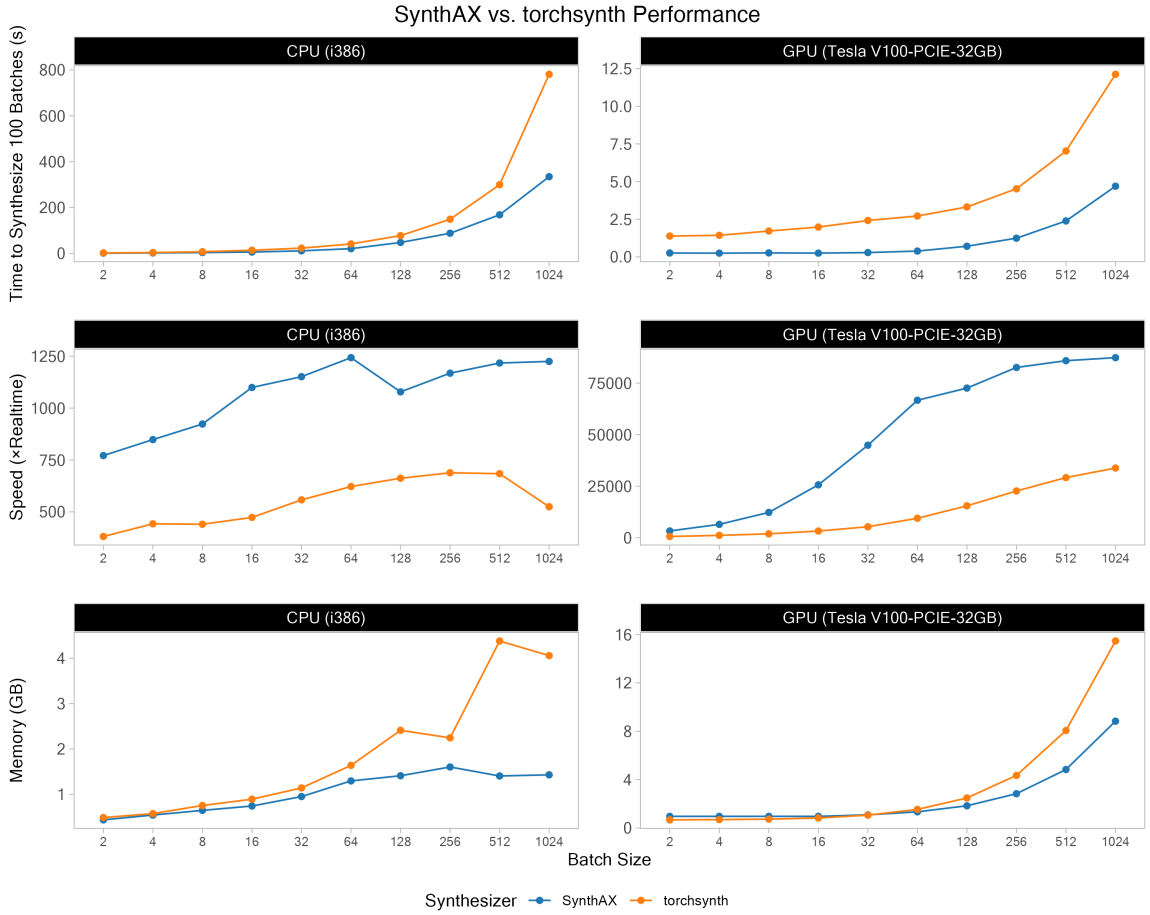

Figure 4-2: Results from performance evaluation, compared with *torchsynth*, on **(Left)** a 2017 iMac with an Intel Core i7-7700K CPU @ 4.20GHz, and **(Right)** an NVIDIA Tesla V100 GPU. Values shown are averaged over 10 runs. We use the *Voice* synthesizer in both SynthAX and *torchsynth*, randomizing parameters each batch. **(Top)** Time to synthesize 100 batches of sound at different batch sizes (given in seconds). **(Middle)** Time reinterpreted as speed × realtime, i.e. seconds of sound generated per second of computation time (see §4.4.1 for details). **(Bottom)** Memory utilization in GB. Overall, SynthAX shows significantly faster performance while retaining a similar memory utilization profile.

For direct comparison, Figure 4-3 plots the speedup over *torchsynth*. This is computed as the ratio of time taken to synthesize 100 batches, computed per 100-batch loop,

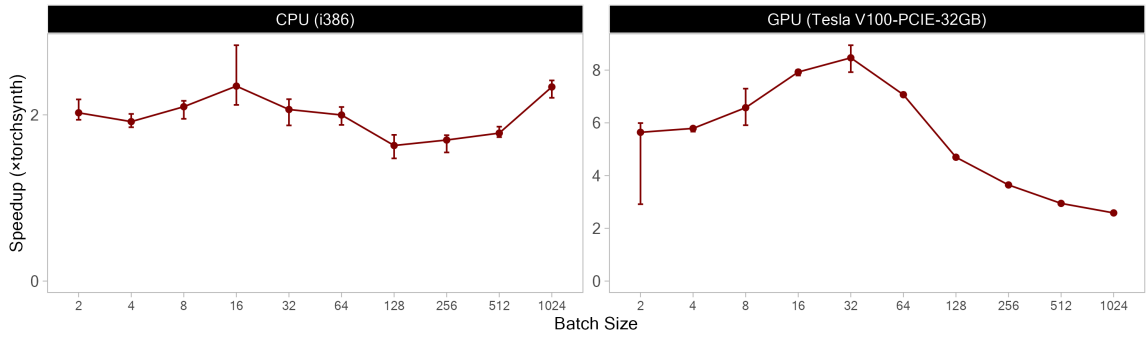

Figure 4-3: A direct comparison showing speedups relative to *torchsynth* [516] per batch size, again for 100-batch total times averaged across 10 runs. Error bars here show min/max results. Overall, SynthAX is more than double the speed in all cases, and peaks at almost 9× the speed of the already accelerated *torchsynth* implementation. As previously, these results are on the *Voice* synthesizer, a 78-parameter synthesizer, where parameters are randomized for each batch.

and then averaged across the 10 runs. We provide min/max error bars to show the full range. This figure shows that the speedups range from just over 2× (some batch sizes on CPU and very large batches on GPU) to almost 9× at the peak speedup level (batch of 32 sounds on GPU).

4.4.2 *torchsynth* Replication

We replicated the examples from *torchsynth* [516] for reproducibility. These include instantiating ADSR envelopes both randomly and with set parameters, VCOs, LFOs, VCAs, mixers, and their synthesizer architecture *Voice*. Figure 4-4 shows some of the resulting spectrograms considering different VCOs and setups in *torchsynth* and corresponding match in SynthAX.

4.5 Applications

4.5.1 Audio Representations

An advantage of synthesized sounds is that they also contain the associated synthesis parameters. In self-supervised representation learning problems, datasets that result from synthesis can be used to formulate parameter prediction problems. For instance, pitch recognition is a prominent auditory processing problem for which synthesized datasets hold significant promise. Recent work on audio representation learning

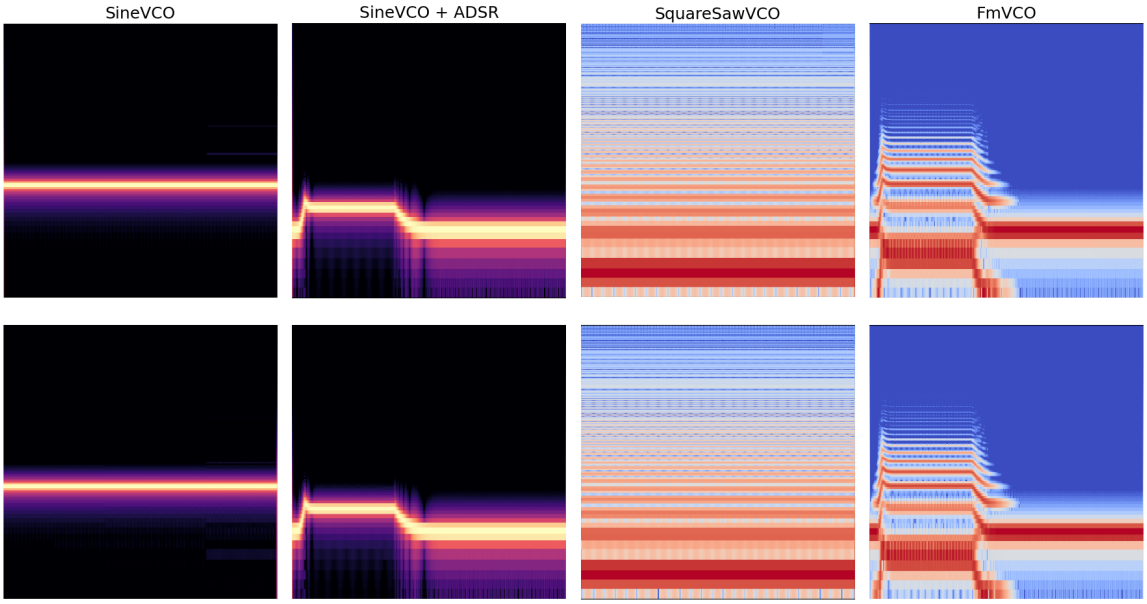

Figure 4-4: Spectrograms for the examples in *torchsynth* (**Top**) and the replication in SynthAX (**Bottom**). From left to right, we show a simple sine wave, a sine wave with an ADSR envelope modulating the frequency, a square wave, and an ADSR envelope-modulated FM patch. The results show clear replication of the output spectrotemporal features.

has employed the Surge XT pitch dataset [516] to evaluate representations on such a task [364, 365]. Many other such prediction problems could be formulated for both training and evaluation, as they expose ground truth information as labels. A synthesizer can generate a large variety of sounds that vary in timbre while holding pitch constant, or conversely which vary in pitch but hold timbre constant for a task such as instrument recognition.

4.5.2 The Synthesizer Programming Problem

One particular area where SynthAX can be useful is in the synthesizer programming problem [461], and specifically the task of parameter inference [171]. A canonical formulation of this asks an algorithm to program a synthesizer to match a given sound. The difficulties of manually programming complex synthesizers are well-established [454], and as such a variety of techniques [142, 171, 203, 310, 334, 335, 417, 453] and even software libraries [462] have been developed to approach this through the lens of automatic matching. Typically, algorithms used are those common to other search and optimization problems, such as genetic algorithms and

even gradient-based optimizers. Given a sound, these algorithms seek to minimize some measure (often perceptually-motivated) of the "distance" between the target sound and a synthesized candidate by tweaking the synthesis parameters. SynthAX can accelerate such applications by speeding up the synthesis, often the most costly step in these problems. Additionally, SynthAX can be combined with other parts of the pipeline written in JAX [59] (such as evosax [281]) to provide a broader speedup for synthesizer programming by matching target sounds.

4.6 Conclusion

In this chapter, we presented SynthAX, a fast modular synthesizer implemented in JAX. We showed that SynthAX generates sounds orders of magnitude faster than realtime, and significantly faster than existing solutions to accelerated sound synthesis. We discussed the possible applications of this synthesizer in research and production problems involving intelligent sound processing and synthesis. In the future, we intend to expand it with more modules and a user interface. By open sourcing this library, we invite contributions towards a high-performance, robust, and well-documented synthesizer that we hope will eventually parallel commercial software synthesizers in the range of possible sounds producible, while retaining the performance benefits which we observe in our experiments on this initial implementation.

5

Creative Text-to-Audio Generation via Synthesizer Programming

Modern generative models have enabled remarkable feats of content synthesis. Yet, these often come at the cost of interpretability and control. When generating audio from text descriptions, for instance, end-to-end approaches can produce impressive acoustic results from even a simple text prompt but offer users little agency over the generation process. Often, this means tweaking the prompt, with no guarantees about preserving any desirable aspects. More recently, it may mean using a small number of post-hoc controls, but these must be learned from correlations at model training time and therefore have significant limitations. There is another limitation of current neural audio synthesis models, in that they are trained largely on recordings, and therefore tend to synthesize acoustically realistic versions of sounds. Often, in creative sound design work, we instead seek abstractions: more sketch-like ways to artistically evoke concepts without replicating what we might get from real-world recordings.

This chapter builds on the synthesizer introduced in Chapter 4 to propose an alternative. Rather than direct waveform synthesis, we generate interpretable synthesizer parameters that users can understand and manipulate. This apparent constraint—working within the bounded space of synthesis parameters rather than unlimited acoustic possibilities—can actually enable new forms of creative expression. Like Chapters 2 and 3, it deals with a cross-modal setting for sound. However, here the second modality is textual, not visual. Additionally, the method uses inference-time optimization rather than pre-training, relying on a pre-trained text-audio joint embedding model to facilitate aligning the modalities.

Abstract

Neural audio synthesis methods now allow specifying ideas in natural language. However, these methods produce results that cannot be easily tweaked, as they are based on large latent spaces and up to billions of uninterpretable parameters. We propose a text-to-audio generation method that leverages a virtual modular sound synthesizer with only 78 parameters. Synthesizers have long been used by skilled sound designers for media like music and film due to their flexibility and intuitive controls. Our method, *CTAG*, iteratively updates a synthesizer's parameters to produce high-quality audio renderings of text prompts that can be easily inspected and tweaked. Sounds produced this way are also more abstract, capturing essential conceptual features over fine-grained acoustic details, akin to how simple sketches can vividly convey visual concepts. Our results show how *CTAG* produces sounds that are distinctive, perceived as artistic, and yet similarly identifiable to recent neural audio synthesis models, positioning it as a valuable and complementary tool.[1]

5.1 Introduction

> *"Of course, bubbles don't make sound, but this is the magic of sound design...you can create the concept of a sound and it seems real."*
>
> — Suzanne Ciani

In creative sound design, realism isn't everything. In the late 1970s, composer Suzanne Ciani famously demonstrated this principle with her iconic *Coca Cola pop and pour* sound effect. This sound, which has become synonymous with the refreshing experience of opening a soda, was not recorded from an actual soda bottle, but skillfully crafted using a Buchla synthesizer. Ciani's work illustrates the immense power of abstraction in auditory representation, where the essence of a concept can be expressed without mimicking real-world acoustic details, while achieving greater impact.

This approach extends beyond single examples into the domain of procedural sound design: creating sounds algorithmically using parameters that can be manipulated to achieve desired sonic effects. By applying procedural techniques, sound designers can often transcend what's physically plausible to obtain by recording real-world

[1] ctag.media.mit.edu

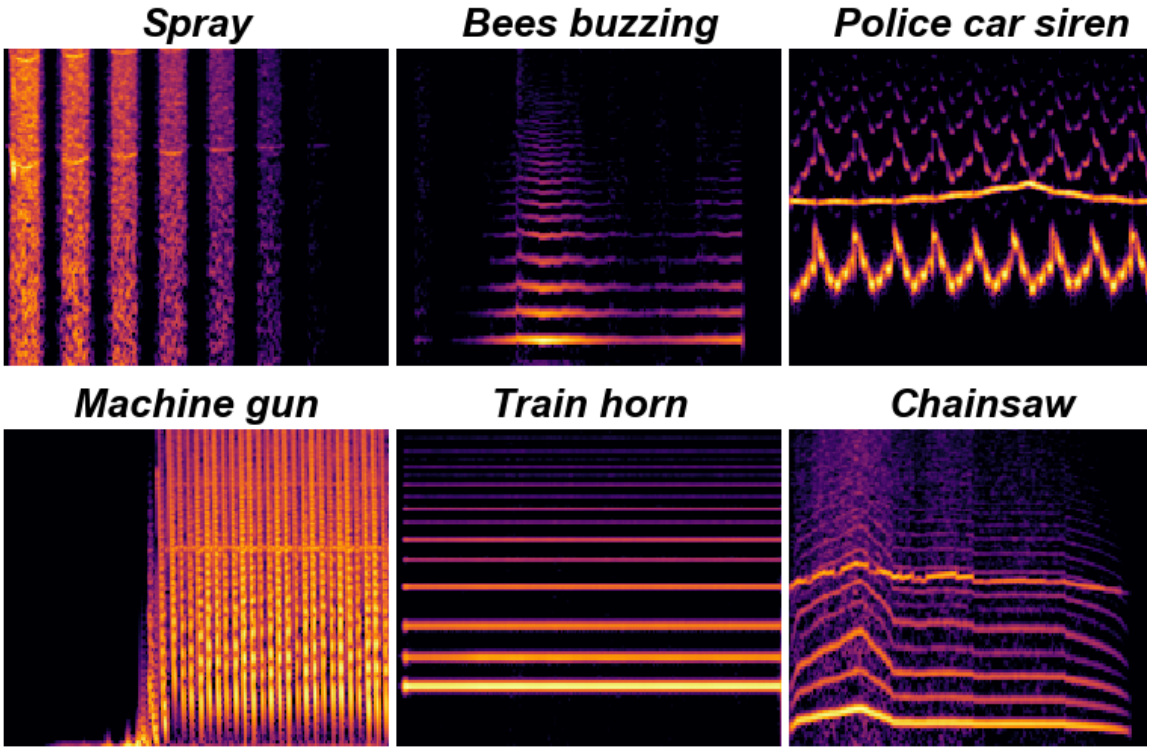

Figure 5-1: *CTAG* leverages a virtual modular synthesizer to generate sounds capturing the semantics of user-provided text prompts in a sketch-like way, rather than being acoustically literal. Spectrograms of auditory outputs corresponding to six text prompts showcase the range of sounds this approach can yield, accompanied by a fully interpretable and controllable parameter space.

events. These methods can lead to highly evocative and expressive sounds in music, film, video games, advertising, product design, and other media.

Neural audio synthesis methods have transformed the state of sound design, enabling specifying sound ideas using intuitive inputs like textual prompts. However, there remains unrealized potential in integrating expressive sound design principles into neural audio synthesis. Current techniques prioritize acoustic recreation and end-to-end application, often overlooking creative possibilities for evoking emotions or concepts, and interactive aspects like manipulating, iterating, and interpolating between sounds. While recent advances showcase remarkable capabilities in replicating real-world sounds, this emphasis can limit the creative palette and expressive potential of generated audio. We propose a method to bridge this gap.

Overall, this work contributes:

- A novel method that integrates a virtual modular synthesizer with a pretrained audio-language model for generating sounds that resonate with human intuition without being literal representations.
- A lightweight, fully interpretable, and controllable synthesizer resulting from our approach, allowing for easy inspection and tweaking for creative purposes.
- Extensive experiments evaluating different approaches to solving this problem, varying optimization algorithms, sound durations, and synthesis architectures.
- Qualitative and quantitative results that highlight how sounds from our method have distinct features from those produced by other neural audio generators, while still being identified at similar rates. We conduct a user study as a gold standard evaluation, given the novelty of the task, which shows the identifiability and potential artistic value of *CTAG*'s sounds.
- Examples of this approach generating several datasets of sounds with their synthesis parameters, and interpolating between different sounds in the parameter space.

We will open-source our approach, both to provide a tool for novices and experts alike to realize their ideas, as well as to provoke future audio generation paradigms that recognize abstraction as an important factor for creative expression.

5.2 Related Work

5.2.1 Sound Synthesis

Neural audio synthesis consists of two main strands: approaches that generate audio waveforms directly in the time domain, and those that do so in the frequency domain. WaveNet [371] notably introduced an autoregressive approach to audio synthesis by predicting one sample at a time. This slow iterative sampling approach, later refined in SampleRNN [339] and WaveRNN [243], reflects the sequential nature of audio data, in contrast to images wherein GANs with global latent conditioning and efficient parallel sampling quickly became a dominant method for synthesis. Later, WaveGAN [127] and GANSynth [136] demonstrated that GANs could in fact be used to synthesize locally-coherent audio, outperforming sequential models' speed

by several orders of magnitude while maintaining a focus on high-fidelity, natural-sounding audio.

A third strand of so-called *oscillator* models, largely propelled by Differentiable Digital Signal Processing (DDSP) [137] is physically and perceptually motivated by the rich history of synthesis and signal processing techniques. Our approach is motivated by this direction, but relies on a simple synthesizer architecture, CLAP [559], for text-conditioning, and gradient-free optimization to provide a simple, training-free solution.

5.2.2 Language-Sound Correspondence

Advances in multi-modal sound-language models have been partly motivated by CLIP [402] for images. Wav2CLIP [553] builds directly onto CLIP by adding an audio encoder, and VQGAN+CLIP [109] generates and edits images guided by text prompts. Audio representation models, such as Microsoft's CLAP [133] and LAION-CLAP [559], emulate CLIP's approach by using contrastive learning on audio-text pairs. We use LAION-CLAP as our audio-language model in this work.

Other recent approaches cast audio generation as a language modeling task. AudioGen [269] is an autoregressive model conditioned on text inputs. AudioLM [58] uses a multi-stage Transformer-based language model. WavJourney [308] uses text instructions to create scripts, which are then used for compositional audio creation. Make-An-Audio 1 and 2 [222, 226] offer text-to-audio synthesis with prompt-enhanced diffusion models, using CLAP to map text to latent representations with a spectrogram autoencoder. AudioLDM [301] learns continuous audio representations from CLAP latents and can perform text-guided audio manipulations. We compare to two state-of-the-art solutions, namely *AudioGen* and *AudioLDM*, in our experiments. Our goals differ significantly from those of these models, as we seek to generate abstract yet high-quality sounds, rather than literal recording-like renditions.

5.2.3 Abstract Synthesis

Visual sketching offers an intuitive analog to abstract sound synthesis. Minimal representations like monochromatic line drawings might use only straight lines and curves with no additional shading or color. These renderings are non-photorealistic; they evocatively convey meaning while emphasizing a subject's essence over its real-world presentation. They can also reveal insights about a subject's underlying geometry,

proportions, and symbolism that may be obscured in more realistic depictions.

The problem of computing recognizable and insightful abstract renderings has seen more progress in the visual than the audio domain. CLIPasso [530] leverages CLIP to distill semantic meanings from images and sketches alike and thereby guide text-to-image generation, varying the number of strokes according to the desired level of abstraction. CLIPTexture [477] enables a user to manipulate a simple sketch or layout through textual descriptions. CLIPVG [478] follows the same progressive optimization approach, but performs image manipulation using vector graphics rather than pixels. ES-CLIP [504] tackles the problem via evolution strategies, generating configurations of colored triangles on a canvas, then assessing their fitness for further iteration. We were inspired by this approach, though we rely on the well-established, easily interpretable, and tweakable paradigm of modular synthesis.

In the auditory domain, the Sound Sketchpad [468] combines sounds together using audio-visual sketches, and the SkAT-VG project [428] applies vocal and gestural manipulation as natural sketching tools. In our approach, we focus on language input, and synthesis rather than the composition of existing sounds.

5.2.4 Interpretable and Controllable Synthesis

Interpretability and controllability of results is essential to human-machine co-creation, in which it is often desirable to closely examine, understand, and fine-tune an artifact. For creative sound design using neural synthesis methods, it can be impossible to retrace decisions made by a complex neural synthesis model en route to synthesizing an output. The model may also not provide any opportunity to iteratively refine the output. Some prior work [571] highlights the potential of program synthesis for interpretability in sequence data, including music. Some neural synthesis models integrate techniques like timbre-regularization [140] to bridge powerful synthesis methods with perceptually-motivated organization of latent spaces. By contrast, our approach offers a fully interpretable and controllable parameter space without requiring us to develop additional neural infrastructure.

5.2.5 The Synthesizer Programming Problem

Despite the near-ubiquitous presence of synthesized sound in modern music, synthesizer programming—that is, the act of creating new sounds through careful analysis and modulation of synthesizer parameters—is a complex task that can often impede

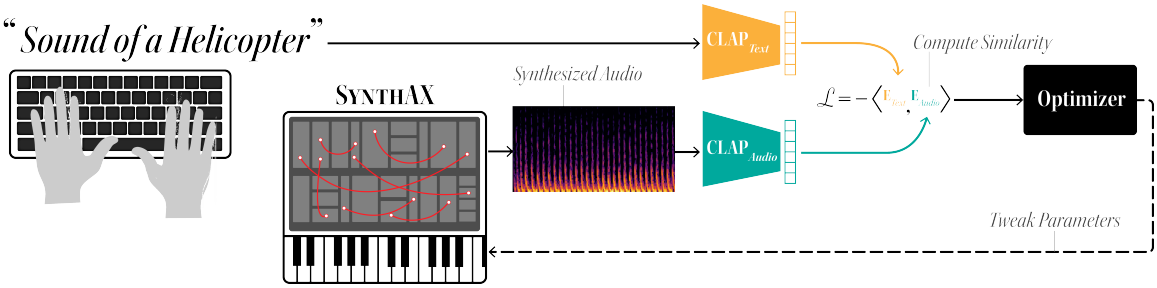

Figure 5-2: High-level overview: we use the LAION-CLAP model [559] to compute the similarity between a user-provided text prompt and SynthAX's [89] output. The optimization procedure iteratively adjusts the parameter settings.

the creative process, if not bar entry entirely. In particular, the conceptual disconnect between parameter settings and the associated auditory output [461] makes synthesizer programming especially non-intuitive without special training. Recent work has investigated techniques for inverse synthesis—given a target sound, infer the parameter setting that will emulate the sound to the closest extent possible—on both musical sounds [569] and real-world sounds, such as animal vocalizations [203], including deep learning methods to learn invertible mappings [141]. However, this task still requires a specific audio clip to start. We provide text-to-parameter inference to bridge this gap, generalizing beyond specific audio files to broader semantic notions of arbitrary sounds.

5.3 Methods

Our methodology hinges on three pillars: a synthesizer, implemented via SynthAX [89], gradient-free optimization methods, implemented via the Evosax [282] evolutionary optimization library, and an objective function based on the LAION-CLAP [559] model, which we use to estimate semantic alignment between the synthesized audio and its corresponding text prompt (see Figure 5-2 for an overview of the pipeline).

5.3.1 Synthesizer

We use a simple synthesizer implementation available in SynthAX, a fast modular synthesizer written in JAX [59]. We specifically use the *Voice* synthesizer architecture, adapted from *torchsynth* [516], which has already been used for programmatic resynthesis of sounds [203]. It consists of 78 parameters for a monophonic keyboard,

two low-frequency oscillators (LFOs), six ADSR envelopes, a sine voltage-controlled oscillator (VCO), a square-saw VCO, a noise generator, voltage-controlled amplifiers (VCAs), a modulation mixer and an audio mixer. All parameters are initialized uniformly, $\theta_i \sim U(0, 1)$.

In addition to this architecture, we evaluate the following variants in increasing order of architectural complexity:

- *ShapedNoise*: An 18 parameter synthesizer consisting of a noise generator, and two control elements to shape the noise amplitude over time: an ADSR envelope, and a low-frequency oscillator (LFO). These are combined into a modulation signal through a modulation matrix, which itself has learnable weights for this combination.

- *OneOsc*: A 23 parameter synthesizer consisting of a sine wave voltage-controlled oscillator (VCO), and the same two control elements as above. These elements are combined into two signals through a modulation matrix, one each for frequency and amplitude.

- *NoLFO*: A 29 parameter two-VCO synthesizer, where one is a sine wave oscillator and the other is a square-saw wave oscillator with a "shape" parameter which controls the degree of "square-ness" vs. "saw-ness". This synthesizer has no LFO components, all modulation is conducted by two ADSR envelopes combined into four separate modulation signals (pitch and amplitude controls for each of the two VCOs).

- *NoNoise*: A 51 parameter synthesizer with two VCOs (as before), and a more complex modulation structure. Here, there is a single LFO, but there are additional ADSRs to modulate the frequency and amplitude of this LFO. The modulated LFO and two ADSR envelopes comprise the inputs to the modulation matrix.

- *Voice+FM*: A 130 parameter synthesizer which adds a frequency modulation (FM) component to the original *Voice* architecture.

For reference, an ADSR envelope is a piecewise control signal consisting of linear or exponential segments: **A**ttack, **D**ecay, and **R**elease, which specify the duration of each envelope segment. The **S**ustain parameter is the level of the control signal

after the decay phase. An LFO is an oscillator whose frequencies are typically lower than audible frequencies, i.e. below 20-40 Hz. These are used for periodic control of synthesis parameters.

In all our experiments, the synthesizer has a control rate of 480 Hz and the audio is generated in batches at a sample rate of 48 kHz. This sample rate is much higher than that commonly used for neural audio synthesis systems (often 16 kHz) and therefore admits much more high-frequency content to be generated.

5.3.2 Optimization

Algorithm 2 Our optimization procedure for producing sounds in *CTAG*. Note: d is the number of parameters of the synthesizer S; for simplicity we omit batches.

Require: Text prompt p
Require: Population/batch size N
Require: Iterations M

Components:
CLAP text embedding model $C_t(p) \rightarrow E^p$
SynthAX synthesizer $S(\Theta) \rightarrow X^a$
CLAP audio embedding model $C_a(X^a) \rightarrow E^{X^a}$
Optimization Strategy: O

Initialize:
Synthesis parameters $\Theta = \{\theta_1, \ldots, \theta_N\}, \theta_i \sim U(0, 1)$
Flattened parameters $\Theta_f \in [0, 1]^{N \times d} = Flatten(\Theta)$

for $i = 1$ **to** M **do**
$\Theta_{f_{\text{new}}} \leftarrow O_{\text{ask}}(\Theta)$ *Generate candidates*
$\Theta_{\text{new}} \leftarrow \text{Reshape}(\Theta_{f_{\text{new}}})$ *Reshape*
$X^a \leftarrow S(\Theta_{\text{new}})$ *Synthesize audio*
$E^{X^a} \leftarrow C_a(X^a)$ *Get audio embeddings*
$F \leftarrow -E^{X^a} E^{pT}$ *Compute fitness*
$O_{\text{tell}}(\Theta_{\text{new}}, F)$ *Update optimizer state*
$\Theta \leftarrow \Theta_{\text{new}}$
end for

$\theta^* = \arg\min_\theta F$ *Select optimal parameters*

During initial experiments, we found the gradients of our differentiable synthesizer to be highly unstable. This instability hindered optimization performance even after attempting mitigation strategies. Recent works in abstract visual synthesis have shown that non-gradient methods can achieve state-of-the-art results without relying

on gradient information [504]. Given these findings, we decided to explore non-gradient approaches which are more suitable for our synthesizer's instability and have demonstrated effectiveness for this task. Focusing efforts here allowed us to sidestep gradient issues while leveraging successful techniques from related synthesis domains.

We experimented with several non-gradient optimization algorithms, using implementations from Evosax [282]. Specifically, we examined simple baselines like random search and a simple genetic algorithm [489], well-known methods like CMA-ES [204] and Particle Swarm Optimization [252], and state-of-the-art methods like Learned and Discovered Evolution Strategies [280]. For each algorithm, we first tuned hyperparameters using Bayesian optimization via the Adaptive Experimentation (AX) platform [27]. We tuned for 50 trials on the ESC-10 dataset, a subset of ESC-50 [391]. Note that the hyperparameter tuning uses no privileged information and can easily be applied downstream on new prompt sets to maximize the performance.

The optimization procedure is specified in Algorithm 2.

5.3.3 Objective Function

We use LAION-CLAP [559] with an HTSAT-based audio encoder [83] and a RoBERTa-based text encoder [309]. We used the *audioset-best* checkpoint for general audio less than 10 seconds long.

The encoders process the audio data X_i^a in batches of size $\mathcal{B}$ where $\mathcal{B}$ corresponds to the optimizer's population size, along with a prompt p. Note that (X_i^a, p) is one particular pair of synthesized audio with input text prompt. We extract the audio embeddings $E_{\mathcal{B}}^a \in \mathbb{R}^{\mathcal{B}\times 512}$ and the text embeddings $E^p \in \mathbb{R}^{1\times 512}$ with the encoders and use them to calculate the similarity score between a batch of audio data and a specific prompt.

$$X_i^a = \mathcal{S}(\theta_i) \tag{5.1}$$

$$\theta^* = \arg\min_{\theta} \; -E_i^{\mathcal{S}(\theta_i)} E^{pT} \tag{5.2}$$

Equation (5.1) shows how the synthesizer $\mathcal{S}$ takes parameters θ_i and produces a sound

(in practice, this is done batched). Then Equation (5.2) formulates the optimization problem to optimize the similarity score between each audio in the batch and one given text prompt using their corresponding embeddings.

5.3.4 Evaluation Metrics

Since we propose a novel synthesis task without existing evaluation metrics, we devise a principled evaluation suite that allows us to quantitatively assess our contributions, in addition to qualitatively reviewing synthesized examples.

Classification Experiments To determine whether our generated sounds are more abstract than neural synthesis methods, we compared results on pretrained classifiers with sounds generated from their class labels. Lower scores can indicate a distribution shift from real audio, despite explicitly optimizing for similarity to the label. We complement with human listener ratings.

Without a perfect synthesis engine, any methods to generate sound will introduce a distribution shift from real audio. In our case, there is a deliberate domain shift to abstract audio. We evaluate on two well-known datasets. The first is ESC-50, a 50-class canonical environmental sound classification dataset [391]. The second is a subset of AudioSet [178]; the full ontology of classes is very large (over 500). We consider classes from "sounds of things" given that this category contains the most sub-classes and sub-selected the top 50 classes by number of annotations, removing duplicates or equivalent classes. We use a pretrained Audio Spectrogram Transformer (AST) model for AudioSet-50, and fine-tune an AST for ESC-50 classification [191]. When evaluating on AudioSet-50, we mask the remaining logits to effectively make it a 50-class classifier.

Synthesis Quality A significant benefit of our approach is synthesizing clean audio using signal generators while keeping attributes like sample rate flexible. We find synthesized sounds also often exaggerate aspects of the prompts, resulting in large variations in acoustic properties over time. Evaluating audio quality reference-free is challenging, so we examine acoustic features that correlate with these aspects (such as high-frequency content and spectral variation).

User Study We conduct a listening test with human evaluators. We ask them to classify sounds, rate their confidence, and rate sounds along a scale from re-

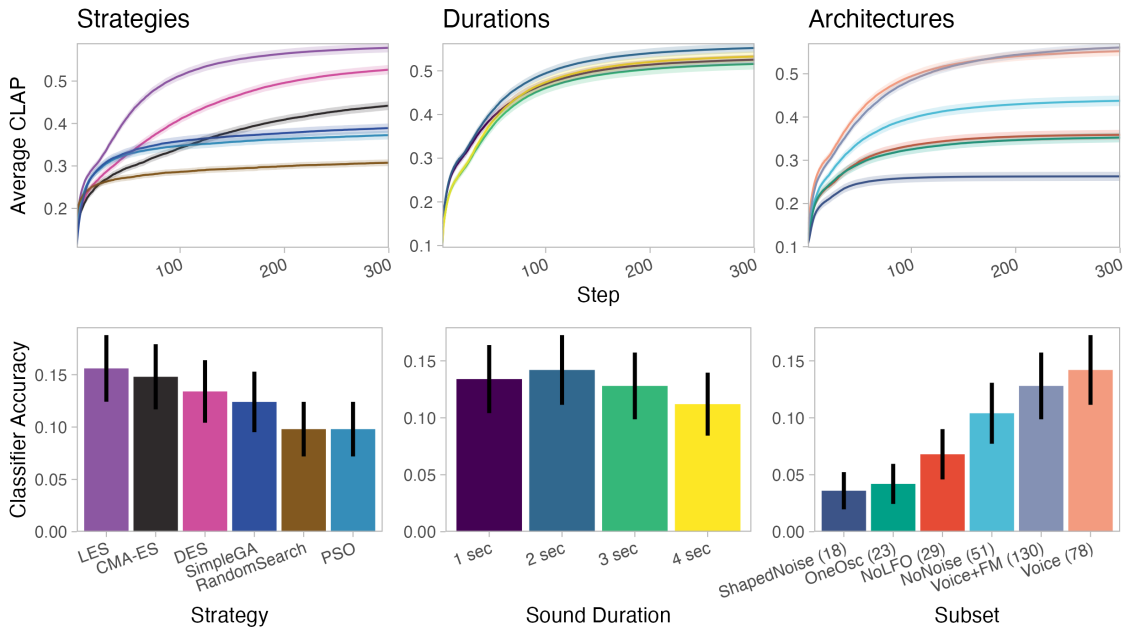

Figure 5-3: Results from our ablation study; all experiments are conducted with ESC-50. **(Top)** CLAP maximization curves, averaged across 10 iterations for each of the 50 prompts. Colored bands show 95% confidence intervals. **(Bottom)** Classification accuracy, with error bars showing 95% confidence intervals. Top and bottom plots share colors. **(Left)** Performance of different algorithms, with hyperparameters tuned on ESC-10. LES is strongest in both optimization and downstream classification. **(Center)** Different sound durations; we find 2 seconds to be strongest. **(Right)** Impact of synthesizer architecture, finding strongest results from the *Voice* model. Parameter counts are given in parenthesis, such as (78) for *Voice*.

alistic portrayal to artistic interpretation. This offers us the most direct signal of our abstraction-related goal. We share details on this study in the next subsection. We compared against the recent neural generation methods *AudioLDM* [301] and *AudioGen* [269].

From our 50-prompt subset of AudioSet [178] classes, we randomly selected 10 for this study. We used text embeddings of the labels with a facility location submodular optimization algorithm from the apricot package [451] to select a modest-sized semantically representative subset. Within each prompt, we randomly sampled two of 10 available *CTAG* sounds. The prompts were: *Truck air brake*, *Water tap*, *Train horn*, *Motorcycle*, *Microwave oven*, *Liquid slosh*, *Chainsaw*, *Airplane*, *Bicycle bell*, and *Machine gun*. For *AudioLDM* and *AudioGen*, we used their default parameters to

generate two sounds per prompt.

This study was determined to be exempt by our institution's IRB. Each participant rated 60 sounds (20 per method) in random order. To examine category-level recognition, participants were asked to select a category given a list of options and rate their confidence. To determine whether our generated sounds were perceived as (abstract) artistic interpretations, we posed the question: "Would you associate this sound more with a realistic portrayal or an artistic interpretation of the label that you selected?" with options on a scale from 1 (realistic portrayal) to 5 (artistic interpretation). We modeled participant responses with mixed-effects logistic and linear regression models and post-hoc contrasts.

5.4 Results

5.4.1 Ablation Studies

Figure 5-3 shows results from our ablation studies, including, from left to right, (1) optimization algorithms with tuned hyperparameters, (2) sound durations, and (3) synthesis architectures. Overall, we observe that the LES algorithm significantly outperforms our other options within the computation budget of 300 iterations (more than needed for several prompts). This experiment was conducted with 2-second long sounds, which we observe in the *Durations* experiment to yield a higher overall CLAP score and classification accuracy than 1, 3, or 4-second long generations. Finally, we see that the *Voice* architecture yields the best results, offering a balance of flexibility in its parameters and modular structure, as well as ease of optimization. However, we note that expanding to larger architectures like VoiceFM could be useful for future work to explore, with more work on the optimization strategy to obtain the best results.

Based on these results, we conduct all additional experiments discussed with the LES optimizer, 2-second sounds, and the *Voice* architecture. We conducted a full hyperparameter tuning run with 50 trials of all ESC-50 prompts to obtain the final optimization hyperparameters.

5.4.2 Qualitative Results

Examples

Figure 5-1 shows spectrograms of sounds corresponding to six text prompts. The "spray" shows bands of noisy bursts, reflecting the short, sharp sound of aerosol being expelled. The "bees buzzing" presents a band of low to high frequencies, encapsulating the vibrant hum of a bee. The "police car siren" is characterized by high-frequency oscillations that sharply rise and fall. The "machine gun" reveals rapid, staccato bursts of energy across a broad frequency range. The "train horn" displays horizontal bands across mid to high frequencies, illustrating the horn's fundamental tone and its partials. Lastly, the "chainsaw" spectrogram is dominated by intense, continuous mid-range frequencies, punctuated by peaks corresponding to the engine's roaring and cutting action.

Interpolation

In sound synthesis, interpretable parameters offer a unique opportunity for deeper insight. Our method provides a fixed set of parameters that possess this property—a salient distinction from contemporary models equipped with high-dimensional latent spaces. This interpretability extends to interpolation between parameters of distinct sounds, granting auditory access to intermediate acoustical transitions. In Figure 5-4, we present a systematic series of spectrograms between pairs of prompts: (1) "Spray" to "Machine gun", (2) "Train horn" to "Chainsaw", and (3) "Train wagon" to "Engine revving," with three intermediary steps linearly interpolated. This discernible gradation corroborates the capacity of our parameter space to retain congruence.

5.4.3 Classification Results

Results are shown in Table 5.2. On AudioSet-50, our results are higher than *AudioLDM*. On ESC-50, the classifier recognizes *CTAG*'s sounds the least, showcasing the distribution shift from its training on realistic sounds. We experimented with constructing concise and descriptive prompts from each sound class from both ESC-50 and AudioSet-50. We used GPT-4 [372] to automatically produce caption-style prompts. We also tried a simple template (i.e. "Sound of a/an ...") to compare. Table 5.2 also shows results for these template (*CTAG+T*) and caption-style prompts (*CTAG+C*). Introducing such strategies does not appear to greatly influence classifier identification. However, in a few cases, we observed the elaborated prompts helped

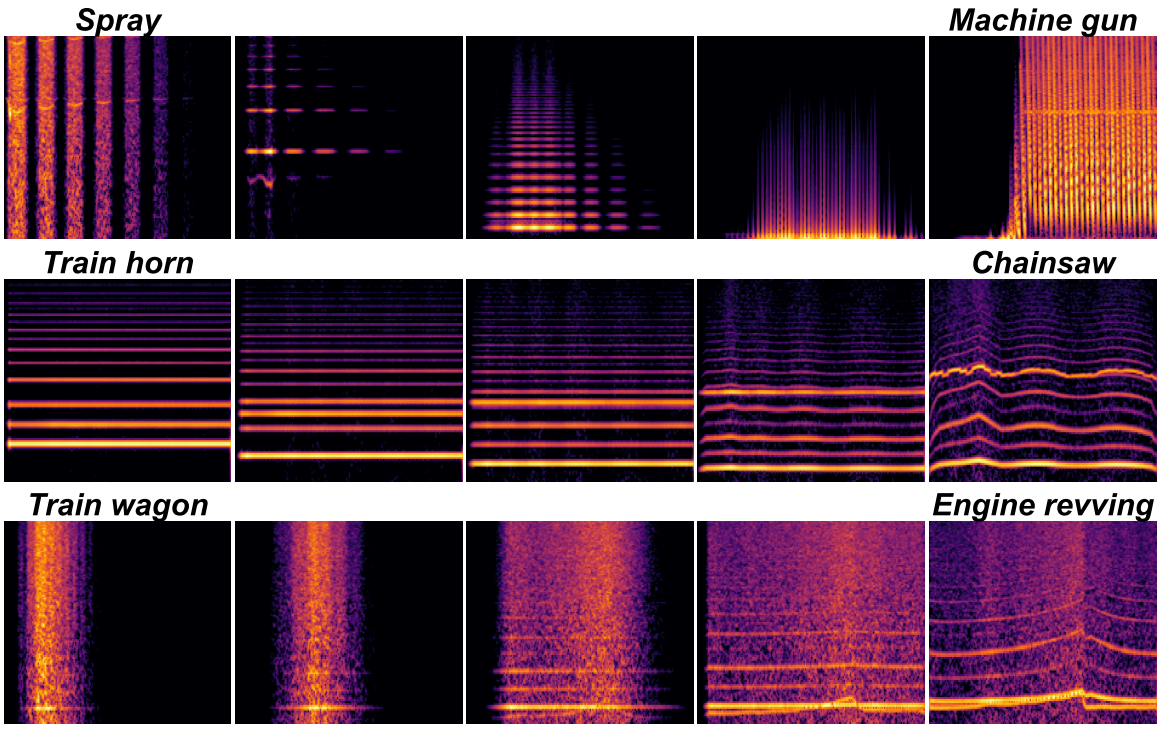

Figure 5-4: Spectrogram series as the result of linear interpolation of the synthesizer's parameters (1) from "Spray" (left) to "Machine gun" (right), (2) from "Train horn" to "Chainsaw", and (3) from "Train wagon" to "Engine revving". Each spectrogram in the sequence represents a step in the interpolation, highlighting the systematic shift in acoustic properties.

to produce qualitatively more accurate results. Overall, *CTAG* sounds are classified correctly significantly higher than chance, and competitively with *AudioLDM*.

5.4.4 Synthesis Quality and Variation

Evaluating the quality of generated examples is challenging for two reasons. First, we lack auditory references to compare against, as we generate from text directly and never use text-audio reference pairs. Most audio quality metrics are reference-based. Second, distance-based metrics such as FAD will likely be confounded by realism. *CTAG*'s sounds are high-quality in that they can be generated at high sample rates and are free of noise or artifacts owing to real-world recording environments or neural synthesis.

To evaluate, we use auditory descriptors (implemented using Essentia [53]) that are plausible correlates of these notions of quality, shown in Table 5.1. Spectral complexity

highlights the presence of more peaks, signaling diversity in the timbral components, while flux shows greater variation of timbre over time for *CTAG* compared with other methods. Following these, HFC (high-frequency content), spectral rolloff, and spectral centroid provide signals of "brightness" or high-frequency presence in the sounds. All of these results show our method's ability to introduce high-frequency content into generated sounds, likely in part due to the higher sample rate we use.

	AudioSet-50			**ESC-50**		
	AudioGen	*AudioLDM*	*CTAG*	*AudioGen*	*AudioLDM*	*CTAG*
Complexity	16.50	17.65	18.06	9.60	12.94	17.76
Flux	0.08	0.11	0.18	0.06	0.09	0.15
HFC	53.25	152.06	427.03	34.49	101.32	380.74
Rolloff	2,487.71	1,628.55	7,031.67	2,254.98	1,647.51	6,996.19
Centroid	1,629.95	1,096.16	4,139.99	1,512.55	1,108.42	4,227.08
Compression Ratio	6.42	7.09	9.51	6.46	7.58	9.57

Table 5.1: Comparison of spectral descriptors—complexity, flux, HFC, rolloff, centroid—and audio compression ratio, across ESC-50 and AudioSet-50. Results are grouped by the evaluation of three methods: *AudioGen*, *AudioLDM*, and our method, *CTAG*.

We also report compression ratio, under variable bit rate (VBR) MP3 compression (quality = 4). Interestingly, *CTAG* achieves a higher average compression ratio. VBR generally works by applying lower ratios to more perceptually complex input. Whether related to high-frequency content or other factors, this suggests *CTAG* sounds contain more perceptual redundancy or are perceptually "simpler".

Note that none of these measures are validated as perceptual metrics of audio quality, and we do not intend to use them as such. Rather, they help us to quantify the qualitative differences we observe between *CTAG*-synthesized sounds and other text-to-audio generation models' results.

5.4.5 User Study

We recruited 10 participants via Prolific at $12/h for a total of $53.33, resulting in a total of 600 observations per outcome variable (i.e. accuracy, confidence, and artistic interpretiveness). Table 5.3 contains the results, which show that our sounds were identified by listeners substantially more accurately than those from *AudioLDM* (odds ratio = 2.72, 95% CI [1.61, 4.58], $p < .0001$), and only slightly less than *AudioGen*

	AudioSet-50					ESC-50				
	AudioGen	*AudioLDM*	*CTAG*	*CTAG+T*	*CTAG+C*	*AudioGen*	*AudioLDM*	*CTAG*	*CTAG+T*	*CTAG+C*
Acc (Top-1)	51.6	17.4	26.2	25.2	23.6	54.0	23.0	16.4	11.4	13.8
Acc (Top-5)	77.4	44.2	45.2	52.2	51.6	71.8	49.4	30.4	26.4	31.0

Table 5.2: Top-1 and Top-5 classification accuracies (%) for pre-trained classifiers with AudioSet-50 and ESC-50. We evaluated both models on results collected using *AudioGen*, *AudioLDM*, and our method with just the class labels (*CTAG*), a simple template (i.e. "Sound of a ...") for each sound (*CTAG+T*) and finally using an LLM for prompt engineering (*CTAG+C*).

on average (odds ratio = 0.85, 95% CI [0.51, 1.42], $p = 1$). Interestingly, though the confidence ratings replicate the ordering of the accuracy results, respondents were significantly more confident rating *AudioGen* sounds, and reported similar, lower confidence levels for both *CTAG* and *AudioLDM*. This underscores the abstractness of *CTAG*'s sounds; despite being identified more correctly, they still create uncertainty.

	AudioGen	*AudioLDM*	*CTAG*
Accuracy	59.5	34.0	56.0
Confidence	3.48	2.95	2.99
Artistic Interpretation	2.32	2.90	3.54

Table 5.3: User study results for sounds from *AudioGen*, *AudioLDM*, and our method, *CTAG*. We report accuracy percentage and confidence (1–5) on label identification, and average rating of the artistic interpretiveness (1–5) of the sound. Overall, *CTAG* retains competitive identifiability while being perceived as more artistic.

Results also show *CTAG* sounds were perceived to be significantly more artistically interpretive than both *AudioGen* (contrast = 1.22, 95% CI [0.93, 1.51], $t(579) = 10.20$, $p < .0001$) and *AudioLDM* (contrast = 0.65, 95% CI [0.36, 0.93], $t(579) = 5.39$, $p < .0001$).

These findings highlight our approach's benefits in capturing artistic interpretation compared to both the existing approaches. All p-values are Bonferroni-adjusted. Full results for post-hoc contrasts are available in the Appendix.

5.4.6 Additional Analyses

In Appendix A.3.1 we provide additional analyses relating to generation time, CLAP scores, prompting strategies for the baseline models, user study results, and a visual-

ization of the parameter space of *CTAG*-generated sounds.

5.5 Limitations

Our method requires iterating for each prompt from random initialization, but techniques like semantic caching to initialize to similar prompts' parameters, predictive methods for prompt-to-parameter derivation, and a user interface extension for tweaking parameters are all potential extensions to make our method more useful in real-world settings. We also focus on brief, non-mixture sounds as these are what the synthesizer is suited to modeling. Future work could explore strategies to extend the duration and complexity of sounds that can be synthesized this way.

5.6 Conclusion

In this work, we proposed a method for text-to-audio generation that offers a fresh perspective on neural audio synthesis by using a virtual modular synthesizer. This approach emphasizes the meaningful abstraction of auditory phenomena, contrary to prevalent methods that prioritize acoustic realism. Our results position this approach as a distinctive tool in the field of audio synthesis, capable of both expanding the toolkit of novices and experts, and stimulating new directions in audio generation research.

6

Contrastive Learning from Synthetic Audio Doppelgängers

Driven by the goal of robust machine perception, we have begun to amass ever-larger datasets of stimuli across modalities. In audio, this often takes the form of recordings, sampled from sources like Freesound and YouTube. Such recordings are not in infinite supply, and the human perceptual apparatus seems capable of learning much more from much less. This chapter asks whether we might similarly endow machines with better auditory representations by focusing not on the quantity of training examples, but on the quality of variation between them.

Chapter 5 observes how simple synthesizers can, surprisingly, recover identifiable concepts related to real-world sound events. That is, the space of sounds implied by even a relatively small modular synthesizer's parameters can be quite rich as it relates to human creative goals. Here, we ask whether this richness might also benefit machines, in a representation learning problem similar to Chapter 3, albeit audio-only in this case. Rather than attempting to replicate the surface statistics of recorded sounds, we construct training data that encodes meaningful variation in the generative process itself. We show that this approach yields representations that generalize remarkably well to real-world applications, despite never seeing real audio during training. Like the dubbed movie pairs in Chapter 3, these synthetic pairs provide a window into how sounds can differ (in a "counterfactual" manner) while maintaining essential relationships. The approach also builds directly on the technical foundation laid in Chapters 4 and 5, leveraging the same synthesis framework and architecture to generate training data at scale.

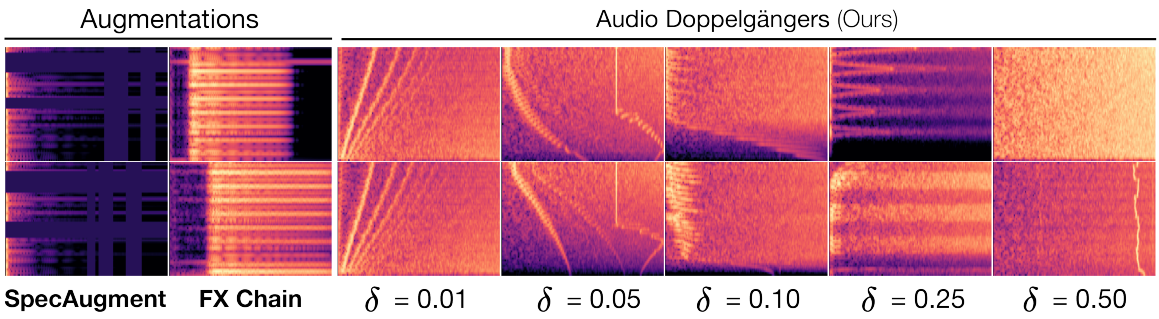

Figure 6-1: **(Left)** Standard data augmentation techniques for contrastive learning applied to audio spectrograms **(Right)** *Audio Doppelgängers*, our approach synthesizing sounds that are controllably different using perturbed synthesis parameters, shown for different factors δ. These sounds can vary in causally controllable ways beyond what data augmentations can achieve.

Abstract

Learning robust audio representations currently demands extensive datasets of real-world sound recordings. By applying artificial transformations to these recordings, models can learn to recognize similarities despite subtle variations through techniques like contrastive learning. However, these transformations are only approximations of the true diversity found in real-world sounds, which are generated by complex interactions of physical processes, from vocal cord vibrations to the resonance of musical instruments. We propose a solution to both the data scale and transformation limitations, leveraging synthetic audio. By randomly perturbing the parameters of a sound synthesizer, we generate *audio doppelgängers*—synthetic positive pairs with causally manipulated variations in timbre, pitch, and temporal envelopes. These variations, difficult to achieve through augmentations of existing audio, provide a rich source of contrastive information. Despite the shift to randomly generated synthetic data, our method produces strong representations, outperforming real data on several standard audio classification tasks. Notably, our approach is lightweight, requires no data storage, and has only a single hyperparameter, which we extensively analyze. We offer this method as a complement to existing strategies for contrastive learning in audio, using synthesized sounds to reduce the data burden on practitioners.

6.1 Introduction

"Noises have generally been thought of as indistinct, but this is not true."
— **Pierre Schaeffer**, *1986*

The success of modern machine learning algorithms for tasks like audio understanding often hinges on both the quality and quantity of available data. Self-supervised learning methods, like contrastive learning, have even been able to leverage unlabeled data, enabling more human-like learning from patterns without needing explicit supervision. However, human perceptual processing is remarkably robust beyond this: for example, the human auditory system can easily recognize sounds across a wide range of variations, such as changes in pitch, timbre, or background noise. Moreover, humans can quickly learn to recognize novel sounds that they encounter in their environment. Replicating this ability to learn from a diverse array of sounds—or "noises," as we might call them—could significantly enhance the efficiency, scalability, and adaptability of machine learning models.

Contrastive learning, which operates by recognizing similarities in the data among negative distractors, often relies on augmentations: transformations of input data that preserve content semantics. This method has been influential in audio representation learning, with specific implementations ranging from spectral masking to temporal jitter to cropping and other methods [4, 224, 326, 363, 439, 481, 534]. Data augmentations, though demonstrably useful, operate at the level of the observed data, not the underlying data-generating process as would be observed in real-world variation. They statistically alter data without directly manipulating the causal mechanisms that produced it, resulting in high correlation between augmented samples, as well as limited control and interpretability.

In our work, we propose a different strategy: using a synthesizer to overcome this barrier, in addition to providing the scalability required for modern pretraining regimes through virtually unlimited data synthesis. A synthesizer can be understood as a system where parameters (relating to psychophysical attributes like pitch, timbre, and loudness) causally influence the generated sound. Modifying these parameters allows us to intervene in the data-generating process in a controllable way to generate positive pairs that vary in terms of their underlying synthesis parameters. Unlike traditional data augmentation techniques, our method generates entirely synthetic

audio data from scratch. This approach allows us to control the underlying data-generating process directly, offering a perspective distinct from augmentation of real data.

We formulate an approach in which we randomly synthesize sounds, and then slightly perturb their parameters to generate positive pairs. We call these *audio doppelgängers* (examples in Figure 6-1); they share a resemblance but are in fact distinct enough to learn from the variation between them. In a way, this approach uses an artificial data source effectively consisting of random synthetic noises but more "natural" differences akin to variation in similar sounds; as Pierre Schaeffer put it, noises are not indistinct. Through a comprehensive set of experiments, we show that models trained this way can yield strong performance on a wide range of downstream tasks, competitive with real audio.

Overall, this work contributes:

1. An approach to synthesizing paired audio examples with a continuously controllable degree of dissimilarity, specified by a simple and interpretable hyperparameter δ.

2. The first study, to our knowledge, of synthetic data methods for audio representation learning.

3. Comprehensive experiments in which we train and compare over 20 model variants across 8 downstream tasks to provide evidence that training with our approach can yield strong results on a wide range of audio processing tasks.

4. An analysis of how these synthetic datasets differ from realistic audio datasets in terms of their auditory features, and how this might contribute to learning effective representations.

6.2 Related Work

6.2.1 Learning from Synthetic Data

Synthetic data, artificially generated information rather than collected from real-world sources, has emerged as a valuable tool for learning across various domains [273, 307, 341, 467]. By addressing data scarcity, privacy concerns [131, 515], or removing

biases [405, 496], synthetic data offers a promising avenue to complement scarce [311] real-world data and further drive progress in machine learning research.

Audio presents unique challenges due to the complexity of waveforms and temporal dependencies. Synthetic data has found applications in subareas like speech recognition [151, 167, 220, 284, 437, 438] leveraging text-to-speech systems for detecting unspoken punctuation [476], recognizing low-resource languages [36], increasing acoustic diversity [88] or detecting out-of-vocabulary words [578]. However, non-speech audio domains can be highly diverse, requiring more complex approaches to data synthesis. In this domain, synthetic data has been used for specific tasks like timbre-text alignment [237] and vocoding [540]. The partially synthetic NSynth [134] dataset has also been used for pitch estimation and instrument classification. In our work, we tackle the general audio domain, proposing a synthetic data approach that can produce diverse sounds for general-purpose audio representation learning.

In computer vision, synthetic data is more popular and has been employed in different tasks to improve performance [87, 129, 229, 337, 416, 434, 457, 528]. While initially focused on using graphics engines to generate photorealistic scenes, recent work has investigated sampling synthetic data from deep generative models [44, 213, 233, 234, 293, 412, 466, 507, 508, 510, 513, 564, 575]. However, these models aim to produce realistic images and still depend on real image datasets for training or synthesis. Thus, recent work has pushed away from realism, generating synthetic data such as fractals [245], or through other procedural noise models [32, 33] to use as training data for visual representation learners. In our work, we also abandon realism and leverage randomly generated synthetic sounds to learn audio representations for downstream tasks.

6.2.2 Contrastive Learning

A common strategy for learning from unlabeled data is *contrastive* learning. In this technique, we seek representations that are *invariant* to minor differences, i.e. they encode a space in which similar objects are closer together, and dissimilar objects are further. A classic strategy for this is to use data *augmentations*, transformations which noticeably alter a datapoint (for example, randomly cropping an image) without changing its essential content (e.g. what the image is of, such as a cat). These transformed versions then become a *positive pair*, while other examples (e.g. an image

of a dog) become *negatives*. In audio, contrastive learning has been used extensively to produce high-quality representations for downstream tasks [4, 158, 411, 439, 534]. Our approach differs fundamentally from data augmentation strategies commonly used in contrastive learning. Instead of applying transformations to existing audio samples, we generate synthetic audio pairs by perturbing synthesizer parameters, creating positive pairs with causal variations that are difficult to achieve through augmentations. This represents a novel application of synthetic data in the context of general-purpose audio representation learning.

6.2.3 Sound Synthesis

The toolkit of sound synthesis has evolved to include a variety of hardware and software [336, 395, 502]. Synthesizers, abstractions of sound synthesis and processing methods often designed to act as musical instruments, are key to this: they expose control parameters that let sound designers guide them to produce desirable sounds for music, film, and many other applications. Accelerated synthesizers [89, 516] have recently allowed much faster-than-realtime sound generation, offering the ability to iteratively tweak parameters to reconstruct sounds [203, 461] and even match textual descriptions [91]. Such approaches highlight the practical utility of synthesizers: lightweight architectures controlled by a limited number of interpretable parameters are capable of producing a diverse array of sounds, often corresponding to well-known categories and concepts (e.g. the sound of waves can often be modeled with time-varying filtered noise). In our work, we leverage SynthAX [89], to rapidly produce diverse training data with controllable similarity between examples.

6.3 Methods

6.3.1 Data Generation

Our data generation pipeline uses virtual modular synthesizers implemented by SynthAX [89] in JAX. By default, we use the *Voice* synthesizer architecture [516], which can generate perceptually diverse sounds. Our synthesizer consists of several common modules (with parameter-counts in parenthesis):

- Keyboard (2x): Controls the sound's fundamental frequency (f_0) and duration.
- Low-Frequency Oscillators (LFOs; 8x each): Two LFOs modulate various aspects of the sound, each with parameters for frequency, modulation depth, initial

phase, and amplitude weights across waveforms.

- ADSR Envelopes (5x each): Six envelopes shape the amplitude and modulation signals, each defined by attack, decay, sustain, release, and curvature (α).
- Voltage-Controlled Oscillators (VCOs): Includes a sine VCO (3x) with tuning, modulation depth, initial phase, and a square-saw VCO (4x) adding waveform shape.
- Noise Generator: Provides broadband noise without additional parameters.
- Modulations (20x): Weight matrix controlling how modulation sources affect destinations.
- Audio Mixer (3x): Combines outputs of oscillators and noise generator.

In total, the *Voice* synthesizer has 78 parameters. Perturbing these parameters allows us to generate a wide variety of sounds with controlled variations, such as slightly lower or higher pitch, a slightly longer onset or release, or a little more or less noise. In our experiments, we investigate two further architectures: *VoiceFM* has 130 parameters and includes a frequency modulation (FM) operator, and *ParametricSynth* has 2 sine and 2 square-saw oscillators, 1 sine FM and 1 square-saw FM operator, 340 in total. Varying the architecture allows us to investigate whether architectural complexity could affect the quality of representations learned. We generate 1-second sounds by default, for compatibility with most encoders (e.g. VGGish [211]). However, this practice can be extended to longer sounds.

Synthesis perturbation factor (δ) A key contribution of our work is synthesizing paired positive samples that sound alike, but are dissimilar due to their synthesis parameters and not only post-hoc effects (e.g. augmentations). This draws on the canonical definition from contrastive learning of positives that are sampled from the same *latent class* [446].

For a single positive pair, we first sample a parameter vector uniformly randomly $\theta \in [0,1]^{m_S} \sim \mathcal{U}(0,1)$ from the normalized synthesis parameter space, where m_S is the number of control parameters in the given synthesizer. Then, we independently sample two isotropic Gaussian noise vectors $\mathbf{z_1}, \mathbf{z_2} \sim \mathcal{N}(0, \mathbf{I}_{m_S \times m_S})$. We define a parameter δ that scales this noise, and then produce two perturbed parameter vectors $\theta_i = \theta + \delta \mathbf{z_i}\ \forall_i \in \{1,2\}$. From these, we clip values back into $[0,1]$ to synthesize two corresponding sounds which serve as positives in the contrastive learning setup.

In principle, δ controls the distance between the positive pairs and therefore the hard-

ness of the contrastive learning task. Practically, we expect there to be a sweet-spot for δ, considering prior work on mutual information and redundancy in contrastive learning problems [506, 509] as with very high δ, the parameter vectors may become dominated by noise, resulting in difficulty effectively aligning their representations. Given this, we extensively study the effect of δ on downstream results.

6.3.2 Real Data

To compare to real audio data, we use sounds from VGGSound [82], a well-known dataset taken from YouTube videos (we only use audio). We use a random sample of 100,000 10-second files and select a random 1-second segment from each file at each iteration to augment. This allows us to fairly compare to our synthetic sounds by keeping duration constant, while still sampling from a variety of real sounds by randomizing the 1-second segments. Though VGGSound has labels included, we do not use them in training these models to keep the self-supervised constraint. Note that VGGSound is currently one of the largest publicly released audio datasets for pretraining, unlike AudioSet (which only releases URLs, not the content itself).

6.3.3 Preprocessing, Data Augmentations, and Audio Encoder

In our experiments, we use VGGish frontend representations [211]. We resample audio to 16kHz and obtain mel spectrograms with 64 mel bands and 96 time steps. We use a chain of effects as augmentations (implemented in torch-audiomentations[1]): a high-pass filter (cutoff frequency range 20–800Hz), a low-pass filter (1.2–8kHz), pitch shift (-2 to 2 semitones), time shift (-25% to 25%, rollover enabled), and finally reverberation for which we sample randomly from a set of impulse responses. All augmentations are applied with probability 0.5. We found that this yielded far stronger results than SpecAugment [382], and so we use this as a comparison point in all our experiments. More details on the augmentation are given in Appendix A.4.2. We also test temporal jitter, wherein different 1-second segments are sampled from within the same source clip and treated as positives [439, 481]. Our audio encoder is a ResNet18 [207], where we replace the initial layer with a 1-channel convolution to account for the effectively 1-channel spectrogram.

[1] https://github.com/asteroid-team/torch-audiomentations

6.3.4 Contrastive Learning

We train for 200 epochs, generating (or sampling) 100,000 sounds per epoch, with a 90%-10% train-validation split. We use a batch size of 768 per GPU with two V100s. The training uses the alignment and uniformity objectives [538] used in prior work on learning with synthetic data [33]. We adopt the default parameters for these: $\text{unif}_t = 2$, $\text{align}_\alpha = 2$, and equal weights $\lambda_1 = \lambda_2 = 1$ for both terms. Following this work, we use stochastic gradient descent for optimization, with a maximum learning rate of 0.72 (calculated as $0.12 \times \frac{\text{total batch size}}{256}$) and weight decay 10^{-6}. The learning rate follows a multi-step schedule with $\gamma = 0.1$, and milestones at 77.5%, 85%, and 92.5% of the total learning epochs. Detailed steps are provided in Algorithm 3. Training with our synthetic data takes approx. 1-2 hours, as the data is generated on the fly in batches, whereas using on-disk datasets with effect chain augmentations can extend training time up to 6-8+ hours.

6.3.5 Evaluation Tasks

To obtain a broad picture of the quality of our learned representations, we conduct experiments on a range of audio classification tasks from the HEAR [517] and ARCH [275] benchmarks. We use evaluation tasks that focus on general audio understanding, rather than tasks that are highly specialized or domain-specific (e.g., tasks exclusively related to speech or music). Our method aims to learn general-purpose audio representations from synthetic data; therefore, we implement tasks which encompass a broad range of everyday sounds. This aligns with our goal of demonstrating the effectiveness of our representations in diverse real-world scenarios.

These tasks cover a wide range of capabilities including sound classification tasks like ESC-50 [393], FSD-50k [157], and UrbanSound8K [441], vocal affect tasks with and without speech like VIVAE [215] and CREMA-D [75], musical pitch recognition via NSynth Pitch (5h) [134], vocal sound imitation recognition using Vocal Imitations [255], and LibriCount [486] for a "cocktail party" style speaker count estimation task. We conduct linear probing experiments using the Adam optimizer for the benchmark-specified epochs with the default learning rate of 0.001 and a batch size of 32.

Algorithm 3 Our contrastive learning procedure with *audio doppelgängers*. In the training loop, we drop the batch index i for simplicity.

Require: Batch size k
Require: Perturbation factor δ
Require: Virtual synthesizer S with m_S parameters
Require: Embedding model M with embedding size m_M (512 in our case)
Require: Total number of training batches N_{batches}
Require: $\ell_{\text{unif}}(\mathbf{X} \in [0,1]^{k \times m_M}) \leftarrow \log\left[\frac{1}{k^2}\exp\left(-t\sum_{j=1}^{k}\sum_{l=1}^{k}\|\mathbf{X}[j]-\mathbf{X}[l]\|_2^2\right)\right]$ where $t = 2$

for $i = 1$ **to** N_{batches} **do**
  $\mathbf{\Theta} \in [0,1]^{k \times m_S} \sim \mathcal{U}(0,1)$ {Random batch of parameters}
  $\mathbf{Z}_1, \mathbf{Z}_2 \in \mathbb{R}^{k \times m_S} \sim \mathcal{N}(0, \mathbf{I})$ {Isotropic Gaussian perturbation noise}
  $\hat{\mathbf{\Theta}}_1 \leftarrow \max(0, \min(\mathbf{\Theta} + \delta\mathbf{Z}_1, 1))$ {Clipped perturbed parameters}
  $\hat{\mathbf{\Theta}}_2 \leftarrow \max(0, \min(\mathbf{\Theta} + \delta\mathbf{Z}_2, 1))$
  $\mathbf{A}_1 \leftarrow S(\hat{\mathbf{\Theta}}_1)$, $\mathbf{A}_2 \leftarrow S(\hat{\mathbf{\Theta}}_2)$ {Synthesize audio from parameters}
  $\mathbf{E}_1 \leftarrow M(\mathbf{A_1})$, $\mathbf{E}_2 \leftarrow M(\mathbf{A_2})$ {Embedding from model}
  $\mathcal{L}_{align} \leftarrow \frac{1}{k}\sum_{j=1}^{k}\|\mathbf{E}_1[j]-\mathbf{E}_2[j]\|_2^{\alpha}$ where $\alpha = 2$ {Alignment cost}
  $\mathcal{L}_{uniform} \leftarrow \frac{1}{2}[\ell_{\text{unif}}(\mathbf{E}_1) + \ell_{\text{unif}}(\mathbf{E}_2)]$ {Uniformity cost}
  $\mathcal{L}_{total} \leftarrow \lambda_1\mathcal{L}_{align} + \lambda_2\mathcal{L}_{uniform}$ {By default, we set $\lambda_1 = \lambda_2 = 1$}
  Update model M using $\mathcal{L}_{total}$
end for

6.4 Results

6.4.1 Benchmark Results

In Table 6.1, we show results across 8 tasks. The top section features external baselines from the HEAR [517] leaderboard and ARCH [275] benchmark results, first the strongest overall and then only self-supervised. It also includes results from MS-CLAP [132] linear probing experiments, GURA [558] (strongest overall model on HEAR), and finally the original ResNet18 trained on VGGSound (supervised) [82]. Note that HEAR leaderboard results may use MLP probes, whereas ours are linear. We add additional internal baselines, including random weights, synthetic data trained without δ but with augmentations, and variants of ResNet18 we trained on VGGSound (with augmentations, and alternately with temporal jitter). Finally, we include a

selection of our results; the best overall score we achieve using our synthetic approach (first row), followed by the best-performing model trained on data from each of the synthesizer architectures (including *Voice* with augmentations). In Appendix A.4.1, we provide a full set of results from all model variants: all synthetic datasets for all values of δ, and further baselines less performant than those we present here.

Overall, our best scores uniformly outperform training on VGGSound with augmentations, and outperform training with temporal jitter (the strongest internal baseline) in 6/8 cases. In some cases, these results are also competitive with strong baselines, such as beating the supervised ResNet18 result on 3/8 tasks, CLAP on 2/5, and GURA on 1/6. Additionally, adding further augmentations to our *audio doppelgänger*-based training does not seem to hold significant benefits, despite being highly beneficial when training with synthetic sounds with no δ, suggesting the δ-based perturbations are already sufficiently strong. All this is accomplished without these models seeing any real sounds during pretraining. Finally, *Voice* with $\delta = 0.25$ is the strongest synthetic-trained model overall, being the top performer on 5/8 tasks, but we note that there is some inter-task variability in the best synthesizer and delta.

6.4.2 Characterizing the Data Distribution

Here, we focus on understanding the distribution of synthetic sounds and how they differ from natural sound properties. We primarily use our alternate training set, VGGSound [82], for these measures. Unless specified otherwise, we use a randomly sampled (for VGGSound) or generated (for synthetic) set of 1000 sounds for each given dataset used for these characterizations. Our goal is to help understand what properties of the synthetic data make it useful for representation learning, given its strong performance.

Embedding Similarity

First, we look at the distribution of synthetic sound pairs and establish that δ meaningfully controls (a proxy for) perceptual or semantic dissimilarity. We operationalize this using LAION-CLAP embeddings [559], since they are trained on a large variety of sounds with semantic descriptions associated. Figure 6-2A shows how the average cosine similarity decreases monotonically with increasing δ for all 3 synthesizers. Figure 6-2B provides a view of how δ affects the geometry of the embedding space. Here, we plot the first two principal components of the CLAP embeddings along with

Data/Model	ESC	US8K	VIV	NSyn	C-D	FSD	VI	LCount
External Baselines								
HEAR/ARCH Top	96.65	79.09	44.28	87.80	75.21	65.48	22.69	78.53
HEAR/ARCH SSL	80.50	79.09	44.28	52.40	75.21	50.88	18.48	78.53
MS-CLAP Linear	89.95	82.29	–	–	23.15	50.24	–	54.51
GURA (HEAR)	74.35	–	–	38.20	75.21	41.32	18.48	68.34
VGGSound Sup.	87.45	77.57	39.38	43.80	54.36	43.76	14.06	56.10
Internal Baselines								
Random Init.	22.45	55.03	33.81	36.20	38.91	9.03	2.43	44.91
Voice (Ours, No-δ, Aug.)	48.65	59.46	36.31	32.80	46.32	16.88	7.12	47.64
VGGSound SSL (Aug.)	48.85	61.91	32.67	39.60	47.86	19.63	6.03	53.46
VGGSound SSL (Jitter)	52.95	63.82	38.12	14.20	**50.03**	24.02	3.43	**69.77**
Audio Doppelgängers (Ours)								
Best Synthetic	**58.90**	**66.71**	**39.45**	**44.40**	**48.43**	**24.12**	**9.15**	**58.60**
Voice ($\delta = 0.25$)	**58.90**	**66.71**	**39.45**	32.20	48.24	**24.12**	**9.15**	52.95
Voice ($\delta = 0.25$, Aug.)	58.75	65.01	34.81	**44.40**	46.17	21.76	8.54	50.70
VoiceFM ($\delta = 0.25$)	57.20	65.11	38.48	35.20	48.43	22.15	6.96	54.00
Parametric ($\delta = 0.25$)	50.55	62.83	37.91	37.60	46.77	18.68	5.70	54.72

Table 6.1: Evaluation results on a suite of tasks including (from left to right) ESC-50 [393], UrbanSound8k [441], VIVAE [215], NSynth Pitch 5h [134], CREMA-D [75], FSD50k [157], Vocal Imitation [255], and LibriCount [486]. For internal baselines, we only bold tasks where the baseline beats the best synthetically trained result. Results for all synthetic variants are in Appendix A.4.1.

the path length for each positive pair of synthesized samples from a *Voice* synthesizer. As δ increases, the path lengths increase and overlap more resulting in less clear separation of positive pairs from negatives. We view this as a signal that we can effectively control the hardness of the contrastive task using δ, the perturbation factor.

Similarities and Differences from Real Data

Next, we compare the synthetic data distribution to VGGSound [82] data. Figure 6-3A compares a selection of features' distributions between several dataset variants. For synthetic datasets, we have *Voice*, *VoiceFM*, and *ParametricSynth* variants. For real datasets, we have VGGSound. We first compare to randomly sampled 1-second chunks. Here, the synthetic sounds match several feature distributions well, such as Inharmonicity [385], Odd-to-Even Harmonic Ratio [329], Pitch Salience [419], and,

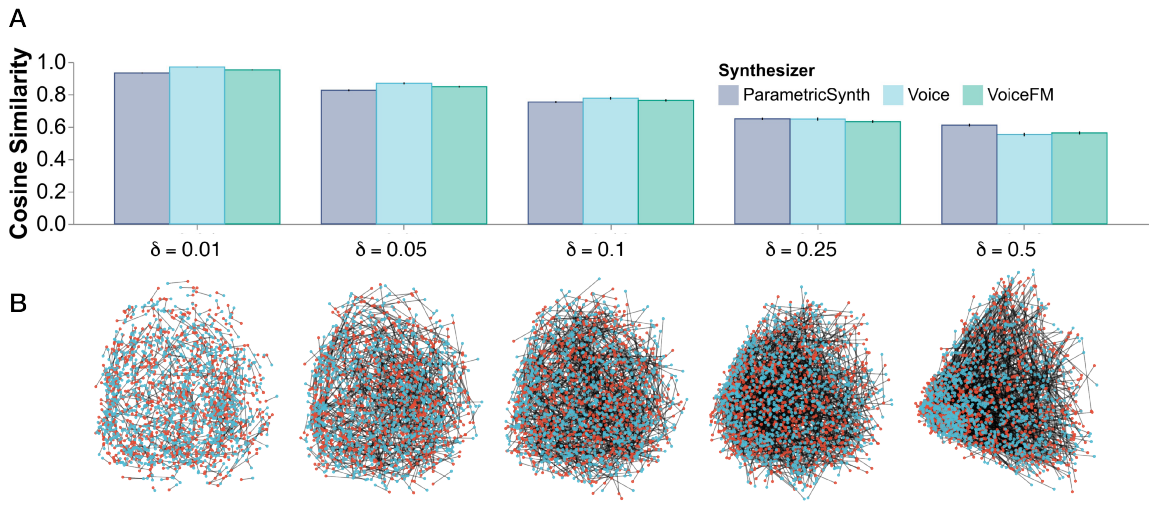

Figure 6-2: **(A: Top)** Average CLAP [559] embedding cosine similarity between positive pairs for different architectures and different values of δ. **(B: Bottom)** PCA of CLAP embeddings for sounds generated with the *Voice* architecture, with line segments showing distances between paired examples. Red and blue points are paired positive instances. Across both plots, as δ increases, the positive pairs systematically become more perceptually dissimilar (via the CLAP embedding proxy).

to a lesser extent, Spectral Flatness [385]. However, the synthetic sounds have higher Spectral Flux [521] and Complexity [285]. Note that *ParametricSynth* also has lower pitch salience. We believe this is due to its larger mixture of sound generators which reduce salience of particular pitches.

Based on these results, we hypothesize that one potential reason the synthetic sounds could be useful for training is the informativeness of the samples. The larger amount of spectral change and higher complexity in terms of peaks could expose the model to more different kinds of sounds rapidly. To try to match these attributes, we produce mixtures of VGGSound, since mixed sounds may have more peaks and variation than individual samples. In VGGSound-Mix 5s, we take 5 arbitrary seconds from each sound and layer them into a 1-second sample. In VGGSound-Mix 10s, we do the same with all 10 seconds available. We show (Figure 6-3) that these get closer to the synthetic distributions on these features, without deviating on other features. These data distributions allow us to assess whether the benefits of synthetic data are largely driven by the change and informativeness of the signals. In Appendix A.4.1, we present results from models trained on these mixture distributions, and obtained mixed results, suggesting other factors of the synthesized sounds may also be important beyond

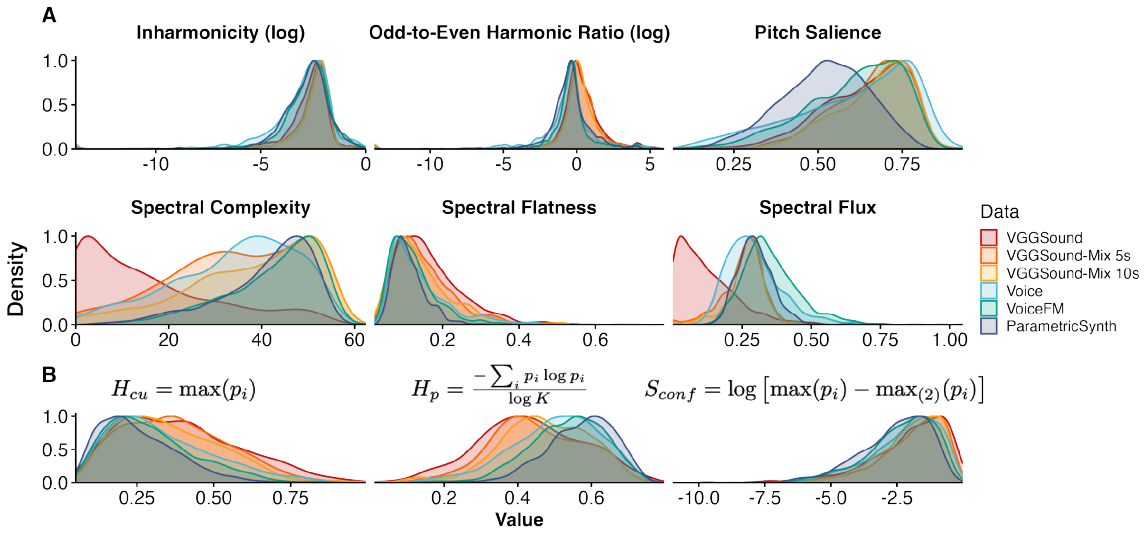

Figure 6-3: Comparisons of synthetic and real sound data (VGGSound [82]) on **(A: Top)** spectral features and **(B: Bottom)** causal uncertainty. Spectral features of synthetic sounds partially replicate real sounds, but exhibit differences in complexity and flux. Synthetic sounds are also more causally ambiguous, indicating a distribution shift. Using dense mixtures of real sounds partially closes these gaps, suggesting the synthetic sounds are different in part due to their density of auditory information.

this.

Causal Uncertainty

We also consider causal uncertainty [10, 28, 54], a factor that we intuitively expect to be different for the synthetic sounds. Helmholtz famously discussed perception in terms of unconscious causal inference from sensory input [210], but the synthetic sounds have no physical causes and do not come from well-understood categories. In Figure 6-3B, we plot 3 proxies for causal uncertainty derived from probabilities of an AST classifier [190] trained on AudioSet. We use the formulation from prior work of H_{cu}, the maximum predicted probability [10, 54]. We also propose two simple metrics to corroborate this: H_p the (normalized) entropy of the output probability distribution, and a confidence score S_{conf}, the difference in probability between the most and second-most probable classes (log-scaled for the plot). Across all, the synthetic sounds are more causally uncertain than the real sounds. However, as with the spectral feature distributions, using mixtures of VGGSound [82] clips moves the real distribution slightly closer to the synthetic distribution per H_{cu} and H_p. We

speculate that exposure to more causally uncertain sounds might be subtly helpful for representation learning; for example, they may contain diverse features that aid generalization to more ambiguous sounds present in downstream tasks. We characterize this as another important distributional difference between the synthetic sounds and realistic sounds from datasets such as VGGSound.

Similarity to Target Distributions

Another lens we can use to understand the effectiveness of training on synthetic data is in terms of the distribution of sounds in the target downstream tasks. A common metric to compare sound distributions is the Fréchet Audio Distance (FAD) [254]. For simplicity, we use the canonical formulation based on VGGish embeddings, though there are some limitations of this [196, 495], and we use either the validation sets or first multi-fold splits of the target task audio. Table 6.2 shows that for ESC-50 [393], VGGSound is closer in distribution to the target sounds, likely due to ESC-50's focus on environmental sounds. For all other tasks, however, the synthetic sounds achieve a lower FAD, suggesting they may better capture task-relevant features for these tasks' sounds. This finding echoes a study of MMDs in torchsynth [516], where the *Voice* architecture shows higher-than-expected similarity to FSD50k sounds. We hypothesize that the synthetic training allows the model to see a wide variety of spectral behavior rapidly, in a way that supports an array of tasks.

Dataset	ESC-50	FSD50k	LibriCount	NSynth	CREMA-D	Vocal Imitation
Voice	17.39	**13.37**	**16.67**	**12.83**	**18.55**	**11.64**
VoiceFM	18.48	15.91	17.67	14.49	21.24	13.66
ParametricSynth	18.75	19.44	21.04	17.42	25.33	17.32
VGGSound	**6.71**	25.33	29.09	27.67	33.83	27.75
VGGSound-Mix 5s	8.81	26.17	30.02	28.70	34.35	29.05
VGGSound-Mix 10s	9.30	26.09	30.15	28.88	34.06	29.16

Table 6.2: FAD [254] scores between different synthetic/real datasets and target downstream task data distributions, computed using either validation sets or the first fold (for multi-fold datasets). For 5/6 tasks, *Voice* achieves the lowest FAD despite containing synthetic sounds. On ESC-50, however, the VGGSound distribution appears to be closest.

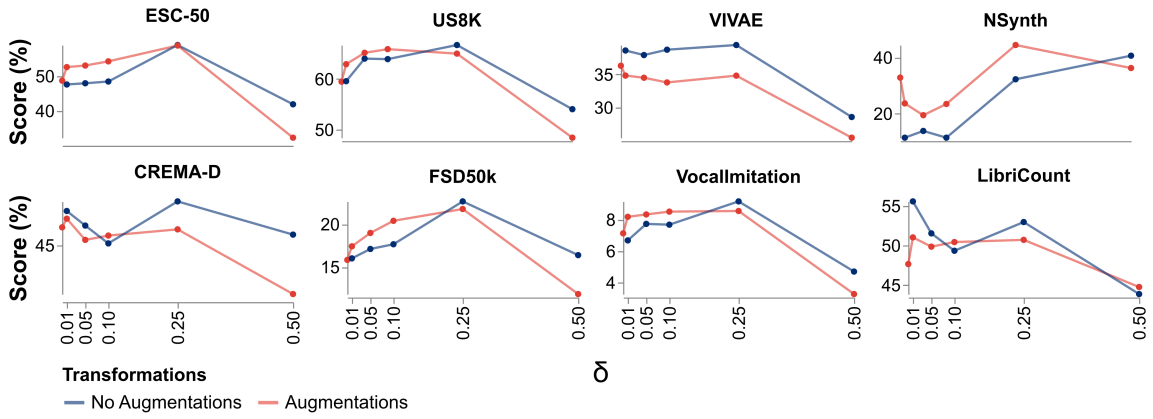

Figure 6-4: Scores with the *Voice* architecture and different values of δ for evaluation tasks in Table 6.1 with and without augmentations. $\delta = 0.25$ tends to give the best results overall.

6.4.3 Ablations and Sensitivity Analysis

In Figure 6-4, we study the effect of the perturbation factor δ on downstream task performance across all tasks for the strongest (*Voice*) architecture with and without additional (FX chain based) augmentations. Overall, we found in early experiments that $\delta = 0.25$ gives the best results. This is observed on the full suite of tasks (with notable exceptions like NSynth without augmentations, and LibriCount overall). In Appendix A.4.1, we discuss why this value might be strongest from the perspective of alignment and uniformity results.

6.5 Limitations

Our study demonstrates the efficacy of our approach using established architectures like ResNets, balancing computational efficiency with the goal of producing generalizable results. While we focused on these architectures, our findings lay the groundwork for future investigations with larger encoders such as AST [190]; this is a straightforward extension. Our research also focused on a clear comparison between synthetic and real data, which allowed us to rigorously evaluate our method's effectiveness. The potential for hybrid approaches, combining synthetic and real data, has a wide range of possibilities in mixing strategies and fine-tuning techniques. These can all be explored without changing our method itself.

The isometric Gaussian noise perturbation proved highly effective, despite its sim-

plicity, since it changes identifiable attributes (e.g. pitch) subtly. This success points to the robustness of our method, while also highlighting opportunities for more sophisticated perturbation strategies to further enhance it. Future work could explore anisotropic perturbations that account for parameter relationships. Adaptive or learned perturbation strategies could also offer significant advancements. Additionally, our evaluation centered on widely used classification benchmarks to create a foundational assessment of the method's performance. Expanding this evaluation to include metrics such as representation disentanglement could offer additional insights into the quality and utility of the synthetic data.

On a broader note, we believe it's important to examine synthetic data-generating procedures for possible biases, similar to the scrutiny applied to real datasets. Though we think this procedure can mitigate some of the biases in real datasets, different synthesizer architectures, values of δ, and other decisions might inadvertently produce performance gaps for different tasks, applications, and downstream populations of use. We evaluated on a wide range of tasks in part to explore this possibility, but further evaluations would be helpful to assess these impacts.

6.6 Conclusion

Further improvements in auditory understanding depend greatly on the data underlying new models. In this work, we examined the value of synthetic data for learning representations of sound. We presented a method that perturbs the parameters of random synthetic sounds to generate *audio doppelgängers*, distinct yet similar sounds that provide a strong signal for contrastive learning. Through a comprehensive set of experiments, we showed how this approach can yield strong results on a wide range of tasks. We will release our code and models to enable the community to experiment with synthetic data sources for audio understanding, and hope this approach will help expand the machine learning toolkit for audio processing.

7

Articulatory Synthesis of Speech and Diverse Vocal Sounds via Optimization

The human voice represents perhaps the most sophisticated and expressive sound-generating system we know. Unlike the abstract synthesis explored in Chapters 4 to 6, vocal sound production is constrained by precise physical mechanisms. Human speakers control such mechanisms with remarkable fluency to produce a tremendous range of sounds. This chapter investigates whether we can computationally reconstruct these physical processes from real vocal samples, not to replace human vocalization, but to better understand and potentially extend it.

We introduce VocalTrax, a system that reconstructs speech and vocal sounds by optimizing the parameters of a physical vocal tract model. Unlike black-box approaches to voice synthesis, this physically-motivated model provides interpretable control. Like Chapter 4, this chapter implements an accelerated computational synthesizer, albeit here for the voice. Like Chapter 5, this work produces desired outputs by inference-time optimization, although here from reference audio rather than text, and using gradient-based optimization rather than evolutionary strategies. VocalTrax extends our exploration of interpretable synthesis approaches to the domain of voice. In principle, this can be embedded into the kinds of generation and representation learning approaches described in Chapters 5 and 6, extending the toolkit of audio machine learning techniques.

Abstract

Articulatory synthesis seeks to replicate the human voice by modeling the physics of the vocal apparatus, offering interpretable and controllable speech production.

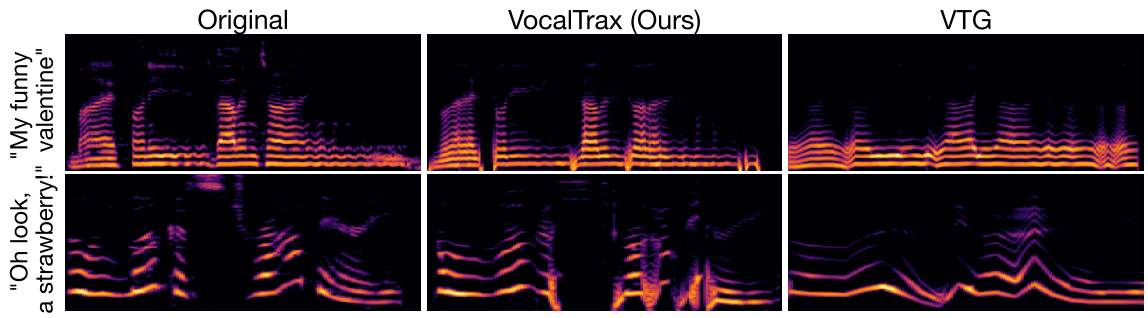

Figure 7-1: Spectrograms showing two target vocalizations with reconstruction via our approach (*VocalTrax*) and prior work [490]. (**Top**) a clip of Frank Sinatra singing *My Funny Valentine*. (**Bottom**) original speech audio from the popular "Oh Look, A Strawberry" meme.

However, such methods often require careful hand-tuning to invert acoustic signals to their articulatory parameters. We present *VocalTrax*, a method which performs this inversion automatically via optimizing an accelerated vocal tract model implementation. Experiments on diverse vocal datasets show significant improvements over existing methods in out-of-domain speech reconstruction, while also revealing persistent challenges in matching natural voice quality.

7.1 Introduction

The human voice presents a formidable challenge for computational modeling, with its complex physiology and acoustics. Articulatory speech synthesis [420], which aims to replicate this complexity by simulating the vocal tract's physical properties, has long been a prized goal of speech technology since it results in interpretable and controllable synthesis. While traditional approaches can be laboriously programmed to construct longer segments [13], they struggle to match the richness and variability of natural speech and even text-to-speech models [270]. As such, there is a need for acoustic-to-articulatory inversion methods that generalize to realistic speech, song, and other vocalizations.

This chapter introduces *VocalTrax*, an optimization-based method for matching arbitrary vocal signals to articulatory parameters using the Pink Trombone[1] (PT) articulatory voice synthesizer. At the core of this is a fast, flexible implementation of PT

[1] https://dood.al/pinktrombone/

which allows formulating the sound matching problem as an optimization task. Given an input voice clip, this approach iteratively optimizes the articulatory parameters of the synthesizer to match the acoustic qualities of a reference clip. This flexibility allows it to tackle a broad range of speech phenomena, from sustained vowels to non-linguistic vocalizations and even singing voices.

Overall, this work contributes:

1. An accelerated implementation of the Pink Trombone (PT) articulatory synthesizer in JAX, enabling efficient differentiation and optimization (planned for open-source release).

2. An approach to reconstructing arbitrary vocal signals, with pre-estimation of only the F_0s and otherwise end-to-end optimization, expanding the range of vocalizations that can be accurately synthesized.

3. Experiments on challenging out-of-domain (real voice) data showing that this approach is significantly more capable than existing gradient-based optimization approaches for vocal tract area function estimation, which are largely limited to synthesizing vowel sounds.

7.2 Related Work

We have been trying to implement machines and anatomical models to emulate human speech for centuries [485]. Computational methods for modeling the human vocal tract to synthesize speech are known as articulatory synthesis [420]. These methods consist of controlling aspects such as the tongue or lips to shape the vocal tract and generate speech by simulating the airflow. These simulations recover the parameters of the tract, which can be controlled and interpreted to assist with pronunciation [303], speech disorders [576], and speech recognition [304].

An important part of articulatory synthesis consists of modeling the human vocal tract [49, 145, 230]. One way of tackling this problem is gathering and training on the combination of speech and corresponding biometric data [95, 114, 306, 554]. Another approach is what is known as analysis-by-synthesis, where parameters are iteratively refined to match the sound target [19, 420, 480]. This can be done using zero-order optimization techniques [73, 91, 105, 170, 232, 335, 422, 449], but it

usually requires more iterations to converge and it's harder to scale. Gradient methods are better at this—and previous research has leveraged neural networks to solve this task [74, 164, 408, 440, 460, 533, 555, 569]—but require large training datasets.

Instead, differentiable vocal tract models can be used to get the best of both worlds: gradient optimization and no requirement for training data. In prior work [490], a differentiable mapping between control parameters and the PT synthesizer is optimized by gradient descent for sound matching vowel sounds. Our approach follows this concept, extending its capabilities beyond vowel sounds generated by PT; synthesizing realistic speech, song, and other vocalizations.

7.3 Methods

The Pink Trombone (PT) is a widely used articulatory speech synthesizer, composed of time-invariant models of the glottal flow derivative (GFD) and vocal tract V. The source GFD is filtered through V, synthesizing the output. To perform end-to-end sound matching, we split our audio input into frames, estimate the fundamental frequency F_0 for each frame, and optimize the vocal tract and GFD parameters for all frames simultaneously. We use a simple objective function involving computing the L_2 distance between the log-mel spectrograms of the target and synthesized audio.

7.3.1 Glottal Flow Derivative

Pink Trombone uses a simplified Liljencrants-Fant (LF) model of the GFD waveform [146]. The LF model is composed of two parameters, the fundamental frequency F_0 and tenseness T, representing the degree of vocal effort. White noise proportional to $1 - \sqrt{T}$ is added to the GFD waveform. We estimate the fundamental frequency (F_0) of each frame using CREPE [257]. The tenseness T for each frame is optimized alongside the vocal tract parameters.

7.3.2 Vocal Tract

The GFD waveform is filtered through the vocal tract, allowing for the articulation of consonant and vowel sounds. PT uses the Kelly-Lochbaum [251] piecewise cylindrical vocal tract model, composed of a sequence of 44 segments of increasing distance from the glottis with cross-sectional areas $A_1, A_2, \ldots, A_{44}$. At each segment junction, the forward and reversed waves are reflected and propagated as described by the

scattering coefficients:

$$k_i = \frac{A_i - A_{i-1}}{A_i + A_{i-1}} \; \forall_i \in \{2, \ldots, 44\} \tag{7.1}$$

To aim for physiologically plausible vocal tracts, we used a simplified physical vocal tract model to determine the diameters $d_1, d_2, \ldots, d_{44}$ shared across all frames. At each frame, two types of transformations are applied: the tongue and two constrictions [490]. The tongue, defined by two parameters, tongue diameter (t_d) and tongue position (t_p), modifies the base diameter into a sinusoidal shape, mimicking the behavior of the human tongue. One lip and one tract constriction, defined by parameters c_l and c_t scale the base diameters of the subset of diameters furthest from and closest to the glottis, respectively, by a factor of $1 - c_l$ and $1 - c_t$. To simplify the gradient-based optimization approach, we keep the constriction indices set at 12 and 39.

7.3.3 Optimization

We use a common mel spectrogram representation of the audio signals, and define our objective $\mathcal{L}$ as the L_2 distance between the target (T) and synthesized (S) audio:

$$\mathcal{L}(\text{T}, \text{S}) = \| \log(|\text{MELSPEC(T)}|) - \log(|\text{MELSPEC(S)}|) \|_2 \tag{7.2}$$

We minimize $\mathcal{L}$ over our parameter space using the AdamW [313] optimizer (with $\gamma = 0.01$), and use a box projection to keep the parameters $\in [0, 1]$. We use a normalized parameter space, back-transformed to each parameter's respective range as needed as has been done in other synthesis packages [89]. We initialize the diameters using the canonical values [490], and other parameters to 0.5 (middle) except T (tenseness coefficients) to 1, to minimize unnecessary noise at the beginning of the optimization. Unlike prior work [490], we do not use inverse filtering to recover any coefficients, and instead perform end-to-end optimization of the full apparatus (except for pre-estimated F_0s).

7.4 Results

We evaluated *VocalTrax* against Vocal-Tract-Grad [490] and ground truth using automated metrics and human evaluations on multiple datasets. Table 7.1 shows results

of automated evaluations on three datasets: TIMIT [173] (subset), AudioMNIST [38] (subset), and VIVAE [215]. We used match error rate (MER) for TIMIT[2] and accuracy otherwise. Given the distribution shift between target audio and even relatively high quality reconstructions, we complement the automated evaluation with a human evaluation. Table 7.2 shows human accuracy responses. Importantly, all our evaluations are on *out-of-domain* data (i.e. data not synthesized with vocal tract models which can be perfectly reconstructed given the same tract model, but rather recordings of real speech).

7.4.1 Automated Evaluations

	TIMIT [173] (MER ↓)	AudioMNIST [38] (Acc ↑)	VIVAE [215] (Acc ↑)
Ground Truth	6.4	73.5	29.2
VocalTrax (Ours)	82.9	20.0	18.4
VTG [490] @ 1024	99.5	9.7	17.7
VTG [490] @ 2048	99.4	12.2	17.3
VTG [490] @ 4096	99.6	10.7	15.6

Table 7.1: Results from automated evaluations. TIMIT uses the match error rate.

In our automated evaluations, we use three datasets to capture different capabilities. AudioMNIST [38] focuses on spoken numbers, which are simple and brief excerpts but do contain semantic information. Given the scale of this dataset, we use a stratified random sample of 600 test-set examples to evaluate across the different methods. By contrast, VIVAE [215] focuses on paralinguistic vocalizations, which do not contain any words but convey affective information. For both datasets, we evaluate using the ARCH [275] benchmark protocol, modified to train on *real* data, and test on resynthesized (or ground truth) data. This ensures a realistic evaluation, wherein models are not trained specifically on the resynthesized speech and can adapt their representations accordingly. Both datasets are for multi-class classification (10 for AudioMNIST and 6 for VIVAE respectively). For AudioMNIST, given the scale (30,000 clips), we do a train-test (instead of cross-validated) evaluation. Finally, we evaluate on a more challenging task: longer-range, higher-vocabulary speech synthesis. We sub-sample 100 clips from TIMIT [173], which contain multi-word phrases or sentences, and aim to resynthesize these fully. We use 2000 optimization

[2]We use MER because word error rate is sensitive to insertions, and thus brief uninformative responses like “thanks for watching” (a common Whisper hallucination given incoherent inputs) result in inflated performance.

iterations for TIMIT to account for its complexity, vs. 1000 for others.

Results are shown in Table 7.1 for ground truth test set data, our method, and Vocal-Tract-Grad [490]. For the latter, we evaluate it at multiple matched hop and frame lengths: 1024 (ours), 2048, and 4096 (their original). Since Vocal-Tract-Grad focuses on vowel synthesis, AudioMNIST and especially TIMIT are likely to be quite challenging for it. Overall, we observe that our method is able to deliver improved reconstructions, judged by their classification and transcription performance, over these baselines. However, for TIMIT and AudioMNIST, our results remain distant from the ground truth results due to the significant distribution shift in addition to reconstruction artifacts present.

7.4.2 Human Evaluations

	AudioMNIST [38]		VIVAE [215]	
	Acc ↑	Conf ↑	Acc ↑	Conf ↑
Ground Truth	100.0 (0.0)	4.9 (0.0)	47.8 (3.7)	3.7 (0.1)
VocalTrax (Ours)	48.7 (2.9)	2.9 (0.1)	23.9 (3.2)	2.5 (0.1)
VTG [490] @ 1024	11.0 (1.8)	1.5 (0.1)	14.4 (2.6)	1.7 (0.1)

Table 7.2: Results from human evaluations (N=10 participants, each rating 30 AudioMNIST [38] and 18 VIVAE [215] samples per source). We show both response accuracy and confidence, each with standard errors (in parenthesis), computed directly from the sample.

To complement automated evaluations, we ran a listening study (results are shown in Table 7.2). We used subsets of AudioMNIST and VIVAE in this study, focusing on (1) how accurately listeners could identify the category the reconstruction (or original example) belongs to, and (2) how confident listeners were about their choices. We recruited 10 participants via Prolific, and estimated that the study took about 20 minutes to complete. The study was determined by our IRB to be exempt. Participants listened and responded to 90 total AudioMNIST clips (stratified random sample of 3 clips per digit category, and the same 30 for each of ground truth, ours, and Vocal-Tract-Grad [490]) and 54 total VIVAE clips (similarly, 3 per affect category, and the same 18 across the 3 sources).

We modeled accuracy using a mixed-effects logistic regression for each dataset, with random intercepts for digit (AudioMNIST) or category (VIVAE) and for participants.

Then, we conducted pairwise post-hoc contrasts. For AudioMNIST, participants were significantly more accurate identifying *VocalTrax*-synthesized digits compared to Vocal-Tract-Grad (odds ratio = 10.7, p<.0001). This was also true for VIVAE, though with a more modest difference (odds ratio = 1.96, p=.019). These p-values were adjusted using the Benjamini-Hochberg correction for pairwise tests. For both datasets, participants were also more confident in classifying our reconstructions. Participants were less accurate and confident with our reconstructions compared to the ground truth clips, suggesting significant opportunities to further improve reconstructions of challenging, out-of-domain samples.

7.5 Conclusion

VocalTrax demonstrates how end-to-end optimization can improve articulatory speech reconstruction of acoustic signals. Our JAX implementation of Pink Trombone and reconstruction approach can rapidly reconstruct a variety of vocal signals, which we hope will open up possibilities in speech analysis, therapy, and voice conversion. However, the quality gap between such synthetic and natural speech persists. Future work should focus on refining vocal tract models, incorporating perceptual factors, and expanding to more complex vocal phenomena.

Part III

Human-AI Interaction and Co-Creation

8

Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence

When a writer sits down to craft a story, the possibilities can feel exciting, but also overwhelming: characters to invent, settings to imagine, plotlines to weave, twists to devise. Even seasoned writers, with the benefit of experience, encounter moments like this. Consider Herman Melville, who famously struggled to write his ambitious novel about a certain great white whale. Melville's challenges were perhaps compounded by his desire to push literary boundaries, often toiling in isolation without much creative support.

Today, assistance for such struggles frequently comes in the form of AI systems capable of generating material on demand. Imagine having such a system as a writing assistant: would you use it to complete a stalled sentence, suggest a new plot direction, or detail a vivid scene? The answers likely depend less on the system's technical capabilities than on your writing process, your current needs, and how you make meaning from external suggestions. In other words, generating content doesn't solve your problem. A successful integration of that content with your story, though, might.

While previous chapters examined how AI can augment human creativity in audio and music, studying real-world creative partnerships requires finding domains where meaningful human-AI interaction is already unambiguously possible. Writing offers exactly such an opportunity. Language models can already engage in the core creative act here (extending a piece of text) in a way that easily composes with human input and editing. This capability creates a useful setting: we can systematically study how humans integrate machine suggestions into their writing process, observing both

the evolving artifact (the story) and the sequence of decisions (integrations) that produced it.

The study detailed in this chapter, conducted in 2020-2021, predates the release of large language model-based interactive systems like ChatGPT. This timing proved fortunate in some ways, allowing us to conduct such an investigation before widespread expectations had been established. The patterns we identified in how writers evaluate, transform, and integrate algorithmic interventions into their stories are, however, not specific to a specific language model or assistant. Rather, they represent a conceptual framework for designers to consider when constructing interactive writing assistants, and more broadly a potentially useful lens for co-creation generally.

Abstract

While developing a story, novices and published writers alike have had to look outside themselves for inspiration. Language models have recently been able to generate text fluently, producing new stochastic narratives upon request. However, effectively integrating such capabilities with human cognitive faculties and creative processes remains challenging. We propose to investigate this integration with a multimodal writing support interface that offers writing suggestions textually, visually, and aurally. We conduct an extensive study that combines elicitation of prior expectations before writing, observation and semi-structured interviews during writing, and outcome evaluations after writing. Our results illustrate individual and situational variation in machine-in-the-loop writing approaches, suggestion acceptance, and ways the system is helpful. Centrally, we report how participants perform *integrative leaps*, by which they do cognitive work to integrate suggestions of varying semantic relevance into their developing stories. We interpret these findings, offering modeling and design recommendations for future creative writing support technologies.

8.1 Introduction

Augmented writing systems pervade human-computer interaction in everyday life, taking various forms to suit specific tasks. From spelling and grammar checkers to tappable word predictions and suggested email completion, these systems are typically designed to enhance human performance and productivity. Recent work in machine learning, intended to improve performance on these tasks and others

(such as machine translation and text summarization), has given rise to formidable natural language understanding and generation models [61, 401]. These are often demonstrated by application to automated or semi-automated narrative generation tasks [9, 330, 394], an essentially creative domain. Given these advances, some recent work has begun to investigate the possibility for such models to be applied to enhancing human creativity, in a machine-in-the-loop setting [72, 101].

Much remains unexplored about how emerging methods in AI, machine learning, and natural language processing might influence creative writing, in part due to the ambiguity and variability of human writing processes. These processes go beyond the linear projection from idea to a full text; research shows how planning narratives, translating ideas into visible textual material, and reviewing are all happening and interacting throughout the process rather than simple sequential stages [155, 366]. However, this is a very familiar process for humans when communicating through writing; as every writer knows, having good ideas does not automatically produce a good text progression. The need for that "good idea" to be anchored and developed so that the reader can be invested takes a great deal of effort. In today's world, language generation models like GPT-2 [401], GPT-3 [61], and new ones coming down the line are typically silent on the inner processes of negotiation and decision that a human writer is working through. Additionally, possible forms contributions from these systems might take to influence writing are not limited to text; writers are able to engage multiple perceptual channels through their work: they may activate multisensory imagination through evocative imagery, invoking auditory and olfactory phenomena, and other forms of sensory description.

We propose to investigate how participants engage with a system that does the following: a multimodal writing support interface that bridges generated writing suggestions with multimedia retrieval to produce concept representations simultaneously in sight, sound, and language. We pair this interface with an extensive study that combines surveys, interaction, and semi-structured interviews during observed, think-aloud writing sessions.

Through this study, we examine and report in detail how participants receive, consider, and integrate suggestions from an intelligent tool into their writing. We explore prominent axes of individual and situational variation in these integrative behaviors, noting the different kinds of "leaps" participants make to understand suggestions and make the necessary compositional decisions and actions to incorporate new

information contained in them, ranging from copying and pasting to re-writing core aspects of their entire story. We are specifically motivated by the following questions:

RQ1 What kinds of assumptions, expectations, and understanding are brought into interacting with an AI creative writing system by different users?

RQ2 How do different users process and integrate different kinds of writing suggestions, and how and why do they accomplish this?

RQ3 How does this suggestion-informed interaction compare with unassisted (or potentially human-assisted) writing?

RQ4 What does the combination and interaction of these three factors mean for intelligent writing support tools?

We design our study from a hybrid *Expectation-Process-Outcome* model (a visual depiction is shown in Fig. 8-1). We seek to capture prior *expectations*, which we do through what we call *explanatory models*, combining aspects of mental models and folk theories of technology. We study the *process* by closely observing participants as they write with the interface, asking questions, and encouraging them to describe and reflect on their thoughts and decisions. Finally, we include an evaluative survey through which participants report on their experience both independently and in comparison to a "blank page" style version of our interface. By combining these sources of information, we seek to document and communicate a range of behaviors, needs, creative processes, and results.

Our findings suggest that (1) different interaction approaches affect writer needs from system suggestions, (2) varied prior assumptions and explanatory models exist and may be both anchored to and adjusted during the interaction process, (3) suggestions support writing in more and less visible and direct ways, and (4) participants perform different kinds of *integrative leaps*, involving cognitive work to make suggestions useful to their writing. We interpret these findings and make commensurate design recommendations for future creative writing support tools.

8.2 Related Work

There is a great deal of related work along multiple axes of this project. Here, we review significant precedents and influences on our work in six disciplinary areas. We

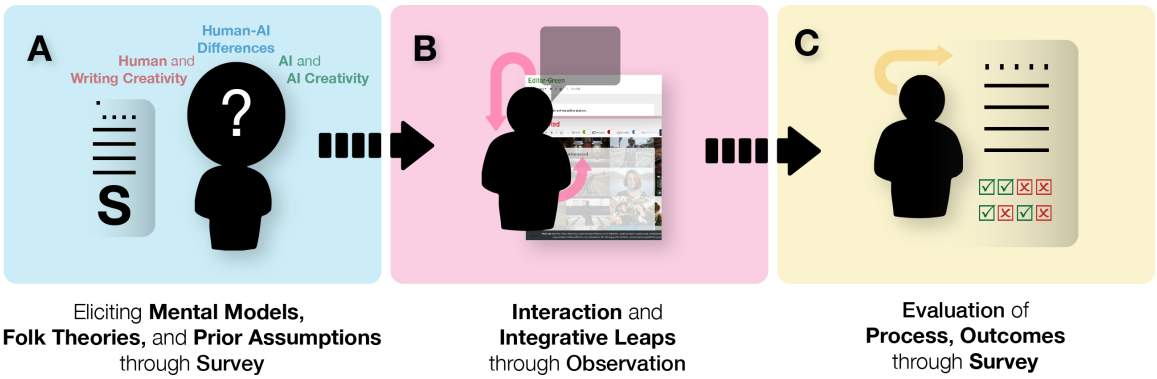

Figure 8-1: **Our *Expectation-Process-Outcome* study model.** We seek to capture **(A)** each participant's "explanatory models" in areas relevant to our system, **(B)** the most salient features of their interaction and sense-making process in writing with it, and **(C)** their evaluation of the outcomes and experience.

additionally review relevant conceptual background areas that inform our empirical methodology and goals.

8.2.1 Studying Writing

Flower and Hayes [155] describe what they term a cognitive process theory of writing. They model several components as part of this: the task environment includes text produced upto a given point, as well as the rhetorical problem at hand, and the writing process(es) involve planning (generating ideas, organizing them, and setting goals), translating (transforming ideas into visible text), and reviewing (evaluating and revising). At a theoretical level, these components are of interest to us because what they seek to model is how writers make decisions while writing and what factors affect this. We similarly seek to understand how writers make decisions and meaning through interaction with a supporting AI tool.

At a methodological level, they rely on *protocol analysis*, wherein participants perform an assigned rhetorical task as they think about their actions out loud and are recorded doing so. They note that this avoids the drawbacks of *introspective analysis*, in which participants report on their actions after-the-fact, observing that this tends to be colored by what they think they *should* have done. Participants are also instructed not to self-analyze under this method. While this provides a helpful starting point, our circumstance is different: participants are not following a task they know how to do and reporting on it. Rather, they are interacting with a new system and engaging

in a new process (or a new version of a familiar process of writing), and as such we need more information from them to adequately understand aspects of their relationship with this system and adapted process. For this, we turn to an interpretive methodology, informed by *thick description* as we will describe later in this section.

More recent psychological approaches to studying creative writing have emphasized the role of retrieval, conceptual combination, and analogical mapping as some of the fundamental cognitive processes that explain creative cognition [154, 246, 543]. Other work on writing has emphasized social-interactive [369] and sociocultural models of writing, studied with a range of social scientific empirical methods that consider how writing activity is "*situated* in concrete actions that are simultaneously *improvised* locally and *mediated* by prefabricated, historically provided tools and practices" [398]. Robertson et al. characterized the conditions under which email-replies-suggestions generated by an AI system are perceived as problematic [425]. They highlight how social context, not just content, can influence how "brief suggestion-like email replies" that ignore social context have the potential to turn otherwise appropriate replies into inappropriate ones. Recognizing that these factors are also essential for understanding writing, especially as writing is *re-situated* and *re-mediated* with new technologies, our approach is informed by heterogeneous empirical studies of writing.

8.2.2 Writing Support

Systems that support writing tasks are ubiquitous, including word prediction, spelling and grammar checking, dictation (speech-to-text transcription), auto-completion of emails, and more. These systems have a substantial history, along with empirical investigations of their effects on writing, productivity, and behavior. For example, Smith and Goodwin [475] investigated lexicon-based single-key vs. double-key typing support for numeric keyboards in context-constrained settings, i.e., for jobs that required such text entry, indicating how computers may help resolve ambiguities arising from the former. Early work on spelling, punctuation, and grammar checking as well as additional textual analysis found that computer assistance could improve writing without substantially increasing writing time, but also found that it prompted users to think critically about their writing when they might not have prior to interaction with such systems [320]. Perhaps in contrast, Woodruff et al. [549] found that their high-level composition assistance tool was absorbed into a less critical

sequence-of-suggestions text composition strategy, despite reports that it was helpful. In modern mobile devices, tools like tap-to-complete word prediction and correction [45] are commonplace, and have been shown to increase typing accuracy [160], though recent work has investigated the effects of such tools on other aspects of the writing experience [72]. Here we will discuss general purpose approaches that are designed to help develop ideas in writing, or otherwise influence or shape various writing tasks.

Moving beyond word-level predictions but retaining their contextual role, Arnold et al. provide methods to generate phrase-level continuations and demonstrate their impact by showing that phrase completions can be offered in a way that they are accepted by the user and interpreted as suggestions rather than predictions [16, 17]. More recent work by Arnold et al. examines the effects of predictive text on writing content, finding that efficiency enhancements apply not just to the process of writing but to the range of content emerging from this process as well [18]. They found that predictive text suggestions—even when presented as single words—are taken as suggestions of what to write. These suggestions often influence the length of the text generated by the user.

Some prior work has also provided multiple simultaneous suggestions to demonstrate different directions [16, 400]. Nicolau et al. identify *cardinality* as an important design factor for non-visual word completion systems [362]. More recent work by Buschek et al., conducted in parallel with ours, has examined the effects of parallel phrase suggestions on writers in an email-writing task [68]. They found that multiple parallel suggestions increased suggestion acceptance, especially for non-native English speakers. InkWell is a writer's assistant designed to help writers augment their creativity by generating various revisions of a given text, employing a synonym-based dictionary and a wide variety of soft constraints [163]. InkWell shows the importance of providing text variations to the user and how this can lead to better writing. Much of this work points to multiple suggestions being helpful, but these are often stochastically (or probabilistically) varying and cannot be reasoned about consistently or causally as complementary channels. Additionally, they do not capture the hierarchical structure of stories. Based on these findings, we decided to add two models to our system that can provide the user with suggestions corresponding to different hierarchical semantic targets in parallel.

For academic writing support, Liu and colleagues [305] introduced G-Asks, a system

for improving students' writing skills, e.g., citing sources to support arguments and presenting evidence persuasively. Their system generates questions with a template-based approach by using Tregex [292], a robust algorithm used to replace keywords in a sentence and the Stanford Parser [116], a natural language parser program that works out the grammatical structure of sentences. As input, the system takes individual sentences and generates questions for the following citation categories: Opinion, Result, Aim of Study, System, Method, and Application. By doing this, the approach supports writing not only to produce content, but also to learn. Inspired by the implication of such a system we consider how suggestions might provoke cognitive processes.

Finally, several projects have considered the role of artificial agents or AI tools in creative story-writing support. Osone et al. found that more writers enjoyed working with a Japanese generative model than not [374]. Roemmele and Gordon's Creative Help system makes suggestions that users can edit to incorporate them into stories [429]. As part of this work, they evaluate how much suggestions are edited as a proxy for suggestion helpfulness. In later work, they additionally study how randomness or unpredictability in suggestion can influence writers' attitudes, finding that increased randomness lowers ratings of factors like coherence and increases ones like perceived originality [430]. Both aspects of this work are relevant to our study, in which we examine how writers edit both suggestions and their stories to integrate suggestions, and additionally the relevance-variety trade-off in a more implicit sense, by observing and reporting the interaction. Gero and Chilton present Metaphoria, which generates metaphors to support creative writing [181]. Their work discusses both ownership and what they call "divergent outcomes" resulting from the suggestions, both of which our study addresses. Clark et al. propose a machine-in-the-loop creative writing system and study its application to stories and slogans [101]. The authors note several findings and make commensurate recommendations, some of which we build on. For example, they note the challenge of balancing between the easily-ignored "pull" interaction, and intrusive automatic suggestions. We build on this by combining direct invocation with a wait-threshold timer-based hint display. WordCraft [102] frames the collaboration between a human storyteller and AI within an open-ended dialog system with more explicit turns and turn types than what is implemented in this chapter, but the effects of these design choices were not evaluated in a systematic way. Finally, like the creators of FairyTailor [41], we also introduce multimodality, which could have significant effects on findings, due to effects on the

content (how are images or sounds translated to text?), the process (how is this information integrated into existing text?), and the overall experience. As such, this work additionally provides helpful background to contextualize our findings.

8.2.3 Language Models

Statistical language modeling has recently made substantial advances through the application of new techniques [529] to very large models [61] and corpora. The predominant paradigm in many natural language tasks has, as such, moved to transfer learning, in which general-purpose models are fine-tuned on downstream tasks. This mitigates issues of data scarcity, reduces training time, and improves performance. In our case, we are specifically interested in causal or autoregressive language models, which probabilistically predict successive tokens from prior ones. We fine-tune two pre-trained variants of the popular GPT-2 [401] language model in our prototype. While these models are no longer necessarily the state-of-the-art language modelers due to rapid developments in a fast-moving field, they are still competitive performers that are state-of-the-art in interactive systems, where other factors including ease of fine-tuning (flexibility), speed of response (interactivity), and open availability are important.

8.2.4 Multimodal Feedback

By presenting various communication channels, multimodal systems are considered to support human information processing by using a range of cognitive resources. This assumption is largely based on cognitive theories proposing multiple, modality-specific processing resources [25, 380]. One goal of a well-designed multimodal system is to integrate complementary input modes to create a synergistic blend, permitting each mode's strengths to overcome weaknesses in the others and support "mutual compensation" of feedback errors [375].

In addition to these cognitive benefits, multimodal feedback offers us a rich window into participants' reasoning and process of sense-making. While language processing alone demands high engagement to process and to make sense, we aim to study how a complementary blend of information representations can allow us to uncover varied aspects of participants' interaction with an intelligent system for creative enhancement. In this section, we look into our two non-textual modalities for feedback: still visual input (images) and auditory input (sound recordings). We review how each of these

modalities has been used to support users on a given task, and consider how these approaches might indicate possible benefits for our task.

Imagery

iTell [279] supports retrospective storytelling with digital photos. It employs a design process based on providing support to help novice storytellers engage in the composition process like experts. To assist users with creating retrospective narratives about their personal experiences, iTell presents the users with four steps to complete: Brainstorm, Organize, Writing, and Add Personal Media. The user must finish each step before proceeding to the next step and cannot skip a step without completing it at least once. One interesting finding from the workshop conducted as part of their user study is the influence of the media modality on the novices' retrospective story development, how novices approach retrospective storytelling, and what is needed to make novices successful retrospective storytellers. In particular, the authors show benefits for novices to have access to mixed media in the story development process. One of the significant differences between iTell and our system is that iTell requires the user to gather any media material beforehand to retrieve and incorporate it during a writing session. Another significant difference is the lack of text suggestions to help the user in their writing.

Another example of a support tool for the development of new ideas is Design Daydreams (DD) [353]. DD is part of a suite of computational design tools that integrate ambiguity and juxtaposition into systems that users can use to discover new ideas. Using a low-tech augmented reality system to overlay digital images on top of objects visually, the Design Daydreams augmented "post-it note" fluidly extends the inspiration designers find online into the physically-interactive and collaborative brainstorming environment. Feedback suggested that the low fidelity of the tool provided a natural ambiguity that left room for interpretation as designers juxtaposed digital and physical concepts together to create new ideas. Like these projects, our visual feedback aims to discover mental constructs related to the story. It does this either indirectly, through the mood created by the image palette, or directly by layering diverse representations and allowing object or concept features to be distinguished and integrated into the developing story.

Audio

Specific attributes of the surrounding environment have been shown to support memory, foster creativity, enhance sensitivity to details, and balance cognitive load [96]. For instance, Mehta et al. found that moderate noise levels, like a coffee shop's ambient sound, facilitate abstract processing [340]. Zhao et al. built a multimodal mediated work environment, where they demonstrated effects on occupants' ability to focus and recover from stressful situations [577]. Sounds with attributable causes (i.e. where humans are able to aurally discern the source) have also been shown to impact memory [409], language learning [550], and, as a feedback modality, attention and information communication [175]. Motivated by this, we integrate an audio feedback system that retrieves sound by concept (rather than by content), to offer a semantically relevant aural dimension that may confer these benefits in the process of writing a story.

8.2.5 Interpretive Approaches

We approach our observation of participants' interaction through the lens of interpretation. Interpretation as a concept has been used in a number of papers in HCI [34, 278, 358, 456]. The interpretive perspective we maintain in this work is informed by anthropological approaches to make visible the alignments of factors of interaction that would otherwise go unnoticed due to common-sense understanding. Our theoretical approach is built on the dichotomy of social theory concepts of understanding as causal explanation (*erklären*) versus understanding as interpretation (*verstehen*).

Following Max Weber's distinction [545] between *explanation* that captures the causal sequence of actions and *understanding* that attends to the meaning of those actions, our research aims to analyze the interaction of the person with the AI system from the perspective of the latter (i.e. "meaning"). More specifically, the meaning of actions from the point of view of the participants, who organically construct meaning in the process of engaging with complex systems. As such, interpretation in this research is a form of understanding that makes it possible to discern the meaning production that occurs within the interaction between the human and the AI system. To that end, we aim to identify and observe how the interaction is influenced by the explanatory models of AI that users have. We look at what type of conceptualization work is done on the part of the users in the process of engaging with AI, both in the world (prior

assumptions and expectations) and locally in our study (impressions, integrative processes, and interactive reasoning), and how they rely on such conceptions to navigate the process and products of co-writing with AI.

"Understanding" implies the meaning of actions can be transferred through co-presence with a participant in one space, being able to build rapport, and engage with the participant so as to understand how people make sense of the world around them. For this research we aimed to complement the quantitative and qualitative data obtained from survey questions with qualitative data from semi-structured in-session interviews, observation of the participants, and what is called, in the social sciences, “thick description” [176]. Thick description allows us to go beyond the observation of causal actions and acquire interpretation by the actors of not only their own actions but also of the context within which they operate. We detail our specific approach in a later section on study design.

8.2.6 Explanatory Models and Expectations of AI

We use the term "explanatory models" to refer to the super-set of two kinds of conceptual representations of computational systems, commonly referred to as "mental models" and "folk theories" respectively. Here we describe each and outline our rationale for combining them in our work.

Mental and Conceptual Models

Human-AI researchers often use the concept “mental model of AI," a term informed by psychology and cognitive science. In the context of human-machine collaboration, and even for human collaboration alone, a great deal of work has illuminated the importance of mental models in promoting team success [118]. In the case of AI, it has been shown that optimal inference does not necessarily yield optimal human-AI team performance. Bansal et al., for example, study mental models of AI performance in the context of human-AI teams [31]. They do note, however, the relevance of other types of mental models (such as those of how the system works) to collaborative settings.

Gero et al. study human-AI collaboration in a game setting, and their results suggest that understanding of the system alone insufficiently develops appropriate conceptual models [182]. The same authors distinguish between mental and conceptual models by indicating that the latter are held by those with extensive knowledge of the

system, e.g., designers and experts. This follows Norman's formulation of these two terms, where he suggests that *conceptual* models are "invented by teachers, designers, scientists, and engineers," [367] noting that researchers then *conceptualize* the mental models through experiment and observation in order to produce systems and conceptual models that direct these mental models to be coherent and usable.

Folk Theories

While mental models offer insight into cognitive representations of a system's operation developed through experience, intuitive theories about the world structure learning, understanding, and cognition more broadly, in diverse ways despite a common psychological substrate [177]. Folk theories are a form of expectations that are based on some experience, but are not necessarily systematically checked [423]. Mental models are structured accounts of a system's mechanics and behavior, but folk theories and implicit beliefs arise from a great many sources of information and interaction, and are not constrained to nor will they necessarily contain an understanding of "the relationship between inputs and outputs" [162]. Folk theories may be especially salient in the domain of AI systems, given their dramatic and continued impact on culture and society. Few kinds of technical systems are as pervasive in the collective consciousness, due to rapid advances, news reports, economic incentives and concerns, and potentially profound implications for human identity and activity.

Folk theories have been captured in the study of cyber-social systems, often relating to algorithms employed in social media platforms. These theories may be elicited through *direct* investigation by researchers, often through interviews and associated methods, or *indirectly*, through inferential procedures applied to data "in-the-wild", such as posts on a social media platform. As an example of the former, Eslami et al. elicit theories about the operation of Facebook's news feed algorithm and a designed alternative [139]. In contrast, DeVito et al. aggregate and analyze over 100,000 tweets to determine user folk theories that contribute to resistance against changes in Twitter's algorithmic content curation system [121].

Why combine user mental models and folk theories?

To understand the prior assumptions, expectations, and understanding that our participants brought into their interaction with our system, we captured their *explanatory models* in related areas. Specifically, we identified related areas as AI and

AI creativity, human creativity in writing, and differences between humans and AI in creativity and writing. Our system is specific enough that they are unlikely to have encountered a substantially similar system before, and are accordingly unlikely to have developed intuitive theories or mental models of our system. As such, these contextually informative areas may have bearing on their experiences, writing and sense-making processes, and evaluations of the outcome.

Creative processes with AI allow varied creators to expressively produce diverse artifacts. We aim to similarly capture complex and multidimensional explanatory models, considering both cognitive and sociocultural factors to obtain extensive representations of participants' prior assumptions, expectations, and understanding. We build on the concept of mental models to consider aspects including the user's beliefs about and attitudes towards AI and creativity, about the production of creative writing artifacts, and consider how these might affect downstream evaluations of the process of interacting with our system. We believe this approach can work inform design processes to yield tools that have clear affordances in creative contexts, and support a diversity of needs and practices.

8.3 System Prototype

Our experimental prototype consists of two writing interfaces: **Editor-Green**, a minimal "blank page" tool, and **Editor-Red**, our augmented multimodal tool. To minimize cognitive bias when conducting our user study, we chose to give names to the editors that would seem roughly equivalent. The system also contains a server that runs language models, as well as a real-time database to track inputs, responses from the server, and interactions, e.g., interface settings. Fig. 8-2 shows both interfaces, including an active multimodal response in **(B)** with images[1] and sounds. Fig. 8-3 shows the underlying data flow through the system architecture that makes these interfaces possible.

[1]Photos by Alan Ren, Matus Karahuta, Javad Esmaeili, Cassie Lopez (1), Christopher Campbell, Brooke Cagle (1), Benn McGuinness, Cassie Lopez (2), Claudio Schwarz, purzlbaum, Matheus Ferrero, Cory Woodward, Tim Photoguy, 2 Bro's Media, Nature Uninterrupted Photography, Fabe collage, Brooke Cagle (2), Ronaldo de Oliveira at Unsplash.

8.3.1 Writing Interface

Editor-Red contains a page-like typing environment, with two suggestion blocks adjacent to it that contain suggestions (these are below the writing page on mobile devices). The two blocks offer two different types of suggestions, corresponding to text generation models fine-tuned on two different datasets. These are returned and presented through images and sounds in addition to suggested text. There is a control panel at the top, which contains some basic formatting features (text styles including heading levels, etc. as well as font formatting), and controls for invoking language model suggestions and switching multimodal response displays. One switch selects between two image presentations: by default, (the "on" position), images are displayed as a "grid" with full opacity behind the writing and response display elements. When toggled off, images are displayed as an "overlay", with multiple images stacked on top of each other and their opacity set low enough that they form an environment together. Two modality switches, one each for images and sounds, turn on and off the inclusion of these modalities, respectively. A slider can be used to adjust the volume of retrieved sounds. Finally, a "Suggestion" button launches a query for a new suggestion, and associated images and sounds, based on the current text of the user's story). Suggestions can also be invoked via the tab key, and after about 10 seconds without any writing activity, a hint regarding suggestion availability appears (indicating that tab can be pressed for suggestions). The suggestion texts are colored with a gradient to clearly distinguish them from user-written text, and are virtually "typed out" over a small amount of time to visually illustrate their narrative structure. A text field at the bottom includes credits for the presented photographs.

Several design features of our writing interface are based on popular word processing platforms. For example, a paper-shaped writing area, and a toolbar at the top for text formatting and other controls. The other design choices we made, for the new features we proposed, were refined through early prototyping and pilot testing with fourteen users, through which we discovered several usability challenges and corresponding solutions. We added tab-based suggestion invocation in addition to re-positioning the button to avoid accidental triggering based on observations made during these sessions. While we initially designed the image display as an overlaid blend underneath the writing area, we found that the grid-style display allowed for more explicit idea borrowing when desired, due to the increased clarity of individual images, thus making this display the default. We selected a number of images (20)

that we found yielded sufficient variety from individual searches, but maintained the individual clarity on typical screen sizes. Finally, we had originally placed suggestions at the bottom of the interface, but found that this constrained space for writing and required more scrolling. We moved it to the right of the writing area to both position it as secondary, and allow quick glancing. We additionally added a gradient to this text to firmly distinguish it from user-written text.

To parallel this augmented interface, we provide another with the same core features, design, and layout that we call **Editor-Green**. This editor includes the text formatting features and the page-like writing environment. We use this interface as a point of comparison, based on the core features of common writing tools. The additional **Editor-Red** features are turned on and off, effectively switching between the interfaces, by clicking the interface title in the top left corner. We did this to allow flexible switching in the study context, while reducing the likelihood of accidental switching, which we observed in early iterations.

8.3.2 Language Models

In order to produce relevant suggestions, we expected that a pre-trained language model would need to be subsequently fine-tuned on a dataset containing many useful examples. However, narratives develop simultaneously at multiple hierarchical structural levels, and single suggestions do not capture any variation in this important property. As noted earlier, prior work has provided and investigated multiple simultaneous suggestions to demonstrate different directions, which points to multiple suggestions being helpful. However, stories allow us to make some domain-specific assumptions that can make these parallel suggestion channels semantically relevant. As such, we produce two variants of the base language model to capture overall plot and local description respectively, offering multiple semantically distinct channels of suggestions. Stochastic variation is also available by simply requesting additional suggestions in sequence without any additional writing (though we note this does introduce additional delay).

We fine-tuned the same language model on two different datasets, producing two final models. The base model is a medium-sized GPT-2 architecture with pre-trained weights obtained from huggingface[2]. The first experimental model is fine-tuned on a corpus of movie summaries [30], which we observe tend to contain high-level

[2] https://huggingface.co/

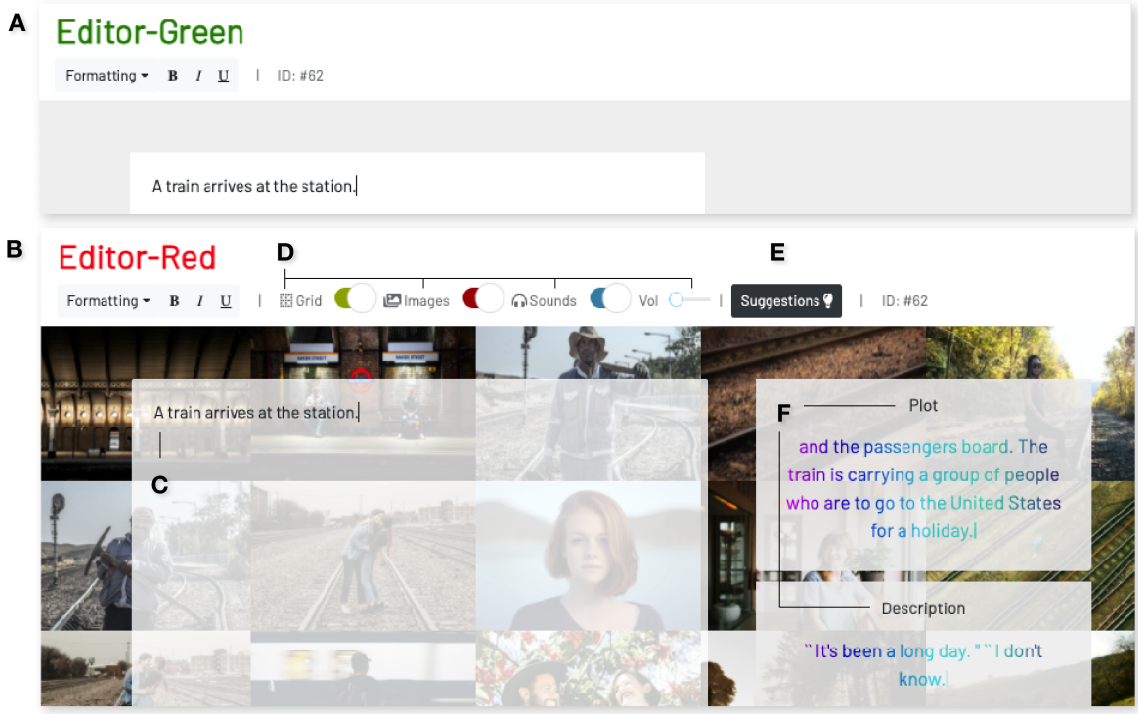

Figure 8-2: **Our experimental writing interfaces. (A)** is a "blank page" editor with only basic formatting features, while **(B)** augments this with generated suggestions and multimodal feedback. In the second interface, users write text **(C)** and can request suggestion by invoking the *Suggestion* button **(E)** or using the `tab` key (a hint is shown after about 10 seconds of inactivity). Two types of suggestions, corresponding to text generation models fine-tuned on two different datasets, are returned **(F)** and presented through images and sounds in addition to suggested text. The user can turn on or off these stimuli, or change the image presentations to an overlay **(D)**.

plot components and event sequences. As such, we label suggestions arising from this model as "Plot" suggestions. The second is fine-tuned on a writing prompts dataset [143], which features prompts and story responses taken from a prominent online forum for amateur fiction. Following the observation by Fast et al. that amateur fiction "tends to be explicit about both scene-setting and emotion, with a higher density of adjective descriptors" [149] as well as our own review of this dataset and the fine-tuned model, we label this second experimental model's outputs as "Description" suggestions.

For each query, our system produces responses from both models. When sampling from the models, we employ a top-k sampling strategy, with $k = 5$, temperature $= 0.5$,

and a repetition penalty of 1.0; these parameter settings were based on initial experiments, i.e. looking at the models' outputs for different combinations of parameter values and making subjective judgements about quality, consistency, and relevance given a variety of input prompts. We decided on a maximum output suggestion length of 40 tokens (tokens are sub-word units, so there isn't a direct relationship to the number of words), finding that this suffices for many scenarios, and weighed against the time and computation needed for autoregressively sampling longer sequences. This cost-detail trade-off is one of many we needed to address in the design process of this prototype. Others included model size, i.e., number of parameters, for which more parameters typically result in greater coherence in modeling long-term semantic consistency but slower performance and consequently significantly greater latency, and the number of model options, with similar constraints. While other work has hosted multiple different-sized models in order to propagate this trade-off to a user decision at the interface's point of querying [40], we wanted to focus our approach on the specific semantic channels of plot and descriptive detail development to support story writing and further reason about suggestion incorporation. As such, we opted to fine-tune two medium-sized models, which balance interactive responsiveness with expressive language modeling.

8.3.3 Multimedia Retrieval

Retrieving visual and auditory stimuli based on natural language descriptions is a challenging task. This is compounded for open-domain text, as in our case. Applications that do this typically need to defer to large internet media databases with search APIs to adequately support the range of possible queries with high-quality media objects. While some recent work focuses on learned approaches to semantic text-image similarity, these approaches are slow, require much data to train, and don't scale well to large databases, and so we opt for simple concept-based (rather than content-based) search of media platforms.

We use two such databases: Unsplash[3] for images, and Freesound[159] for audio. Concept-based searches for media on these platforms are typically performed with keywords rather than long-form text, and so we use the RAKE algorithm for automatic keyword extraction [435] as a preprocessing step, pooling the keywords from both model outputs (*plot* and *description*). We then query Unsplash with the output

[3] https://unsplash.com/

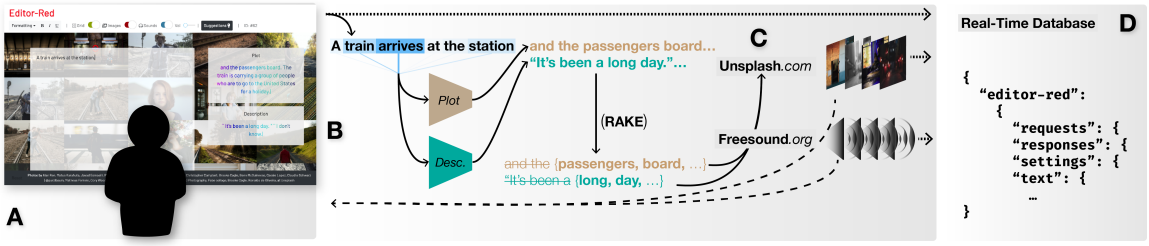

Figure 8-3: **Flow of data through our system. (A)** The user enters text into the interface which is, upon request, transmitted to **(B)** a backend application. This operates two causal language models, fine-tuned for plot-level and description-level suggestions respectively. The text is tokenized and input to both, and generated suggestions are captured. Keywords are then extracted (using the RAKE algorithm) for use in the multimedia search queries: **(C)** calls to the Unsplash and Freesound APIs retrieve semantically associated image and audio content respectively, and these are sent back to the interface along with the suggestions to be presented to the user. In parallel, all use data is logged into **(D)** a real-time Firebase database. We track requests (including the state of the story at each request time), system responses (suggestions, links to media), the latest story state, and changes in settings (e.g., turning any specific modalities on and off). The logging system is replicated, for text only, in **Editor-Green** as well.

keyword list. We observe that Freesound is sensitive to multi-keyword searches and often returns no sounds in these cases, so to avoid rate limitation problems we supply only the first extracted keyword to its API to search for sounds. We apply three content filters to Freesound queries. First, we limit the duration to be between 10 seconds and 30 seconds to allow for sounds that are both long enough to contribute to an acoustic environment, and not so long as to extend retrieval time. Second, we filter out results marked as containing explicit content, after noting that these sometimes appear even when not necessarily suggested by the query. Finally, we apply a filter on the "dissonance" feature[4], which is extracted directly from each audio signal, so it is ≤ 0.4. Since sounds need to be combined together, constraining the sensory dissonance [396] of each independent element helps to layer them effectively into a coherent soundscape.

[4] https://essentia.upf.edu/reference/streaming_Dissonance.html

8.3.4 Data Logging

To allow detailed logging of user interaction with our prototype as well as generated suggestions, we use a real-time Firebase[5] database. This database keeps track of user-typed text in real-time as well as any changes to settings in the interface (e.g., turning sounds or images off or on), time-stamped requests with their associated input text, and time-stamped responses with text suggestions and links to retrieved media. This allows us to approximately reconstruct a sequence of writing events from a session, which we refer to in order to build the later section on *integrative leaps*.

8.4 Study

To investigate user interaction with our prototype and the role of user explanatory models of AI within that interaction, we designed a mixed-methods study consisting of two observed writing tasks, a four-part survey, and extensive logging of interaction data.

8.4.1 Formative Study

As part of our initial exploration, we conducted a formative study with 14 participants and an earlier version of the prototype. We called this interface MLVille, an homage to Herman Melville who famously struggled with writer's block before it was a well-documented phenomenon. With 4 out of 14 participants we conducted open-ended, interpretation-focused interviews, which allowed us to get in-depth data. These participants interpreted their actions and interactions with our prototype while performing the study tasks to provide additional context and insight. Through a broad set of survey questions, analysis of the produced text both computationally and qualitatively, and extensively documenting usability feedback, we performed several updates for our second study iteration, which is described in sections to follow. Specifically, we re-designed our interface, fine-tuned new language models on different datasets, re-oriented our study around thick description (noting the range of information and useful perspective it generated), and re-designed our survey to capture identified factors of variation and interest.

[5] https://firebase.google.com/

8.4.2 Recruitment

Potential participants were invited through large mailing lists associated with various departments and living groups at R1 universities, including one social sciences department and several Computer Science-adjacent lists, as well as a post on reddit. As a pre-condition for being recruited, applicants filled out a short survey to confirm that they meet the requirements of being fluent in English and being over 18 years old. The survey also contained a video tutorial that explained the features of **Editor-Red** and in order to filter for applicants who watched and paid attention to the video, they were asked to answer two screening attention check questions about the system's features. Those applicants who met the requirements were invited so as to have a balanced pool of participants who identified as native and non-native English speakers. We also made sure to have a balanced pool of participants with and without Computer Science backgrounds.

8.4.3 Participant Demographics

27 participants completed the writing task. Data from 4 had to be excluded due to firewall-related issues, mid-session server problems, and unwillingness to complete the task as instructed. All participants reported having at least a high school diploma. When asked about their disciplinary affiliation, 35% replied *Another STEM field*, 21% *Computer Science*, 13% *Life Science*, 9% *Business*, 13% *Humanities*, 4% *Social Sciences*, 4% *Medicine*. Participants' ages ranged from 18 to 45, with 48% of participants in the range of 18-22. 65% of participants reported that English is their first language. When asked "Do you struggle with writing?", 78% of the participants responded yes.

8.4.4 Study Structure

Our study design follows the structure shown in Fig. 8-4. We began by sending each participant a consent form in advance of the scheduled session, allowing them enough time to read and ask questions. They gave verbal consent at the beginning of the session with the interviewer and were then given an introductory overview of the study procedure, which took 3-5 minutes. Participants then completed a 10-minute introductory survey ($\mathbf{S^I}$), designed to elicit the prior knowledge, conceptual frameworks, and beliefs that participants had about AI and its application in writing, human creative writing, and their own previous writing experience. Participants began the first 20 minute writing task, with either **Editor-Green** or **Editor-Red** depending

Figure 8-4: **Study design.** Our study consists of two writing tasks, one each with **Editor-Green** (no augmentation) and **Editor-Red** (with augmentation) for 20 minutes; which interface participants used first was counterbalanced across subjects. For each task, the participant is given one of two prompts (in randomized order) to then create a story with. The two writing tasks are interlaced with sections of a four-part survey, with introductory and background components, as well as one for each writing task. The study takes approximately 75 minutes in total.

on their group assignment. They were instructed to write a story using one of the following prompts: *The phone began to ring* or *A train arrives at the station* (alternating prompts between groups to control for the effect of the prompt). Both prompts were designed to be short, somewhat vague, and contain the beginning of some action (phone call and train arrival).

Most participants, once given the task ("write a story using the following prompt"), began writing without asking any questions. Some participants asked if there were any requirements in terms of genre, structure, or length, and we informed them that there were none. Participants were informed that they should use suggestions only if they find them helpful. Deciding when to stop writing was completely up to participants and we clearly stated that at the very beginning of the task.

After each writing task, participants completed the corresponding follow-up survey, i.e., $\mathbf{S^G}$ (< 5 minutes) for participants who wrote in **Editor-Green** or $\mathbf{S^R}$ (~ 10 minutes) for participants who wrote in **Editor-Red**. In accordance with standard order-counterbalancing, participants completed a second 20 minute writing task with whichever editor they had not yet experienced, followed by its corresponding survey.

Finally, all participants completed a survey that invited them to compare the two writing experiences they had during the session, as well as provide some additional demographics/background information ($\mathbf{S^C}$). The overall duration was about 75

minutes, and participants were compensated with a $25 Amazon gift card.

Two researchers separately conducted study sessions via Zoom videoconferencing. The sessions were recorded with permission, and the researchers took notes throughout the session. Participants shared their screens during the writing sessions, and they were asked to switch this function off when answering survey questions. While writing, participants were explicitly encouraged to comment and react aloud as they wrote, processed information, and responded to incoming suggestions and media. During the sessions, interviewers observed participants' interactions with the prototype and writing process. Additionally they prompted participants to communicate about their thought processes and experiences periodically throughout each writing session.

8.4.5 Survey

The survey consisted of four blocks. The Introductory block of questions ($\mathbf{S^I}$) contained open questions and multiple choice questions. It was designed to elicit the prior assumptions, expectations, and understanding that participants had about Artificial Intelligence and the possibility of its application in writing and creative writing, specifically. Participants were also asked about their own writing and their thoughts about creativity in human writing.

The block of questions after writing in **Editor-Red** contained five open questions on the experience of the interaction, which was followed by a longer section that contained two grid sections with 7-point Likert-type items relating to general usability. After this, there were six multiple choice questions asking participants to provide more detailed information on their experience (e.g."When **Editor-Red** was giving its ideas, what were you paying attention to? Text, Images, Sounds, None", "I think I will enjoy using **Editor-Red** more, if..."). Finally, there was also one more 7-point Likert-type grid of items asking participants to rate statements on suggestions provided by **Editor-Red** (e.g."The suggestions made by **Editor-Red** were creative", "The suggestions made by **Editor-Red** were coherent", "I enjoyed co-writing with **Editor-Red**", "I enjoyed collaborating with **Editor-Red**"). The block of questions after writing in **Editor-Green** ($\mathbf{S^G}$) contained two grid sections, with all items relating to the augmentation features omitted and the rest replicated.

The block on comparison ($\mathbf{S^C}$) between **Editor-Green** and **Editor-Red** consisted of

Yes/No and open questions on creative writing ("Do you consider the text that you wrote in **Editor-Red**/**Editor-Green** creative?", "If yes, in a few words, explain how it was creative. If no, explain in a few words, why not?"). There were also 7-point ordinal items asking to compare **Editor-Green** and **Editor-Red** in terms of creative writing (e.g. "In which editor was the text that you wrote more creative?") and four items on cognitive load adapted from the NASA TLX survey [206] (e.g. "Where did you feel more focused when writing a text?").

The final section contains demographic questions asking participants about their highest degree, disciplinary affiliation, age, and gender identification. Those participants who identify themselves as non-native speakers of English are asked to provide more information about levels of self-reported proficiency of various skills and depth of exposure, which we assess based on pre-existing instruments [247, 328]. In this block, there are also supplementary questions asking participants what kinds of writing they struggle with and how often they do creative writing.

All the questions throughout the four blocks are meant to elicit data for the key phenomena we were interested in: participants' prior understanding and anticipations of AI and writing using AI, how participants understand creativity and creative writing, participants' interpretation of the system's work and their explanation of engagement with the system's suggestions. Additional concepts of usability of the system, cognitive load, and agency were also included. The questions were strategically phrased in different ways (open questions, closed questions, multiple choice, Likert-type items).

8.4.6 Observation and Thick Description

Participants spent about 20 minutes writing in each editor (within each 75 minute session). This gave researchers an opportunity to capture a wide range of phenomena: participants would comment on how they usually write outside the study and how they are writing within the study, explain their process of coming up with ideas, their opinions and judgements of the system's suggestions, talk about how they were making decisions to incorporate or not incorporate suggestions, and give their reasons. The ability to be there with participants when they were writing and to observe immediate reaction and, to the extent possible, raw and unmediated answers, allowed us to produce "thick description" [176], as noted earlier.

Observing the interaction with the system allows us to capture the reasoning of

participants for incorporating or ignoring suggestions, and also glean how participants make sense of the interaction with the system and their strategy on structuring this interaction to support their writing. In describing the interaction with **Editor-Red**, we note that the interaction is not reduced to just getting suggestions from the system. Participants perform a task of writing a text while existing within a particular space, which is defined not only by the interface of the system and its multimodal suggestions, but also those expectations and explanatory models that participants had prior to the study and were constantly adjusting during the study. As such, we seek to capture detail about aspects of their experience that go beyond just suggestion incorporation behaviors.

8.4.7 Data Analysis

Writing sessions

The data from writing sessions consisted of (1) logged data of the texts participants wrote and suggestions they received, (2) transcripts of sessions where participants thought out loud during the interaction with the system and answer interviewers' questions, and (3) the notes that interviewers made during the sessions. After we finished running the study, interviewers watched the session videos, making additional notes and comparing them to the notes they made during the sessions. Then we entered all the data from the writing sessions into a shared document and MAXQDA[6] (software for qualitative analysis). In MAXQDA, we first used a deductive approach to code the data: we employed pre-existing concepts from research questions (such as conditions for acceptance of a suggestion, creativity, agency and ownership, etc) as codes. In the second stage of the analysis, we applied an inductive approach to code the data: in particular, the *in-vivo* method (using the words of the participants to create codes), so as to let the voice of the participants and their actual concepts structure the themes. Two rounds of inductive coding were done, followed by a process involving rearranging codes and turning in-vivo codes either into new themes, or adding them to existing codes. At this point, a second researcher did their round of coding and partially re-coded the data. The two coders discussed and reached agreement on the codes. A third round of coding by a third researcher was done to align and streamline all the codes.

[6] https://www.maxqda.com/

Survey responses

For the answers to open-ended survey questions (expectations), one researcher performed an initial open coding (*in-vivo* method, i.e. using the words of the participants to create codes), followed by a second cycle that involved deductively applying domain concepts associated with posed questions to the initial codes. For example, in asking about differences between human and AI text production, we relied on some concepts from prior literature (such as statistical vs. symbolic processing or novelty, value, and surprise in creativity) that were closely related to the *in-vivo* codes. This was accompanied with a values orientation (i.e. trying to infer participants' values and beliefs). Then a second researcher reviewed and partially re-coded the same data, and disagreements were resolved through discussion of instances and codes themselves (labels and definitions), as well as including secondary codes for individual responses where appropriate.

8.5 Results

We detail findings from the three primary components of our study: our survey to capture participants' prior assumptions and pre-existing explanatory models, observations and responses during the semi-structured interview process that accompanied their writing, and questions posed afterwards about their final thoughts and experiences during the sessions. In this section, we focus on detailing each independently before examining the synthesis of their respective data. We explicitly review specific examples of how these data interact, but note that our broader findings are informed by all three sources as they represent different means of inquiry and perspectives on the experiment.

8.5.1 Prior Assumptions, Explanatory Models

Detail in explanation vs. technical depth and accuracy

We assessed the structure of participants' prior explanatory models of AI through one open-ended question, i.e., "How do you think AI works? (For example, where does it get information? How does it produce information? How does it understand what you ask it?)", expecting a range in the responses. We observed during the first coding cycle that the results seemed to actually vary in more than one way (rather than being more or less structured overall), and so we model this as a two-dimensional

construct. During the second (deductive) coding cycle, we adjusted code labels to relate them to prior and parallel work:

1. **Type of Explanatory Model**
    (a) *Abstract*: What AI does
    (b) *Operational*: How AI works
2. **Technical Depth**
    (a) *Sparse*: Vague or inaccurate description
    (b) *Sophisticated*: Low-level and accurate description

We based these labels on prior and parallel work. Specifically, DeVito et al. describe *abstract* and *operational* algorithmic folk theories, noting that the former "do not include specific attempts to theorize how an algorithm might actually operate" [122] (their sub-codes for these are not applicable to our case). Interestingly, in a study of mental models of adversarial machine learning, Bieringer et al. found that their participants' prior knowledge did not necessarily determine the technical depth of elicited mental models, pointing to a possibly multidimensional space. They apply the labels *sparse* and *sophisticated* to describe the technical depth in these mental models [47].

The majority of participants in our study gave *Sparse-Abstract* models ($N = 14$). For example, P3 wrote "I think it gets info from devices and uses language features to understand us." See Table 8.1 for the full distribution over these label combinations, and additional examples.

The second most common explanatory type was *Sophisticated-Operational* ($N = 5$). For instance, P1 alluded to both generalization and optimization in their explanation: "I think that it works by feeding it data. It is then able to use the data that it is fed and apply the given outcomes for the provided data to novel situations. It is accurate as it continues to learn and reduce loss between the real and given answer."

Sparse-Operational and *Sophisticated-Abstract* explanations each occurred twice. We identified P16's description as an instance of the former:

	N	Example
Sparse-Abstract	14	"I think AI is a software that attempt to resemble the way an intelligent brain works. I suppose it bases its decision on a set of situations that are used as possible scenarios."
Sophisticated-Operational	5	"Different kinds of AI work differently. If we're talking about machine learning systems, they are trained with large corpi of data that are curated by data scientists or machine learning engineers. The algorithms for these systems find and exploit patterns in the data to then accomplish tasks. If I ask an AI something, it will look for a way to take my input and compare it to patterns in the corpus of data it was trained on."
Sparse-Operational	2	"i think it works by first being given a set of instructions, or base algorithms, which are then trained by feeding it various data sets/ user inputs. For example I'm pretty sure visual captcha companies use user input data from instructions like"select all the traffic lights in this image" o train or test their own image recognition algorithms. The data is relayed to the computer in a format it can interpret, such as code for matlab, and the computer then recognizes which configurations of code evaluate to true given the desired condition. For text, it can also scan the input for key words/tags, that make it branch down a certain path in the algorithm."
Sophisticated-Abstract	2	"There are different kinds of AI. The most simple is a series of if/else statements, more complicated AI might use neural networks and deep learning. AI could get information from any source a computer can: file input, cameras, microphones, etc.It produces information by taking some input, processing it in some way, and outputting it.It does not "understand" anything in the same way a human does, but rather algorithmically processes data it is given."

Table 8.1: **Labels for types of elicited explanatory models of AI systems.** N is number of responses, and Example contains a quote associated with the label. *Abstract* theories communicate what AI does vs. *Operational* theories which emphasize how AI works. *Sparse* and *Sophisticated* refer to low and high levels of technical depth and accuracy in elicited explanations reflectively.

> “ Usually it gets information from big datasets of training examples, similar to ones it needs to act on. There are different ways for it to produce information - it may do some clustering algorithm, use neural network, genetic algorithms or simply build decision tree based on previous answers. Most AI do limited number of tasks, so they recognise one of few commands. The ones which recognise speech likely try to represent sentence grammatical structure and use previous users’ dictionaries and set of predetermined algorithms. But i really do not know ”

P16’s explanation is long and contains examples of machine learning methods and speech recognition systems, the description of the latter however is ambiguous and likely inaccurate (by comparison to most existing speech recognition approaches); moreover, the participant explicitly indicates that they don’t know how it works despite offering an account.

By contrast, a *Sophisticated-Abstract* explanation provides accurate description coupled with description of what AI does but not how AI works (despite the questions specifically asking "How"). For example, P15 wrote:

> “ AI is trained on a large amount of data. The training will usually tell the machine what it needs to know and it will then produce information based on its training. It will identify similar features from the training dataset and the test dataset to make an analysis. ”

This participant, like P1, refers to generalization (similarity between train and test features), which requires knowledge about how machine learning models can be useful for real-world tasks, but provides no theory about how this is accomplished despite the posed question explicitly asking for it.

Human creativity in writing

Our two questions relating to human creativity in writing allow us to elicit unconstrained thoughts through open responses as well as anchor to classical constructs such as Novelty (historically new), Surprise (unexpected), and Value (useful to people) [51] via a multiple-choice item. We additionally included an *Other* field in the multiple-choice item, to allow participants to specify a different dimension if they felt that their concept was not adequately represented by these three, especially given the domain constraint. A majority of participants ($N = 12$) indicated that Novelty

was most important, followed by Value ($N = 6$), Surprise ($N = 4$), and one custom response: "evocative use of language", a domain specific attribute.

In the open responses, participants identified several features of human creative writing they considered important, which we coded as follows:

Freedom/Expression Five participants commented on personal expression and expressive freedom (P1, P3, P10, P20, P22). For example, P3 simply stated "freedom of expression", while P20 remarked on both aspects: "Creativity in writing for me is putting down a **personal**, immersive response to a prompt. So taking a spark of direction and going **wherever I want** from there."

Imagination/Fiction/Inspiration Five participants commented on imagination and fictive writing (P5, P6, P8, P9, P11), such as P5 who wrote "Letting my imagination go free, creating worlds and scenarios that don't exist." P6 illustrated this by comparison: "Creativity writing is the type of writing used in **stories, novels, poems, journals**. I keep it **separate from scientific writing**, which I don't consider as creative writing."

Additional thoughts: Structure/Clarity/Goal-directedness, Novelty, Unexpectedness, and Truth Still others commented on form, structure, flow, and direction. P2 indicated that creative writing involves "having a clear goal and many possible ways to accomplish that goal." The familiar dimensions of Novelty and Unexpectedness (surprise) also appeared in several comments. Finally, one participant alluded to "truth", perhaps indicating not the idea of verisimilitude (or similarity to reality) but "truth in fiction."

These features of human creative writing participants considered important are also summarized in Table 8.2 with additional examples.

AI creativity

When asked whether they thought AI could be creative, the majority of participants ($N = 17$) indicated Yes. We assessed this through analysis of an open-ended item, and some participants did indicate uncertainty by using words like "probably." We obtained the following codes through our analysis:

	N	Example
Freedom/Expression	5	"Creative writing, to me, is writing that embodies one's own novel artistic expression and is not primarily functional."
Imagination/Fiction/Inspiration	5	"It is an enjoyable activity that involves imagination and allows one to express his/her feelings."
Structure/Clarity/Goal-directedness	4	"Having words flow out and describe things in a satisfying way"
Novelty	4	"For me its coming up with new, innovative and engaging ways to write a story."
Unexpectedness	4	"Creativity in writing is using tropes and ideas in uncommon ways."
Truth/Reality	1	"Being able to capture truths about the real world through words"

Table 8.2: **Codes re: human creativity in writing.** N indicates number of participants, Example shows a corresponding quote.

Human-based Five participants (P3, P6, P8, P17, P19) noted that ostensibly creative AI is somehow modeling human creativity, and used this point to indicate that AI can be creative. For example, P3 wrote "probably, because it can mimic other human features."

Combinatorial Creativity and Uniqueness/Randomness Others pointed to the notion of combinatorial creativity [51] (P4, P9, P16, P21), suggesting that "It can be creative if it happens to combine things in a way that people wouldn't naturally consider" (P9). Relatedly, participants noted the opportunity for creativity arising from randomness. P15 notes that AI-generated ideas "can be completely illogical which is sometimes the best creativity."

Novelty/Surprise Three participants (P7, P10, P12) implicitly made the connection to novelty and surprise, remarking that AI "can be creative in the sense that it can produce novel solutions to problems," but also that "this presupposes a narrow conception of creativity" (P12).

Additional thoughts: Future creativity, and uncertainty Still others pointed toward future creative ability, due to the improvement of AI, such as P5: "Eventually, yes, but I'm not sure it can make big leaps in novelty in a single go." P18 expressed uncertainty about the question of whether AI can be creative, writing that they were "unsure what makes humans creative." Other participants expressed different kinds of uncertainty; P1 wrote that they believe that AI 'can be accurate" but they would

	N	Example	Yes	Unsure	No
Human-based	5	"Perhaps. The AI itself is a work of human creation. It might help a person to be creative in the same way that automatic writing, randomness, or writing constraints do."	4	0	1
Combinatorial	4	"I think that AI can create from a broad set of information that it has been given, but I do not think it can make up something new"	3	1	0
Uniqueness/Randomness	4	"Yes, it may suggest ideas or connections that a human might not usually make"	4	0	0
Other	3	"I do believe that AI can not be creative based off of my understanding. I believe that it can be accurate but I have a hard time imagining what a creative AI would look like."	0	1	2
Novelty/Surprise	3	"Yes, in the sense that AI can generate novel and surprising ideas without input from a human."	3	0	0
Gradually/Future	3	"Yes definitely! But to a certain extent. Because technology keeps on evolving and the internet is a very good example of it where it really helps in building creativity. But nonetheless, i think there is a limit to its creativity as compared to human but of course it will help a lot in enhancing creativity"	3	0	0
Unsure what makes humans creative	1	"I'm not sure because I'm not sure what processes allow humans to be creative"	0	1	0

Table 8.3: **Codes from open responses about AI creativity elicited from participants.** N is number of responses, Example contains a quote associated with the label, and Yes/No/Unsure are counts of categorical responses, for each label, from participants about whether they think AI can be creative. Some responses are labeled with more than one.

have a hard time "imagining what a creative AI would look like." P22 emphasised the fact that AI depends on "the input humans give to it" and noted that "if humans don't keep updating the inputs, it may not be creative anymore."

Table 8.3 provides other examples for codes mentioned above.

Human-AI Differences

We also elicited participants' thoughts about qualitative differences between human and AI text production. We assessed this through two items: one multiple choice to indicate the presence of a difference ("Do you think the way AI produces text is different from humans?"- Yes/No/Unsure), followed by an open-ended question prompting them to explain how and why (or why not) human and AI text production

mechanisms are different.

For the multiple choice item, 16 participants answered Yes (they think the way AI produces text is different from how humans do it), 5 answered "Unsure", and 2 participants answered "No." In surveying qualitative examples, we found diverse concepts of what differs between how humans and AI produces text, as well as effects of such differences.

Statistical/Data vs. Symbolic/Mind Eight participants (P1, P2, P7, P8, P, P15, P17, P18) pointed to the contents and mechanics of the text producers. They differentiated between data-driven and mind-driven text production, for example P1 wrote: "AI is taught to produce text based off of given **data**, an algorithm is used to produce text while a human creates text based off of their **mind**." Participants phrased this by contrasting generating text "statistically and linearly" with "using mental hierarchy of words" (P1), or indicating rule-based vs data-driven constraints. For example, P8 noted that "people can produce infinite correct and comprehensible outputs based on their knowledge of their native language's grammar and vocabulary, while AI will only be able to produce content based on its input."

We label this by combining the cues from participant responses and the traditional AI notions of symbolic processing vs. statistical modeling, two dominant paradigms in natural language processing.

World model and understanding Another thought from Five participants (P4, P5, P9, P11, P12) was that AI is lacking sufficient understanding of context, which comes from experiences in the world. P5 points out that AI "has not learned language by interacting with society and cultures, learning from family and personal experiences, or have the ability to draw on memory when responding in the same way humans do" while P4 appeals to sensorimotor functions:

> " AI produces numbers based on drawing patterns and similarities from the numbers in the dataset that it has been fed. It **doesn't understand and have visual representations in the brain that it then produces into motor action**, it's just reproducing what it's already seen. "

Not sure how humans do it Four participants (P9, P12, P16, P18) remarked that either they were unsure, or not much is known, about how humans produce text.

They differed in whether they thought this would extend to similarity or difference between humans and AI. For instance, P9 answered: "I'm not really sure how people "produce text"" and continued saying that since "AI learns from previous patterns" then maybe people in a similar way "learn from past experiences and instructions." P16 noted: "Not that is known how humans produce texts, so mechanisms developed independently are unlikely to be the same."

Complexity, performance Three participants (P6, P7, P11) commented on expected difference in the complexity of the produced text, or performance factors that might affect quality. P6 emphasized that "it depends on the level of development of the AI" explaining that "in the ideal case one should not be able to recognize a human produced text from a machine produced one." P11 used an example of google translate and it being unable to translate complex terms or understand the context, to point out that the difference between human and AI producing text might have to do with the complexity of the language .

Additional thoughts P14 argued that AI lacks "intentionality" due to not having "desires or beliefs", while P19 relatedly noted that AI cannot be "spontaneous" or "irrational" in its behavior as compared with humans. Two participants also made comments about formality in language. For example, P21 noted that "AI can only use what it has been taught or can access via some database while humans may access more informal or colloquial writing patterns."

Two comments noted that AI language systems are based on human-provided data. For example, P6 didn't expect a difference "because the information is mainly fed by humans." Finally, three participants made seemingly contradictory or unclear statements. For example, P22 indicated no difference, but then expressed an opinion on the difference of a somewhat ontological difference:

> " I think in some ways, each AI and humans communicate with our own languages and it's a mean of mutual understanding between them, so it's not that different. **It's just, humans don't operate the way AI does, and vice versa.** "

Table 8.4 provides other examples for examples and ratings mentioned above.

	N	Example	Yes	Unsure	No
Statistical/Data vs. Symbolic/Mind	8	"from my understanding, computers generate text statistically and linearly (vs. humans using mental hierarchy of words)"	7	1	0
World model and understanding	5	"It has not learned language by interacting with society and cultures, learning from family and personal experiences, or have the ability to draw on memory when responding in the same way humans do."	5	0	0
Not sure how humans do it	4	"Not that is known how humans produce texts, so mechanisms developed independently are unlikely to be the same. Also humans seem to be much better at text generating but may be it is because they have more and more diverse experience"	3	1	0
Other	3	"In a sense, yes, because it‘s more or less about encoding connections in memory, then following the connections through to retrieve this memory, or making predictions, based on what you already know."	1	1	1
Complexity, Performance	3	"I guess it depends on the level of development of the AI. In the ideal case one should not be able to recognize a human produced text from a machine produced one."	2	1	0
No intentionality	2	"It seems not as structured as humans? And it doesn't seem to have "intentionality" (e.g. they don't appear to have desires or beliefs when they try to make an argument)"	2	0	0
Formal vs. informal language/behavior	2	"AI can only use what it has been taught or can access via some database while humans may access more informal or colloquial writing patterns"	2	0	0
Human-based	2	"Because the information is mainly fed by humans"	0	1	1

Table 8.4: **Codes re: expected differences between human and AI text production, *before* writing.** N indicates number of participants, Example shows a corresponding quote, and Yes/No/Unsure are counts of categorical responses, for each label, from participants about whether they think AI text production is generally different than that of humans. Some responses are labeled with more than one.

8.5.2 Interacting with the system

All the data discussed in this subsection was received through observation of participants' interaction with the system and through verbal comments that participants made during the study. The comments were made either when participants were thinking aloud during a writing task, or as a reply to the interviewer's questions. Throughout the session interviewers asked general questions, like, "What are you thinking?" and "How is the writing going?": either once every 2 minutes or, if the participant seemed to be disturbed by being interrupted often, when the participant stopped writing. interviewers also asked more specific questions, like, "What do you think about that suggestion?", "Why are you laughing?", and "Tell me how you incorporated that suggestion."

We noted broadly different styles of overall participant writing and engagement with suggestions, described in detail below. We provide examples of participants who most clearly embodied the associated characteristics of each.

1. *Reactive writing*. Through observation, interviewers identified four participants for whom suggestions were actively shaping their story and helping them decide where the story was going (P17, P7, P10, P23). They wrote in a way that looked like a reaction to either suggestions of the system or as a reaction to the task. There were clearly effects from the pressure of time, conditions of the task, and their habits of writing. Some participants also mentioned that what they wrote was more like a "stream of consciousness" (P23).

2. *Proactive writing (with suggestions)*. Participants with clearly proactive writing (P2, P11, P15, P18, P20) wrote having a clear idea of what they wanted to write (having some horizon of their story) and how they wanted to do it (having their own process). They incorporated suggestions of **Editor-Red** at some particular points of their stories, either when there was the end of the scene or after they had exhausted the story horizon they had in mind. They did not let **Editor-Red** take over their process of writing. This type of writing was characterized by longer writing periods and hitting suggestions fewer times. For example, P18 requested suggestions only two times, and had a 15 minute writing process non-stop. P19 requested suggestions three times, and had longer writing periods, with one period being 9 minutes.

3. *Actively opposed to suggestions*. Four participants (P6, P8, P16, P22) were

generally *not willing to incorporate* the suggestions of **Editor-Red**. For example, P8 had their own idea of how they wanted to do things and explained the resistance to incorporate the system's suggestions, saying "I'm not a super suggestible person." P16 and P22 did not like the suggestions and did not include them in the writing. P6 wrote so as to improve the suggestions by the system, waiting for this to occur, and so didn't engage with the suggestions in the duration of writing their story with it.

We begin by exploring overall reasons, given certain types of suggestions and contexts, to incorporate or not to incorporate suggestions in the view of our participants. Subsequently, we describe and characterize instances of suggestion integration, often from more reactive and proactive writers (who did accept suggestions), in detail through the lens of *integrative leaps*.

Reasons to incorporate suggestions

Making a judgment as to whether the system's suggestions are in line with the participant's writing or too "out there" seemed to be an important axis along which participants, in the process of writing, were constantly making decisions about suggestion incorporation. Five participants specifically commented on the system's suggestions being in line with their writing (P1, P3, P4, P20, P21, P5). For example, P1 explained their decision to incorporate a suggestion because it was "thematically accurate and kind of good-to-keep-the-story-going description." P20 set a scene in their story and then hit the suggestions, as they wanted to see how **Editor-Red** "would interpret that." One of the suggestions of **Editor-Red** was "...I'm calling to let you know that you've been selected to the next round of the lottery," and P20 exclaimed: "Wow... it's a bit scary because I had thought of the lottery idea or just some other kind of news... yeah, so it's interesting that it immediately followed that train of thought about a lottery."

At the same time, some participants appreciated that the system was providing suggestions that were unexpected and not immediately related to their previous writing. For example, one participant explained that some suggestions, even though they seemed "absurd," were also so "detailed" and "specific" that it was "inspiring." As a result, even though some suggestions were "a little bit out there" to the extent that they would make them laugh, the suggestions would still give them "something to go" (P1).

We observed a subtle and variable trade-off between how creative or unusual suggestions were thought of as being by participants and how easy it was to incorporate them. The possibility of an easy transition towards incorporating the suggestions seemed to be a crucial factor. For example, P1 commented on one of the suggestions:

> " Some of the suggestions, would be either so similar to what I wrote that it doesn't seem worth incorporating or too creative, or I guess too hard to transition to. But that [suggestion] would logically be the next thing that I write about. "

Along these lines, some suggestions that might have been incorporated under the right circumstances by the corresponding writers were not integrated because of the time and effort it would require. P9, when choosing one of the two suggestion types (*Plot* and *Description*) that they equally liked, acknowledged that though they liked the suggestion that was "more fun, weird, crazy," it was such "a divergent shift" that the suggestion looked "too effortful to incorporate." Another participant explained that if they were to take "a much longer route" they might follow the system's suggestion of monster hunting (*Description* suggestion: "You are a member of a group of monster hunters.") as "it seems fun" but since they were short on time they decided not to explore this plot line (P5).

Reasons to not incorporate suggestions

Participants gave a wide variety of reasons as to why they might have been unwilling to incorporate suggestions. Some participants commented that the textual suggestions of **Editor-Red** looked "basic" (P15), "plain" (P6), "redundant" (P4), or were not "picking up the tone" of the story they were writing (P6). At some points in their writing, six participants commented that suggestions were not in line with what they wrote (P2, P6, P7, P8, P12, P22), complaining that the system was not able to see that "this is not where I'm going" (P7). Some of the suggestions also did not make sense to participants and were repetitive (P2, P6, P17), whereas for some participants the fact that some of the suggestions were not coherent was not an obstacle to engaging with their content. For example, when P7 was interacting with the system, it experienced a number of delays in producing suggestions, and finally a plot suggestion came out as: "not sure where to end Train to MIT for the first time not sure where to end Train to MIT for the first time not sure where to end Train to MIT for the first time not." The participant said laughing: "That's fine. I don't necessarily need it to be coherent" and expressed the readiness to carry on the writing.

One of the four participants who did not incorporate the system's suggestions found that the suggestions contained tropes and led to more "stereotype writing" (P8). Four participants also specifically pointed out that the system's suggestion distracted them from pursuing their own ideas (P2, P8, P12, P14). P8 explained that they felt they had "to tune out" the system's suggestions as they already had a picture of what they wanted to write in their head and it was easier for them to write "without the extra stuff." Some suggestions also did not come at the right time in the narrative: "Oh, wow, I would not think of cryogenic sleep! That's an interesting idea, but I didn't use it. I don't think it came at a good time in the story." (P1).

Visual suggestions were not incorporated if they were perceived as unrelated to the current writing (P14, P15, P6, P4). Participants also commented that some of the images were not only unrelated to the writing but also seemed arbitrarily constrained or homogeneous (e.g., demographically): "It's kind of strange, there's just a bunch of white guys staring at me and I don't know why" (P2) and "I'm confused ...And I'm curious as to why all the suggestions are very similar, and they are all images of straight blonde Caucasian women" (P5).

P14 considered the images "aesthetic" and "cute" but was following the idea they had in mind already. Some of the image suggestions, similarly to the text, did not come at the right time: "The pictures are cool. . . but this doesn't really fit with the character I have right now" (P15). P3 commented that even though they were not using visual suggestions, having them seemed "less daunting than having a white space in front of you."

Sound suggestions were the least used, sometimes due to a lack of relevance to the writing, either in content or in tone and style, and sometimes for other reasons. P5 described the sounds being not relevant and "random" (P5), and P9 explained why they were going to switch off the sounds:

> "The sounds are a bit dystopian. I feel if I use the sounds as an inspiration, I'd end up thinking of some sort of totalitarian government that's using lots of walkie talkies all the time and tracking people. It sounds very different to the vibe I was thinking in my mind. "

The sounds were also described as "aggressive" making it hard to focus (P8), "too much" (P9), and "distracting" (P12). All participants, at least once, switched off the

sounds at some point of their writing, although we observed instances where sounds were related to the story and/or resulted in incorporated suggestions.

Editor-Red as "support system"

Some participants described the overall experience of the system, which is not limited to just the relevance of the suggestions. For instance, some reported that they felt Editor-Red supported the writing process. P3 commented that the system was giving "good lines" and admitted that it helped "to continue along, where otherwise I think I will just stop writing." P3 continued saying that when they had absolutely no idea what to write, taking a word or a line from Editor-Red gave them "something to add" and then they "kept going, and kinda went from there." Two participants explained that Editor-Red helped them to feel less stuck in their writing (P12, P16). Even though P16 did not incorporate any of the suggestions, and during the experiment commented on the suggestions being "dumb," they expressed their surprise that, in the end, writing in Editor-Red did seem to help them feel "less stuck." P12 explained their feelings about one of the suggestions: "Although I wouldn't word-for-word take that, it, at least, redirects my attention from just being stuck in the kind of the crucial little loop to having somewhere else to go. So that's helpful to get unstuck, I suppose."

P23, reflecting on their experience of writing after the end of the session, admitted that they felt that inspiration came not directly from suggestions but rather suggestions made them think of something else and this is where ideas came from. P6 admitted that even though they did not use the suggestions of Editor-Red, writing in it actually "relieved some of the stress of writing." P6 further explained that even though the suggestions were not helpful to them personally, the system was still "creating that distraction, that was good for making the task a little bit more relaxing." They noted that interacting with Editor-Red really helped in mitigating the stress of writing, comparing it to a feeling of "petting a cat."

P1 described how interacting with the system changed their process of writing as they would "write for the suggestions." P1 explained that when they didn't want to continue writing as they could not think of what to say, being in the system would be a motivation to "write a few additional sentences in order to get a better suggestion" and to continue writing if they were unhappy with the suggestions that I received till the get a better suggestion. P1 found it "very helpful."

All in all, observing participants' interaction with the system enables us to map out the multiplicity of tactics that participants engage in while performing the task of writing a text. Participants borrow words and lines from the system's suggestions, get inspired by the system's suggestions directly or indirectly, use ideas from the suggestions as reference points to produce their own ideas, or receive psychological support from the system (e.g. reducing the stress of writing by, for example, giving them something to focus on besides their own feelings of getting stuck).

Willingness to cooperate with Editor-Red

Participants practiced different patterns of engagement with the system hoping that it would provide suggestions that better served their purposes. Some were willing to wait longer for the system to start giving better suggestions, occasionally under the assumption that allowing more time would give the system an opportunity to catch up with the participant's writing. For instance, P1 said: "This makes me think I didn't wait long enough because now everything's about phones. So maybe I should wait a little longer."

Some also decided to keep writing in order to give more information to the system (P1, P19, P8). P1 reasoned that the information that they were "feeding it" might be "not substantial now". When P7 received another round of suggestions, one sentence that particularly got her attention was the phrase "I could feel the horn blaring in the distance." This phrase was almost identical to what the participant had already written, with the system having changed precisely one thing: the original sentence was "I *hear* the horn blaring in the distance." P7 pondered:

> " I see it changed some of the words around. *I could feel the horn blaring*, which is interesting. It's much more visceral than *I can hear it blaring*... Yeah, *hear* is not quite the right word... I will just put *feel* for now. "

This is an example of how a participant is willing to cooperate with the system and make sense of its contributions, even when someone else might consider the re-phrasing to be overly subtle (merely a lexical substitution) and not valuable to developing the story further.

Integrating the system's suggestions

To provide a more granular exposition of the suggestion integration patterns we observed, we enumerate and detail a collection of *integrative leaps*. These leaps describe the different kinds of interpretation and expression involved in incorporating aspects of suggestions into the developing story, in particular how and how much participants alter the meaning and structure of their narratives when doing so. We use them as fine-grained windows into the mechanics of the most visible examples of this incorporative process, those we are able to access through our observational methodology.

Our data on suggestion integration contains 47 instances of integrative leaps from 19 (out of 23) participants; P6, P8, P16, and P22 did not appear to incorporate **Editor-Red**'s suggestions in any identifiable way). These are examples that the researchers conducting study sessions identified of participants engaging with and actively incorporating suggestions from the system. Participants often explicitly commented and explained why and how they incorporated suggestions, as they were encouraged to do, and we report on their interpretation of this process in addition to our observations and analysis.

Types of integrative leaps

The integrative leaps can be analyzed along a number of axes. First, we consider the "edit" distance (e.g. lexical, semantic, etc.) between the suggestion as presented to the user and as incorporated into the story. We broadly characterize these as *direct integration* ($N = 30$; e.g., verbatim or restructured verbatim for a textual suggestion or a textual analogue of the object or idea represented in a visual or auditory suggestion) or *indirect integration* ($N = 17$), where it often would be impossible to capture this integration if we did not have the participants' explanations, due to the modifications they made in the process of suggestion incorporation.

Second, we look at how incorporated suggestions relate to global aspects of their story's direction and most prominent elements. When participants used suggestions to explore new lines of narration, we call it *exploratory integration* ($N = 28$, shown on left half of both figures), in contrast to taking suggestions to continue with their chosen narrative by adding more details, which we call *confirmatory integration* ($N = 19$, shown on right half of both figures).

Finally, with the view that suggestions are intended to ease cognitive inertia in the writing process, we attend to the role they play in creative problem solving. Do they simply solve a localized problem by "closing" some aspect of the narrative in a necessary, analytical, or expected way? For example, naming a character that has already been described, or explaining why a character went from place A to place B if both of those events have been established. Or do they "open" up options to consider, resulting in abstract, novel, or unexpected events, patterns, or directions? We describe these as *convergent integration* ($N = 31$, shown on the bottom half of both figures) and *divergent integration, shown on the top half of both figures* ($N = 16$) respectively. While these often overlap with confirmatory and exploratory integrations respectively, there were a few cases in our coding process where we found it useful to explicitly make a distinction between these two dimensions, in order to better explain behaviors that we observed. For example, two of the six integrations we detail in the following section are ones we labeled, through an iterative process, as *exploratory* and *convergent*. In these leaps, participants may use suggestions to both pivot at a narrative level, and solve a local problem within this context. Although our categories are still relatively broad and cannot cover all the differences between integrations that we observed, we sought to sufficiently represent the most prominent aspects of integrations with these labels.

Integrative leaps

In this section we review several examples of *integrative leaps*, identifying them along the aforementioned axes as well as describing the participants' interpretation and comments. We summarize each instance in a discrete box that clearly identifies the input text (before the suggestion), the suggestion at hand, the text after integration, the participants' explanation, and our labels (for example, Integration 1). When participants identified that they were prompted by visual or auditory suggestions, we include thumbnails or links for the reader to review.

P3, following the "The phone began to ring" prompt, was writing an intense story of a mother getting a phone call from her estranged son. Through a number of previous suggestion interactions, the participant wrote a story where the son on the phone call was in trouble, as some people were holding a gun to his head and demanding some information he didn't have. The next round of suggestions contained "I'm just a normal person who is in a hurry to get home." Following that, the participant

wrote "She freezes. What is he talking about? This isn't making any sense... yes, she has an estranged relationship with her son, but they are normal people." As the participant explained, the phrase in the suggestion "I'm just a normal person" stood out to them and prompted them to develop it into the mother's inner thoughts trying to come to terms with the fact that her son and she herself are probably in big trouble. We labeled this example as *direct* (almost verbatim integration: **normal person** to **normal people**), *exploratory* (the participant did not have a clear idea of the narrative) and *convergent* (solving a local question of how the main character reacts to the news that her son is in trouble). See integration 1 for more details.

Integration 1

Input (summary): [Emotional dialogue, son is held captive...] ..."What?" She replied back. "Who are you talking about?" "It's them," he whimpered. "But I-I don't have anything to tell them. I don't have the information they're looking for."
Suggestion: *Plot.* I'm just a **normal person** who is in a hurry to get home...
Integration: She freezes. What is he talking about? This isn't making any sense... yes, she has an estranged relationship with her son, but **they are normal people.** "You're not making any sense." "It's **not normal. None of this is normal**" he responds shakily. She hears a scream and the phone cuts out.
Explanation:

> "...I'm just thinking about how to continue this story but I don't really have much... but the suggestion under *Plot* is giving me some... you know, "**I'm just a normal person**" line...I still don't have any sort of direction with the story... this feature seems to be good to help me, like, continue along, where otherwise I think I will just stop writing... "

> " ...it just kinda stood out to me in relationship to this story... cause this story, it seems like again the mom is just a **normal person**, so if she is getting this phone call from her son, it doesn't make any sense, we are just **normal people**, so I thought I would incorporate that" "

Our labels:

- ***Direct***: the incorporation is almost verbatim (**I'm just a normal person** to **they are normal people**).
- ***Exploratory***: the writer, from their own remarks, does not have a clear narrative direction that this suggestion would reinforce. Rather, it gives them a possible next step to build on.
- ***Convergent***: the suggestion helps to solve a local problem in a concrete way (continuing the story further).

P21 was developing a story from the prompt "The phone began to ring" and was

describing a call from the best friend of the main character. P21 wrote the first part of the dialogue "“Wait why were you in the hospital?” I asked my friend” and the subsequent round of the suggestions contained images with cars. The participant immediately took on the idea: “I’m seeing cars, so maybe he was in a car crash.” and to continue the dialogue, P21 wrote: “My sister was in a car crash. She’s okay, but she broke a rib.” Since the suggestions helped to keep the writing going and did not prompt the participant into a new avenue of thought, as well as being a textual representation of a suggested visual object, this entry is labeled as *direct* (**images of cars** to **car crash**), *confirmatory* (reinforces the existing narrative), and *convergent* (closes a local question of why the person is in the hospital). We report details in integration 2.

Integration 2

Input (summary): [Best friend phone call. . .] . . . “I ran into your ex-boyfriend at the hospital”. I was in shock. I hadn’t seen him since 4 years ago when he left me to run away to Cuba with some new woman.

Suggestion: (Images) {  ,  , . . . }

Integration: “Wait why were you in the hospital?” I asked my friend. “My sister was in **car** crash. She’s okay, but she broke a rib.” I completely forgot about what she said about my ex being in the area, assuming it was hours ago, and rushed to the hospital. We were neighbors growing up, so I was pretty close with her sister too.

Explanation: “I’m seeing **cars**, so maybe she was in a **car** crash.”

Our labels:

- ***Direct***: direct representation of visually represented object.
- ***Confirmatory***: reinforces the existing narrative.
- ***Convergent***: closes a local question: i.e. "why?", "what happened?" regarding a character in the story.

P4, following the prompt "The phone began to ring," was developing a story about a police detective who called the main character and asked to come to the police station because their sister was in trouble. P4 felt unsure as to how to continue and what it could be that the detective could have been accusing their sister of. This participant was really perplexed with what in their previous writing could have prompted the subsequent suggestions involving zoos, animals, and tropical places (these were in the retrieved images) but still decided to go ahead and integrate the suggestions into

their story. In the end, P4 wrote (about the police detective): “He appeared a bit nervous. He told me that he suspects my sister may have stolen an elephant from the zoo when she was studying abroad in India. I felt shocked.” P4 explained their reasoning in integrating the system’s suggestion: “I don’t know why these images popped up and how they are related to what I wrote before. But I saw the elephants and some kind of more tropical places and so it kind of made me think of ...I don’t know I was thinking what could it she possibly have done wrong that she could be in trouble and so the elephant was standing out to me, so I chose to say that “she stole an elephant” and I was thinking where elephants are and I know that there are a lot of elephants, like Indian elephants, so that’s why I said that.” See integration 3 for more details.

The participant concluded their story by writing, in an attempt to rationalize and make sense of the participation of the elephant in their story:

> “ I knew my sister loved animals, especially larger ones, but I never would have expected this. Where would she have left it? I had so many questions. I asked if I could talk to my sister. "Did you steal an elephant??" "I don’t know what he’s talking about. I’ve never seen it before." ”

Later on in P21’s story (previously described in integration 2), they were describing a character driving to the hospital and the system gave auditory suggestions that P21 described as chanting and explained: “There is chanting happening, it makes me think she got into traffic because there’s a protest happening, ...or a parade.” So P21 wrote in their text: “In my mad dash to get to the hospital, I forgot that the 4th of July parade was happening today just blocks down from the hospital. I’m stuck at an intersection where the parade is passing by...” In this example, sound suggestions prompted the participant to think about what could have caused the traffic, so call the integration indirect. The integration of this suggestion also significantly altered the course of the plot (exploratory) creating new avenues of the story development (divergent). More details are in integration 4.

P5 was writing a slow-paced descriptive story using the prompt "A train arrives at the station." At some point, the protagonist was stopped by an officer and told that the train would not be boarding as there was some issues. P5 requested a suggestion and one of the suggestions was “I had been waiting for this moment for years." The participant continued developing their story and wrote: “The train was already late

Integration 3

Input (summary):[Going to meet detective] . . . I took the train to get to the police station. When I arrived, the detective met me at the door. He appeared

Suggestion:

Plot. to be a little **bit nervous** but seemed to calm down when I asked him. . .

(Images) {  ,  , . . . }

Integration: He appeared a **bit nervous**. He told me that he suspects my sister may have stolen an **elephant** from **the zoo** when she was studying abroad in India. I felt shocked.

Explanation:

> " I guess I used the first section of the plot to write "he appeared a **bit nervous**". . . these images I don't know why they popped up and how they are related to what I wrote before. But I saw the elephants and some kind of more tropical places and so it kind of made me think of ...I don't know I was thinking what could it she possibly have done wrong that she could be in trouble and so the elephant was standing out to me, so I chose to say that "she stole an elephant" and I was thinking where elephants are and I know that there are a lot of elephants, like Indian elephants, so that's why I said that "

Our labels:

- ***Direct***: **elephant** (from images), **nervous** (from text)
- ***Exploratory***: the elephant, India, studying abroad are substantially new aspects of the plot at this point
- ***Divergent***: does somewhat "close" a local question (what did she do?), however in a very unexpected way that raises many more questions than it answers

Integration 4

Input (summary): [Continuation (see integration 2] . . . We were neighbors growing up, so I was pretty close with her sister too.
Suggestion: Sound 1 (crowd call and response); Sound 2 (crowd cheering)
Integration: In my mad dash to get to the hospital, I forgot that the 4th of July **parade** was happening today just blocks down from the hospital. I'm stuck at an intersection where the parade is passing by, so I have no choice but to watch the high school band and floats made by various organizations go by.
Explanation: "There is chanting happening, um It makes me think she got into traffic because there's a protest happening, . . . or a parade."
Our labels:

- ***Indirect***: no words or concepts are directly applied; abstract link must be explained by participant (**sound of crowd chanting** to **4th of July parade**)
- ***Exploratory***: altered the course of the plot significantly; narrator eventually turned around and went home after several experiences in the parade
- ***Divergent***: not necessary or expected; creates a twist to develop further; opens up new questions for story

and now this; who knows how long before I get on board?! I can't be late... maybe if I start now, I can drive over to... no, no, no. I'll never make it that way." To the interviewer who ran the session, there was no obvious connection between the suggestion and what the participant subsequently wrote. However, P5 explained that the suggestion "I had been waiting for this moment for years" made them think "more of a frustration for the train being late" and they imagined that there was something that the character was supposed to get to on time in another city. So this idea was translated into making the character impatient.

Following the prompt "A train arrives at the station" P9 started writing a fantasy story about frogs waiting for their tadpoles to get back from Tadpole Kindergarten. Another round of suggestions read: "sound of a bell ringing and the frog who was holding the bell was holding a tray of frogs..." As the participant explained, the specific "sound of a bell ringing" in the suggestion made them think about sounds in general and what kind of sounds can be in the setting of their story. The participant wrote "Once inside the parlour they were all taken back by the ringing and clattering of dishes and trays. " Here, the participant took a concrete description of sound (sound of a bell ringing) and then made a shift from concrete description to the general concept of sound and made a decision about what kind of particular sound will be in their

Integration 5

Input (summary): [Train is running late…] "…There is a matter we have to attend to first before we will let anyone be checked in," said the officer calmly.
Suggestion: *Description.* I had been waiting for this moment for years.
Integration: The train was already late and now this; who knows how long before I get on board?! I can't be late… maybe if I start now, I can drive over to… no, no, no. I'll never make it that way.
Explanation: "So in this case rather than "waiting for this moment for years," I'm thinking more of, like, a frustration for the train being late and now more delays and there's like something that character was supposed to be trying to get to on time in another city. So it's going to [make] the character impatient."
Our labels:

- ***Indirect***: **waiting** for **years** to frustration, impatience
- ***Exploratory***: switches from describing scene and events to narrating internal dialogue about the character's feelings
- ***Convergent***: an expected reaction to the situation that describes the effect of the train's lateness

story ("clattering of dishes and trays").

We summarize these axes of integrative leaps in two figures. Fig. 8-5 shows *direct* integrative leaps, and Fig. 8-6 shows *indirect* integrative leaps. In both figures, left is *exploratory*, right is *confirmatory*, top is *divergent*, and bottom is *convergent*. Participant IDs are noted along the horizontal axis, aligned to the corresponding instances. We can see a few patterns when surveying these leaps in total. For example, participants generally made more *direct* leaps than *indirect* leaps, but these are also related to the other dimensions: most direct leaps were also *convergent*, addressing necessary and local narrative features, though there are several exceptions. Conversely, *indirect* leaps are slightly biased toward *divergent* integrations. Similarly, the *exploratory* label often coincides with *divergent*, but we can see several exceptions to this. On the sides, we include high-level descriptions of what each integration contributes to the developing story, with illustrative examples provided for each quadrant.

8.5.3 Outcome Evaluations

In addition to participants commenting on their experience during the interaction with the system, we were also interested in capturing overall impressions and specific

Integration 6

Input (summary): [Tadpoles taking the train back home from Kindergarten...] ...Once inside the parlour they were all taken back by the

Suggestion: sound of a bell ringing and the frog who was holding the bell was holding a tray of frogs and he was holding a tray of tadpoles who were all waiting for the new tadpoles

Integration: Once inside the parlour they were all taken back by the ringing and clattering of dishes and trays. Frogs and toads were excitedly gulping down the various fly filled delights inside. “Georgia! Barry! Tadette!” beamed Mr Willeker. “You all look so well!” Please take a look at the menu.

Explanation:

> “ I found that interesting as I guess it made me think more of like the sounds that could be inside this parlor or something ... because, basically, I was going to end up doing another long description that’s probably quite boring. Probably similar to my previous thing I was writing, but I could then think about the sounds like clattering plates. ”

Our labels:

- ***Indirect***: **sound of a bell + tray** to **ringing and clattering of dishes and trays**; tray is shared, but most of it is indirect
- ***Exploratory***: participant uses it to lead in a different description (see their explanation)
- ***Divergent***: a level of description that opens up commentary about the food and environment, etc.; not necessary or expected

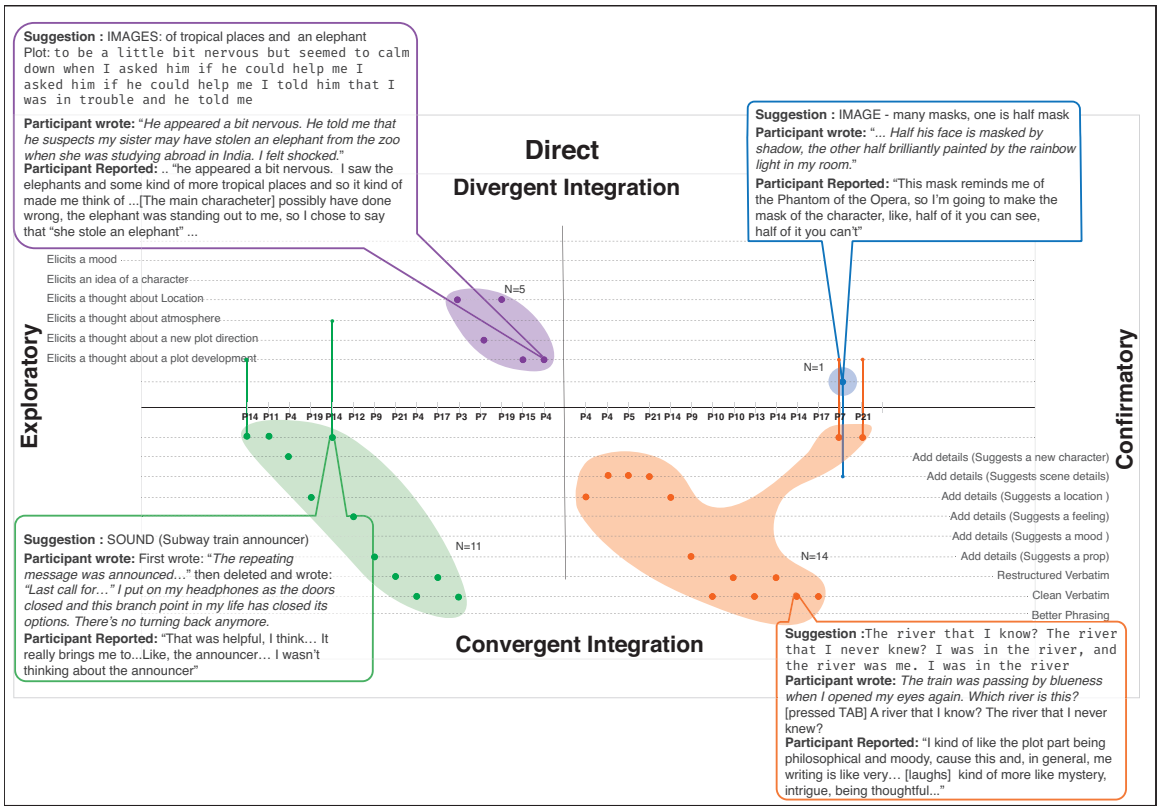

Figure 8-5: Diagram of exploratory/confirmatory and divergent/convergent *direct* integrative leaps made by the participants.

thoughts post-writing related to aspects of the prior assumptions and participant explanatory models we captured.

General Impressions

We obtained general impressions with one open-ended question ("What are your impressions from using **Editor-Red**?"), and subsequently tagged these with overall sentimental valence (summary in Table 8.5). We found that a majority of participants ($N = 13$) noted largely positive experiences with the interface, offering considerably different reasons. Some participants felt the suggestions were impressive or surprisingly relevant; for example, P20 noted that they were "pleasantly surprised by how spot on the predictions were at times," and P12 wrote "I was impressed by its ability to generate sentences based on the context."

Some participants indicated that suggestions were directly helpful. P1 made an explicit connection to writer's block, saying "I really enjoyed the visuals and suggestions.

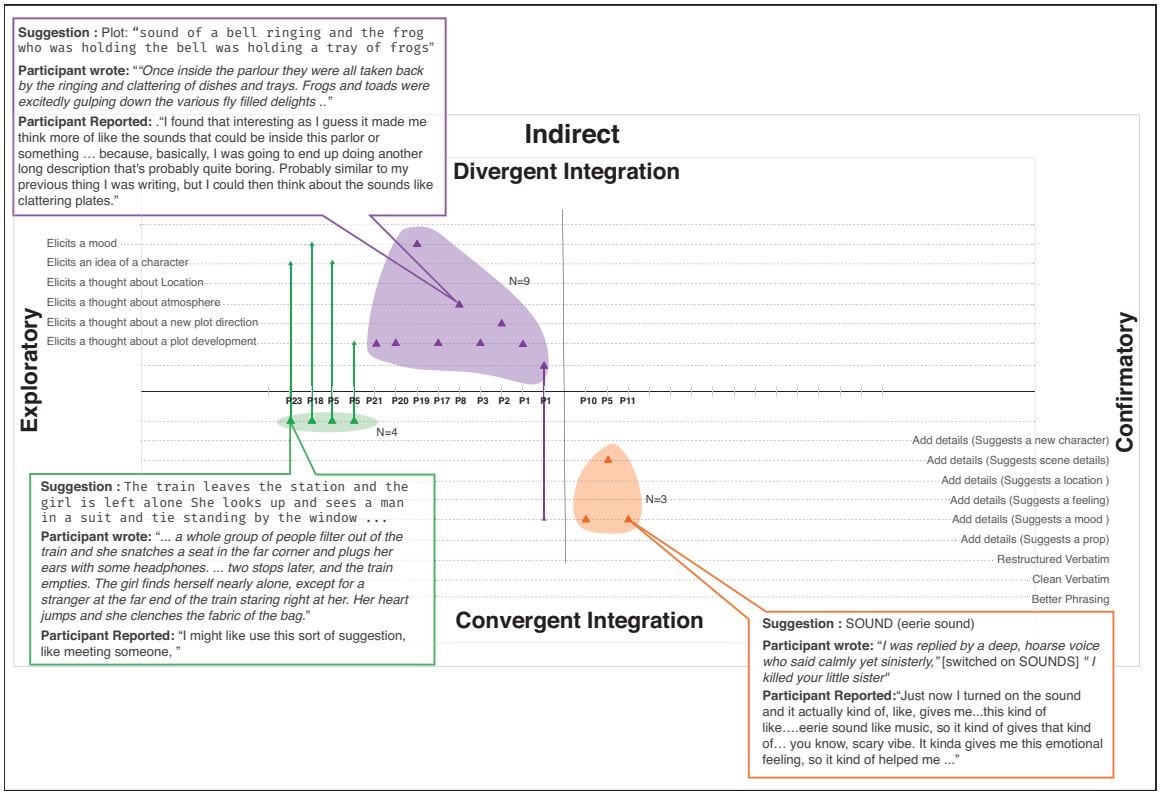

Figure 8-6: Diagram of exploratory/confirmatory and divergent/convergent *indirect* integrative leaps made by the participants.

I'm not a very visual person and struggle with writer's block, so this was very helpful."

Others found suggestions indirectly helpful, due to mood enhancement or some other affective effects. P11 wrote that it was "Very helpful in bringing the mood up in writing. It helps create the ambience and emotions needed for the writing." P5 and P9 both noted that the interaction was "fun" in addition to noting other effects beyond helpfulness of suggestions. P9 wrote:

> " It was fun/funny. The plot suggestions often didn't make sense but the description ones **were either useful/thought-provoking or amusing to read** even if I didn't use any part. The sound wasn't directly influencing my ideas but having background noise was **relaxing**. The pictures sometimes were relevant and sometimes not, so I didn't stare at them too long when they changed. "

P15 indicated that suggestions weren't always relevant, but that they were willing to do the work to find ways to incorporate them, which ultimately did make them helpful:

> " It was definitely a place to draw inspiration from. **I would not use the suggestions as they are, but with some modifications, the ideas presented were definitely helpful.** The sounds are calming and peaceful. The background images are a hit or miss because they don't tell the same story as the text I am writing, but **if I thought about the images a bit more creatively rather than literally, they were a bit more helpful.** I don't ever see myself using the overlay version of the images though. "

We annotated several responses as being overall negative ($N = 6$). Most of these also included comments about suggestions being less relevant, but participants did not indicate that they were necessarily able to integrate them despite this (in contrast to P15). For example, P16 wrote:

> " good idea but implementation worse than I expected. Images and text suggestions did not match storyline well, some proposed options were too exotic like "ustrophobia", selection tool to find images was highlighting parts of word not whole word, sounds were not very relevant, "

P18 also commented on suggestion relevance, but focusing on the writing context and timing rather than the overall relevance, noting that "I think it's interesting, but I don't see the plot or descriptions as being particularly helpful unless it's for a writing prompt. Once you get into the story, the suggestions are not very relevant."

Still other participants seemed neutral about their experience with the interface ($N = 4$), either providing little direction in terms of what worked for them and didn't, or explicitly accounting for both strengths and weaknesses in a balanced manner. P14 wrote:

> " It was sometimes helpful when I don't have any ideas, but not super helpful... Sometimes the text might not make sense... The images are usually slightly off topic, but that could be helpful in giving me new ideas (they are very aesthetic/instagram-feel)... I liked the sounds when they existed (it really does bring me to the place) "

Suggestion Helpfulness

Participants' overall impressions contained general indications of suggestion helpfulness, but we were also interested in obtaining fine-grained reflections to better elucidate the conditions and motivations involved in this. Again, participants diverged in what they found helpful and why. Here, we coded responses as indicating

	N	Example
Positive	13	"In Editor-Red I felt like I was writing in a time travel machine rather than staring at a blank page. I think it helped me feel a lot more grounded, present in the moment and in my body rather than just a disembodied brain trying to force words on a page."
Negative	6	"I didn't find it very helpful, but I could see how some features might be useful if developed further. The suggestions didn't seem to take into account the total content of what I had written, and so they seemed irrelevant and even distracting (for example, a phone scrolling instragram on a beautiful summer day when the phone in my story was a handset phone in a hotel room)"
Neutral	4	"It seemed fairly easy to use. I primarily was looking at the Plot section, possibly because it was more relevant to the given prompt while the Description section seemed to be more like different prompts. "

Table 8.5: **Our sentiment labels for overall impressions from participants of writing with Editor-Red.** N indicates number of responses, Example shows a corresponding quote.

suggestions were **D**efinitely helpful ($N = 2$), **H**elpful ($N = 5$), **S**omewhat helpful ($N = 11$), **R**arely helpful ($N = 1$), or **Not** helpful ($N = 4$), see Table 8.6 for full details. In addition to this, we found several different ways that suggestions were or were not helpful.

New ideas, phrases, words A common indication was that suggestions yielded new ideas, possibly in phrase form but sometimes even as a word that was contextually useful. Although the suggestions typically contained full sentences, subsets of these were more often described by participants as being helpful. The helpfulness of these ideas also extended to their presentations in other modalities (images and sounds). P21:

> " The suggestions were helpful. At one point, I felt that there had not yet been enough action in the story and the Editor suggested an event to happen. Another time, the sounds that were playing triggered an idea for the next setting of the story. Other times, even if I didn't like the suggestion I was given, it would still give me an idea for how to proceed. "

Relevance Another reason suggestions were or were not helpful had to do with how relevant they were regarded as being in the context of writing. Most participants were mixed on this. For example, P8 wrote that the sounds seemed in-tune with the tone of the story they were writing, but described challenges to overall relevance of

suggestions:

> " I was impressed that the sound suggestions seemed to pick up on the creepy, suspenseful tone of the story right away, and it could be helpful if the image suggestions followed the tone more closely. It kept showing me pictures of smart phones, which was not helpful. It would have been more helpful to see images of places and people for inspiration about how to describe their features. it would also be more helpful if the suggested images were more varied rather than all being pretty similar. That way the writer could choose from potentially useful images (and maybe even indicate which ones were more useful to see more like that?) "

P2, by contrast, found the sounds very distracting but the plot-level suggestions occasionally helpful:

> " the plot suggestions were occasionally helpful, but the descriptions were usually completely off the mark; the background images were aesthetic but not totally related, and the sounds were very distracting "

Mood, continuity, and flow Some participants expressed effects on process rather than on content, as we also observed in their comments during the interaction. They suggested benefits beyond the direct application of suggestions, for example P1 noted:

> " For the most part I think that they were helpful. Even if I didn't use every idea that was suggested to me, I was inspired by their mood. "

P11 also pointed to this, writing about flow, mood, and ambience (the latter of which did seem to depend on relevance to the general topic of their progressing story):

> " Yes! They are definitely helpful! I think it helps to prompt me to think about what to write next and keep the mood on writing. It helps to keep the ambience according to the topic Im writing as well. "

Redundant and repetitive suggestions Several participants noted the presence of redundant or repetitive suggestions, either repeating the content of the text in some form, or containing internal repetition. P15 wrote:

> “ The suggestions could definitely be improved. For example, the first suggestion I had kept repeating the same things in the plot box and the description box was very plain and essentially what I had already written. However, the second time I used a suggestion went better and I was able to draw inspiration from the images and plot box. ”

Additional thoughts Two participants found suggestions not or rarely helpful during the course of their writing, even when they might identify their potential value. Table 8.6 contains such an example, as well as a summary of all the other themes.

Outcome Ownership

Almost all participants ($N = 22$) indicated that they felt the outcome text was primarily or entirely theirs, citing a few different factors to reason about their ownership. No participants reported not feeling ownership, while one expressed a little uncertainty.

Only took some phrases/words/ideas Most participants pointed to their cognitive and creative work in absorbing suggestions in the form of phrases, words, and ideas into their stories. P5 made a comparison to a collaborative process with another writer:

> “ I would because I didn’t take suggestions word-for-word except for one short phrase. Otherwise, it was like me bouncing ideas off a friend rather than the friend actually writing prose for me. ”

Ideas were primarily theirs Other participants made the perhaps related argument that general, global aspects of the stories were their own. From P13:

> “ Yes because while I used the suggestions somewhat, the general storyline was my own. ”

Autonomy and authorial discretion P22 pointed to authorial discretion, i.e. "final cut", as the source of their ownership over the outcome: "I think whatever platform you use to brainstorm, in the end of the day, you are the decision maker to put that into your writings or not."

P21 referenced their work in not only deciding *how* a suggestion might be integrated, which we have discussed earlier, but perhaps whether or not to integrate it altogether:

	N	Example	D	H	S	R	Nt
New ideas/phrases/words	10	"The plot descriptions and text suggestions were helpful and creative. Some of the images, however were a bit generic. I mentioned a phone and the grid overlay just shoved several iterations of smartphones/, it would be nice if it could show different types of telephones, and in that way, allow the writer to have a visual reference, in case the writer wants to describe the object in more detail."	1	3	6	0	0
Relevance	6	"They were sometimes helpful and sometimes not. The suggestions were sometimes incoherent sentences, and they sometimes would not fit in well with the rest of the story."	0	0	4	0	2
Mood, continuity, and flow	4	"The suggestions were helpful. At one point, I felt that there had not yet been enough action in the story and the Editor suggested an event to happen. Another time, the sounds that were playing triggered an idea for the next setting of the story. Other times, even if I didn't like the suggestion I was given, it would still give me an idea for how to proceed."	1	3	0	0	0
Redundant/Repetitive	3	"I took some of the suggestions, but there was a lot of repetition (it kept wanting me to include a "hospital" scene, for instance). I mostly used it before a change in the narrative and it helped me think about plot development. "	0	0	2	0	1
Other	2	"For the suggestions menu, it didn't really work with me, only 1 out of 3 chances did I get to use it but it's quite similar with what the introduction video explained and I think they're helpful too in times of writer's block appear. "	0	0	0	1	1

Table 8.6: **Codes from whether and how suggestions were helpful.** N indicates number of participants, Example shows a corresponding quote. Here, we coded responses as indicating suggestions were **D**efinitely helpful ($N = 2$), **H**elpful ($N = 5$), **S**omewhat helpful ($N = 11$), **R**arely helpful ($N = 1$), or **No**t helpful ($N = 4$). Some responses are labeled with more than one.

> " I would call it my own. While I did receive inspiration from the Editor, there was nothing that I took verbatim from the Editor. I didn't include anything without putting thought into whether or not it would add to the story "

Similar to real-world encounters Some participants compared the references obtained through working with **Editor-Red** to real-world encounters or analogous explorations of open domains like the internet. P11 wrote:

> " . . . Just like when we try to find ideas through browsing the internet, or just **having a walk outside to create the mood and inspiration to write.** But editor Red makes it more efficient and **easier to find idea since it is all in one platform.** "

Suggestions were not helpful Some participants felt ownership due to not incorporating any suggestions. P8 very simply stated: "I didn't really use any of the suggestions."

Additional thoughts P1 indicated "I would call it my own, but acknowledge the suggestions that were made to me." They were the only participant that suggested such an acknowledgement. All these codes are summarized in Table 8.7.

Differences from Initial Expectations

To capture the ways in which writing with our interface qualitatively differed from their expectations of it, we posed an open-ended question. We coded the responses both for overall difference (Yes/No/Unsure) and themes that explain how or why, if so (see Table 8.8). A majority of participants ($N = 13$) indicated a difference from their expectations, with many also indicating that it was overall similar ($N = 9$). We coded one response as being unsure: P10 wrote "It was a novel experience. I had no expectations because I wasn't sure what to expect."

Similar 6 participants noted that the experience was overall similar to their expectations. P14 wrote:

	N	Example	Yes	Unsure
Only took some phrases/words/ideas	10	"Since some of the suggestions were pretty far out there, I would say that what I wrote is my own. However, there were interesting moments where I did copy a phrase that the system proposed. This felt more like plagiarizing than, say, picking a different word from a thesaurus. But I don't think the system can take too much credit since it was generating ideas based on my writing."	10	0
Other	3	"It depends. If I only used it to provide visual references and sounds for describing a scene, then would still call it my own, but if i got vital plot points from the editor suggestions, then I wouldn't fully call it my own work."	2	1
Ideas were primarily theirs	3	"It definitely feels very much like my own because almost all the words were mine, the story and progression is mine, the tone is very much mine."	3	0
Autonomy and authorial discretion	3	"I would call it my own because even though I did use the features to brainstorm, but I mostly write it on my own and I think whatever platform you use to brainstorm, in the end of the day, you are the decision maker to put that into your writings or not. "	3	0
Similar to real-world encounters	2	"Yes. Because it helps me find an idea, but I was the one who developed the story and make the story coherent. I think Editor Red is just a platform that helps a lot in creating ideas and mood in writing, not to help in writing the whole story itself. Just like when we try to find ideas through browsing the internet, or just haviing a walk outside to create the mood and inspiration to write. But editor Red makes it more efficient and easier to find idea since it is all in one platform."	2	0
Suggestions were not helpful	2	"Yes, because it was produced by me entirely "	2	0

Table 8.7: **Codes from open responses regarding ownership over outcomes *after* writing.** N indicates number of participants, Example shows a corresponding quote, and Yes/Unsure are counts of participants reporting a feeling of ownership or being unsure (no participants responded no).

> “ I have worked with language models before (I’ve played around with Writing with Transformer-type websites, using GPT2 for applications, and I’ve actually done an NLP externship that involved making image recommendations based on text haha using Unsplash too) so it was similar to what I expected in that it sometimes doesn’t make sense, says things that are not super related, but could be coherent/interesting sometimes. ”

Less relevant Participants remarked that suggestions were less relevant than they imagined initially, if not always then some of the time at least. P4 remarked:

> “ Sometimes I had no idea where the pictures in the grid came from, because sometimes they seemed relevant to what I had written and then sometimes it seemed like they were completely random. I expected the plot suggestions to be a little less repetitive. ”

Less out-there P2, among others, noted that suggestions were less big-picture or less out-there (which we use to imply further from the written narrative) than they anticipated:

> “ fairly differently... I guess I was expecting some bigger-picture feedback, like larger plot suggestions or thematic images (like outer space for a space-related story ... though how clear would it be to AI that the story is about space?) ”

More creative/intelligent Two participants noted suggestions being more creative or intelligent than they expected. P1 wrote:

> “ It was surprising to see the intelligence of the AI and the creativeness of the suggestions, for example "cryogenic sleep" was a very novel idea suggested to me. ”

Less subtlety/control P7 expected and desired differences in what parts of their writing the system attended to, indicating that they might like to do this through some control input:

> “ I thought it would just look at the last few words that I had written and it would ideate on those ideas. Sometimes that was the case, particularly for the image suggestions. However, the text suggestions would sometimes go all the way back to the beginning of the story. I think that I wanted more control over where the AI was paying attention, but it otherwise did what I thought it would do. ”

Slower P16 remarked that the interface was slower to respond than they expected, in addition to suggestions being less relevant: "its interface acted as expected but I expected it to give suggestions faster and those be more relevant"

Not as directly usable/helpful P9 and P3 noted that suggestions were not as directly usable or helpful as expected. P3 wrote: "I thought it would give me more suggestions/sentences that I would just copy into my writing directly. It was more of abridged words/phrases."

Human-AI Differences

We assessed differences in participants' practical expectations of our system and human writing partners with a counterfactual item: how did they think writing with such a partner might be different?

More collaborative/communicative Most commonly, participants expected greater communication and a more collaborative interaction from a human co-writer. P7 noted this:

> " I think I would want another human to be more coherent. I would expect them to make more meaningful contributions to the work and I think that would feel more like a collaboration. This felt like I was immersing myself in a writing environment where the things were tangentially related to what I was writing, but not exactly relevant. "

More understanding/experience Other participants alluded to experience of the world or understanding about more general aspects of narrative development. P6 wrote that "I could explain to the person the story I am thinking about. I could convey the tone and the feelings I am trying to infuse in my text"

Speed (slower/faster) or effort Participants either thought writing with a human would be slower or faster, or require less or more effort in comparison. More commonly, participants thought it would be slower and require more effort. P22 reasoned: "I think, it will require more efforts because mutual agreements between the other humans are needed and the outcome really depends on how you and that person's relationship, their background, mutual understanding, etc."

	N	Example	Yes	No	Unsure
Similar	6	"Yes, it did. It is similar with the introduction video and I can use it easily by watching that."	0	6	0
Less relevant	4	"Sometimes I had no idea where the pictures in the grid came from, because sometimes they seemed relevant to what I had written and then sometimes it seemed like they were completely random. I expected the plot suggestions to be a little less repetitive."	4	0	0
Less out-there	3	"I expected the sounds to be more ambient sounds that contributed to a certain vibe of the story, but they seemed much more random and chaotic, which maybe had to do with the content of my story. I didn't expect that the editor would be able to draw upon elements of the story I had already written. I expected it to imagine new storylines that may have taken me in a new direction."	2	1	0
Other	2	"It was a novel experience. I had no expectations because I wasn't sure what to expect."	1	0	1
More creative/intelligent	2	"It was surprising to see the intelligence of the AI and the creativeness of the suggestions, for example "cryogenic sleep" was a very novel idea suggested to me."	1	1	0
Less subtlety/control	2	"The plot suggestions were not as subtle as I expected, like I thought it could help lead into plot points but instead the suggestions were often far off things I'd have to work toward and might take a bit of time to write to that point to incorporate the plot suggestions and make it make sense."	2	0	0
Slower	2	"see answer above for details. its interface acted as expected but I expected it to give suggestions faster and those be more relevant"	1	1	0
Not as directly usable/helpful	2	"The sounds were not what I was expecting. They sounded quite eery and scary, and I was trying to write something more lighthearted. Some of the suggestions were surprisingly good (like including characters that I'd mentioned and dialogue). Some of the suggestions also looped around though, and didn't really make sense (& I maybe expected them all to be kind of ok). I liked that the suggestions were slightly longer than I expected though."	2	0	0

Table 8.8: **Codes from how using Editor-Red compared with each participant's expectations.** N is number of responses, Example contains a quote associated with the label, and Yes/No/Unsure indicate whether the experience was overall different than expected.

P11 similarly expected writing with a human to be slower and/or require more effort, especially one who isn't a professional author. They attributed this to time needed for information processing, searching with other tools, etc.:

> " I think it would take more time if I write with another human since he/she would have to think for ideas and suggestions as well, or if not, he/she might also use another artificial intelligence like Google to find more ideas. Unless, the human is a professional author. If not, I think it will take more time to write as I would have to discuss as well. In addition, another human would not be able to create the mood/ ambience that I would like to have while writing. "

More questions P3, among others, expected more questions along the writing process:

> " I think we probably [would've] asked each other questions back and forth like "why is the character doing this? What does it sound like?" etc. Writing with humans tends to involve a more "question-based" approach. "

Less self-driven P14 and P15 expected the process to be less self-driven or self-owned, P14 wrote: "It might also feel less introspective (I enjoy the space from being alone)," while P15 alluded to ownership (see Table 8.9.

Additional thoughts P23 expected that another human might face similar challenges to the writer (see Table 8.9 for quote) and P16 simply indicated a general difference, without specifying their reasoning about why.

8.5.4 Relating participant expectations, processes, and outcomes

Our three-fold study data collection generated a great deal of information from relatively few participants, describing each one's interaction with our system in substantive detail. Although our study design's primary goal is for these three types of data to collectively provide insight, here we review a few examples of instances where explicitly combining these sources of data at the level of the participant or sample provides additional information.

	N	Example
More collaborative/communicative	10	"I think that another human being would have asked questions along with suggestions in order to better tailor their suggestions."
More understanding/experience	8	"Yes. I think the human would be more helpful because they could suggest if more description should be added to the setting or if the character needs to be developed more or suggest a direction for the plot, none of which I felt I got from the editor-red suggestions."
Speed (Slower/Faster) or Effort	4	"Slower going. Plot suggestions would have made more sense, but I don't think description ones would have been much difference in help. A human might have been more helpful with naming things in the story."
More questions	3	"I think we probably wouldn't asked each other questions back and forth like "why is the character doing this? What does it sound like?" etc. Writing with humans tends to involve a more "question-based" approach."
Other	2	"Another human might have challenges coming up with ideas just like the writer. The suggestions might have been different which would have geared my story in a different direction altogether."
Less self-driven	2	"Writing with another human would have definitely changed the course of the story. It also would have felt less like my writing because of the other person's ideas. "

Table 8.9: **Codes re: expected differences from writing with a human co-writer, *after* writing.** N indicates number of participants, Example shows a corresponding quote. N.B. some responses are labeled with more than one code.

Anchoring to prior expectations

Regarding the possibility of AI creativity, P6 noted that they thought it "depends on the amount and type of data that will be available to the AI to create something new," emphasizing that the "broader and [more] various the set of data the more creative the AI could be." During the interaction, we observed P6 not engaging with the suggestions to advance their story, but rather, as discussed earlier, attempting to improve the system suggestions with their writing instead. In this case, an inaccurate expectation resulted in the system being unhelpful to them, due to their behavior anchoring to this expectation rather than adjusting to the system's behavior during the process.

By contrast, some participants who were optimistic about the ability for AI to be creative managed to find utility in suggestions that may even have reflected poor or less coherent language-modeling behavior. For example, P15 wrote that "...information/ideas provided by AI can be completely illogical which is sometimes the best creativity" and, after writing, indicated their willingness to make sense of and incorporate possibly irrelevant suggestions: "with some modifications, the ideas presented were definitely helpful... images are a hit or miss because they don't tell the same story as the text I am writing, but if I thought about the images a bit more creatively rather than literally, they were a bit more helpful."

Adjustments to prior expectations

Some participants appeared to adjust their prior expectations after interacting with the system. A particularly clear case is P1, who initially expressed a belief that "AI can not be creative," but could be "accurate." During the interaction with the system, having received a suggestion, P1 was impressed with it and was contemplating whether to characterize it as "accurate" or "creative," finally coming to the conclusion that "this is a really good plot and it's creative enough." After the interaction, they noted that "It was surprising to see the intelligence of the AI and the creativeness of the suggestions." This participant initially identified that *Value (useful to people)* was the most important aspect of human creativity to them, and described the suggestions afterwards as "helpful. Even if I didn't use every idea that was suggested to me, I was inspired by their mood." The suggestions were useful to them in their process of writing, perhaps demonstrating the characteristics of *Value*.

At a sample level, a majority of participants ($N = 14$) initially responded that they would consider the final text to be "co-written by myself and AI," however afterwards, almost all participants ($N = 22$) indicated that they would call the written text their own (with one participant unsure). In addition, all those who responses Unsure or No to differences between human and AI text production initially ($N = 7$; P23, P17, P10, P6, P22, P9, P3) were able to communicate clear expected differences after the interaction. For example, P6 initially indicated that they were unsure, suggesting that "...it depends on the level of development of the AI", but finally wrote that with a human they "could explain to the person the story I am thinking about... convey the tone and the feelings I am trying to infuse," which is about explicit communication and intuitive influence rather than modeling performance. Participants had viewed a video of our system before the initial response, indicating that their perception was informed by actually interacting with the system rather than its overall design and method of suggesting.

Do more accurate mental models of AI improve the experience or outcome?

To examine this question, we consider the two opposed categories of mental models of AI: Sparse-Abstract and Sophisticated-Operational. These groups respectively had the least and most detailed and accurate expectations of AI. We can examine how their outcome evaluations varied across the dimensions of overall experience, suggestion helpfulness, and differences from expectations.

Sparse-Abstract. 8/14 participants reported overall positive experiences, with 2 neutral and 4 negative. 10/14 total reported suggestions being at least sometimes helpful (1 rarely helpful, 3 not helpful). 9/14 indicated that the experience was different from their expectations, with 1 unsure and 4 reporting no or not much difference.

Sophisticated-Operational. 3/5 participants reported overall positive experiences, with 1 neutral and 1 negative. All 5 indicated that suggestions were at least sometimes helpful. 3/5 indicated that the experience was different from their expectations, with 2 reporting no or not much difference.

The data in this case indicate a complex relationship between the depth and accuracy in explanatory models of AI and experiences with our interface. A simple assumption might be that having more well-calibrated expectations of AI systems might arise from technically deeper and more accurate reasoning about how it generally works, which

may appear to be supported by how helpful participants thought suggestions were. However, our observations and participants' comments about suggestion helpfulness suggest that the differences have more to do with styles of writing and openness to narrative change than accurate expectations of the system's behavior. This is reinforced by the lack of clear difference in whether the system behaved as expected or not between these two groups.

Even so, we can explore this further by considering whether accurate expectations themselves were clearly associated with positive impressions or indicated suggestion helpfulness. 9/13 of those who reported a difference from expectations indicated an overall positive experience, compared with 4/9 who didn't. For helpfulness of suggestions, 9/13 who reported a difference from expectations found suggestions at least somewhat helpful, as compared with 8/9 who didn't. Again, there is a divergence between the two outcome variables, suggesting a complex relationship between expectations and outcomes.

8.5.5 Usability and overall experience

Usage data

In all, participants requested 165 suggestions and received 162 (3/165 requests were not resolved, for example due to requesting a subsequent suggestion while one was already processing). The median number of suggestions requested and received per participant over the 20-minute session was 7 (min = 2, max = 15). Participants wrote 229 words on average in **Editor-Red** compared with 296 in **Editor-Green**. With regard to suggestion modalities, on average participants had images toggled "on" for 99.8% (min 97.5%, max 100%) of the duration of their active engagement with the system, and sounds for 77.1% (min 25.2%, max 100%), with this duration measured from either the first suggestion request or alterations to these controls, to the last. During this period, participants toggled images on and off between 0 and 9 times (median 1) and sounds between 0 and 11 times (median 1.5).

System usability

As part of the post-task survey, we presented the participants with a battery of questions tailored to understand their experience using our interface, as shown in Fig. 8-7. The post-task survey contained 42 questions and produced more data than

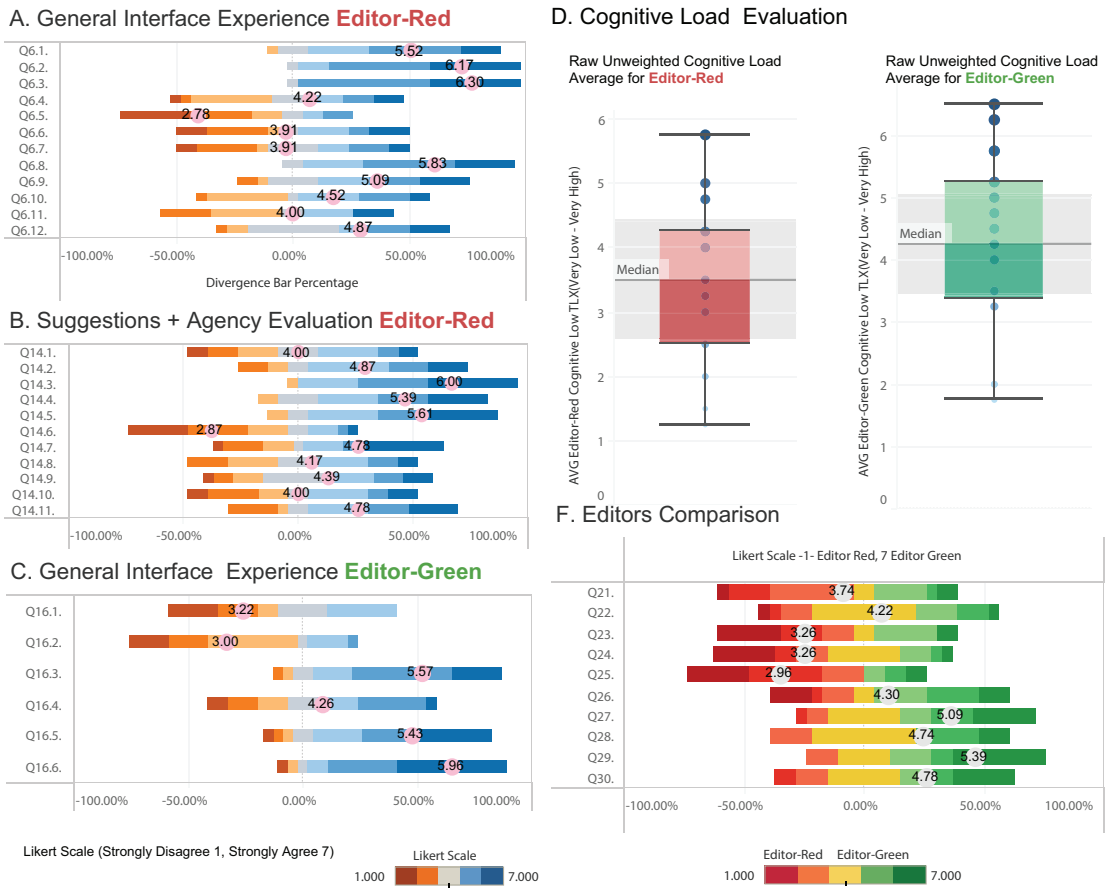

Figure 8-7: **Post-task questionnaire focusing on user experience and editors comparison**. Full list of questions in Appendix A.5.1.

could be analyzed within the scope of this chapter. We report those questions that examined the issues that are the focus of this chapter.

Participants responded to **Q6.4** where they were asked to score using the likert scale Strongly Disagree (**DDD**)=1, Disagree (**DD**)=2, Somewhat Disagree (**D**)=3, Neutral(**N**)=4, Somewhat Agree (**A**)=5, Agree (**AA**)=6, Strongly Agree (**AAA**)=7 to the statement "The pictures used in **Editor-Red** were helpful." The responses were almost evenly distributed with one more participant disagreeing than agreeing (*AAA*=3, *AA*=3, *A*=3, *N*=4, *D*=8, *DD*=1, *DDD*=1). However, when asked to rate the statement **Q6.5** "The pictures used in **Editor-Red** distracted me from my task," the majority of participants disagreed (*AA*=3, *A*=2, *N*=2, *D*=3, *DD*=6, *DDD*=7).

We also asked participants to rate "The sounds used in **Editor-Red** were helpful." **Q6.6** (*AAA*=4, *AA*=2, *A*=5, *N*=1, *D*=2, *DD*=6, *DDD*=3), as well as "The sounds used in **Editor-Red** distracted me from my task," **Q6.7** (*AAA*=2, *AA*=4, *A*=3, *N*=5, *D*=1, *DD*=6, *DDD*=1) on both of which participants were somewhat evenly distributed between overall agreement and disagreement, with 5 participants vs. 1 participant neutral respectively. The distribution to these questions can be found in Fig.8-7A.

Participants responded to **Q6.11** "Using Editor-Red was intuitive" and **Q6.11** "Using Editor-Red was easy" we found that **Editor-Red** was considered intuitive to use (*AAA*=4, *AA*=9, *A*=6, *N*=3,*DD*=1) and easy (*AAA*=9, *AA*=10, *A*=3, *N*=1).

Finally, We were also curious to know how participants felt towards the different modalities available to them in the interface shown in Fig. 8-7B. When asked to rate **Q14.5** "I mostly used the textual suggestions and not pictures or sounds" participants agreed (*AAA*=8, *AA*=2, *A*=4, N=1, *D*=3, *DD*=4, *DDD*=1).

Effort and cognitive load

To understand the cognitive load imposed by writing with our system, we sub-sampled a group of 4 items from the NASA TLX (removing physical effort and performance, which are less relevant in our case) and then took the raw TLX score, which is simply the mean rating across all items per participant. We found that 17/23 participants rated the cognitive load of **Editor-Red** ($\mu = 3.41,\ \sigma = 1.15$, *Mdn.* = 3.50, *Min* = 1.25, *Max* = 5.75) lower than **Editor-Green** ($\mu = 4.13,\ \sigma = 1.45$, *Mdn.* = 4.25, *Min* = 1.75, *Max* = 6.50) Fig.8-7D.

Evaluation of outcome creativity

We captured the participants' perception of the creative output and of **Editor-Red**'s creative support by asking them to rate the following statement: **Q14.1** "I did most of the creative writing, using **Editor-Red** just for suggestions." Almost all the participants (22/23) agreed with the statement with the exception of one participant (*AAA*=9, *AA*=7, *A*=6, *DD*=1) . When asked to rate **Q14.8** "The suggestions made by **Editor-Red** were creative" we found that 16/23 participants agreed the suggestions were creative, 2/23 were neutral and 5/23 disagreed (*AAA*=4, *AA*=4, *A*=8, *N*=2, *D*=2, *DD*=3) Fig.8-7C.

We further asked for them to evaluate whether or not the final text created during the

task was creative, using **Editor-Red**: **Q17** "Do you consider the text that you wrote in **Editor-Red** creative?" Most participants responded "Yes" (N=20), with 3 participants answering "No." Similarly, we asked them to rate the text they wrote with **Editor-Green**: **Q19** "Do you consider the text that you wrote in **Editor-Green** creative?" 14/23 participants responded "Yes" and 9/23 responded "No." We found that 13/23 participants preferred **Editor-Red** as the text generated during the task was perceived as more creative, and when asked: **Q21** "In which editor was the text that you wrote more creative?" (from **Editor-Red**=1 to **Editor-Green**=7), 2/23 indicated Neither or Both, with 8/23 participants towards **Editor-Green** (1: (N=1), 2: (N=4), 3: (N=8), 4: (N=2), 5: (N=5), 6: (N=1), 7: (N=2)). Participants show a preference towards **Editor-Red** for using that type of editor for writing creative text. **Q25** Which editor did you prefer for writing a creative text? **Editor-Red** (17/23) **Editor-Green** (6/23) (1: (N=6), 2: (N=7), 3: (N=4), 4: (N=0), 5: (N=2), 6: (N=2), 7: (N=2)) Fig. 8-7C.

8.5.6 Agency and ownership

We found that participants generally enjoyed writing with the help of suggestions from **Editor-Red** and were enthusiastic about the concept of writing with a "collaborator," especially once natural language generation capabilities improve and the suggestions are closer to their own writing style. From observing the writing session and post-survey conversation, it was unclear that the issue of ownership and agency in co-writing with AI was something that participants were at all concerned with or gave any thought to.

When prompted to rate the statement **Q14.2** "I enjoyed co-writing with **Editor-Red** " in our post-task survey, participants responded with 17 participants agreeing, 4 participants neutral, and 2 participants disagree (*AAA*=6, *AA*=5, *A*=6, *N*=4, *DD*=2). A similar response when asked **Q14.3** "I enjoyed collaborating with **Editor-Red**" was shown with 19/23 participants agree, 2/23 neutral, and 2/23 disagreeing (*AAA*=8, *AA*=4, *A*=7, *N*=2, *DD*=2) and when asked to rate the following statement **Q14.6** "The final product of writing is a result of joint efforts of **Editor-Red** and myself" participants responded with 10/23 agreeing, 4/23 neutral, and 9/23 disagreeing (*AAA*=2, *AA*=2, *A*=6, *N*=4, *D*=4, *DD*=3, *DDD*=2). Almost all the participants (22/23) responded that they only relied on **Editor-Red** for suggestions and did most of the creative writing **Q14.1**. (*AAA*=9, *AA*=7, *A*=6, *DD*=1) Fig. 8-7C.

Two participants admitted they were primed by questions in the pre-session survey to

think about agency and ownership in writing using the suggestions of the system. P3 explained they had thought that “using AI interface would make me feel that I wasn’t even doing my own writing," but ultimately P3 felt that the system “helped along” but “didn’t tell me what to write.” P22 (one of the four participants who didn’t visibly use **Editor-Red**’s suggestions) explained that even if they had used the suggestions, they would still call it “using my own creativity” as they believed that “even deciding to use it or not, is actually really a choice for me” and is seen as a part of “creativeness.” The participants seem to think that this type of system can improve creative writing by being supported by the system and less cognitively demanding than the simple text editor.

8.6 Discussion

8.6.1 Suggestion quality: relevance, coherence, and variety

Several participants rejected suggestions for a perceived lack of coherence or relevance to their developing texts, which comports with prior work on language model assisted writing [72, 101]. Building on this, we have also shown that several others in our study did not see this as an obstacle to working with the system and in some cases appreciated less immediately semantically relevant suggestions and were able to incorporate ideas from less linguistically coherent suggested sentences. While we attempted to trace this difference to expectations, model behaviors, and other potential predictors, our expectation from observing participants is that this has primarily to do with a difference in participants’ approach to creative writing. As such, the relevance of suggestions may not be a simple variable to always aim toward maximizing; rather, the optimal level of relevance might vary by writer. Sometimes, it might also vary depending on other circumstances; for example, some participants noted that less relevant suggestions likely required more time to integrate, and that they might do so given additional time to write. This may also be reflected in the fact that on average, participants wrote less text in **Editor-Red** than in **Editor-Green**, though we note this is also related to other aspects of the interaction in our study (the novelty of the interface, participants talking more while using **Editor-Red**, etc.).

The ambiguity in assessing relevance extended to the multimodal concept representations; even when they were not used directly, their contribution to the environment might vary with their relevance. For example P8, who didn’t visibly incorporate any

suggestions, noted that they were "impressed that the sound suggestions seemed to pick up on the creepy, suspenseful tone of the story right away, and it could be helpful if the image suggestions followed the tone more closely" as compared with P5 who wrote that the "sound wasn't directly influencing my ideas but having background noise was relaxing."

Balancing relevance with variety is likely to be important in making suggestions useful to participants, in our assessment. Participants especially noted the homogeneity of images: "I mentioned a phone and the grid overlay just shoved several iterations of smartphones, it would be nice if it could show different types of telephones" (P20). This homogeneity also extended demographic factors: "there's just a bunch of white guys staring at me and I don't know why" (P2) and "they are all images of straight blonde Caucasian women" (P5). We noted that these instances were not directly related to query material, indicating that these might reflect broad biases in available images.

Technical approaches to generative modeling and information retrieval to support creative processes should, in our view, be intentional in handling these parameters (relevance, variety) and consider individual and situational variation in their optimality criteria. Modeling this is likely non-trivial and raises questions such as: what is relevant when and to whom? When are precise, logical suggestions needed, and when are surprising, unusual suggestions needed? The integrative leaps we have reported on suggest the practical challenges in automatically inferring this trade-off, or even reducing it to a simple, one-dimensional control. A helpful source of information in our case is the writers; they often have strong intuitions about both. Finding channels for writers to communicate their personal stylistic and contextual narrative needs to both interfaces and the underlying models, for example in natural language or by providing examples, may help these systems robustly support creative expression by both being flexible and allowing users to clearly and naturally communicate their needs and intentions.

8.6.2 Editor-Red beyond writing suggestions

A supportive writing environment

Though what **Editor-Red** seems to be doing on the surface is providing words, lines, and ideas to borrow and rely on, we observed much more than just that in the

interaction settings we studied. A wide range of participants' comments highlight that the system acted as a support tool in diverse ways. Those participants who actively integrated the system's suggestions admitted that **Editor-Red** was structuring their process of writing. For instance, P1 admitted that they found themselves at a certain point "writing *for* the suggestions," seeing **Editor-Red** as "a form of motivation to continue writing" in order to get better suggestions. P3 commented that **Editor-Red** helped them "keep going" and "continue along" with their writing when they otherwise would have stopped.

Editor-Red redirected attention from being stuck (P12) and helped feel "less stuck" even when the participant was not taking the system's suggestions (P16). Writing as a process is fraught with self-doubt, anxiety, and feeling overwhelmed, systems like **Editor-Red** can mitigate stress by being a comforting distraction like "petting a cat" (P6).

Some participants used suggestions as just a starting point for the participants' own individual creative journey. For example, P23 explained that often **Editor-Red** suggestions gave them a different idea rather than taking the suggestion right directly. As P23 further explained, getting inspiration from something can be unrelated to what that inspiration was. The integrative leaps (see §8.5.2) that participants made when engaging with **Editor-Red** illustrates a wide range of examples of what users are capable of and willing to do when integrating with a writing system.

Personal and cultural references versus AI-generated references

In the "blank page" writing with **Editor-Green**, 10 participants out of 23 visibly relied on cultural (books, TV shows, music videos) and personal references (memories, personal experiences, and immediate surroundings, e.g. describing what one can see from the window). For example, P8 writing in **Editor-Green** with the prompt "A train arrives at the station," explains that they are thinking about "the train station and Anna Karenina, kind of thing." P9 writing in **Editor-Green** with the prompt "The phone began to ring" explains that "the phone" made them think about a landline, a landline made them think about a hotel, and that, in turn, made them think about the last trip they had when they were staying in a hotel, which prompted a subsequent description they made in **Editor-Green** writing (after completing **Editor-Red** writing).

When writing in **Editor-Red**, 5 participants out of these 10 did not visibly use any cul-

tural or personal references in their writing, but instead relied on the **Editor-Red** suggestions. Their story development was structured and oriented by the suggestions they incorporated. This is not that surprising on its own since participants were asked to use the features in **Editor-Red**. However, it is possible to imagine that writing with systems like **Editor-Red** allows a user to rely less on one's own cultural and personal references and one's "self" (an individual's interiority, a means and ends of one's own actions [261, 262]). Rather, it provides an opportunity for a user to interact with the "self" of a system, deriving references from its suggestions.

We hypothesize that systems like **Editor-Red** can be used also when users for situational or psychological reasons do not want to engage with their own experience and inner thoughts and can be used for building systems that can provide therapeutic support for a user. This type of psychological support function has been recently identified in other human-AI creative interaction domains [491].

8.6.3 Dynamics of suggestion integration

Ideas for writing often came not through directly applying **Editor-Red**'s suggestions, but as a result of active engagement with the system from the participant's side and their readiness to do cognitive work in extending, adjusting, and altering suggestions and/or prior text to better suit the combination of text they had written and either any thoughts in their mind about how to proceed (confirmatory) or ideas about altering the narrative to lead in a new direction (exploratory).

The comments that participants made explaining these integration examples provide an insight into the multiplicity and multidimensionality of practices involved in human interaction with generative language systems, and especially how users create new meaning through this interaction. We specifically did not aim to do a linguistic or semiotic analysis of the integrative leaps that we documented, which we argue would require a great deal more data in order to yield generalizable insights. Instead, we aimed to document some orientation points that reflect structural differences in participant behaviors during our study.

Creativity, inefficiency, and synthesis

When interacting with **Editor-Red** , participants' concepts of creativity in suggestions often seem to be constrained by the possibility of an easy transition. The possibility of an easy transition, in turn, is individually and contextually varying. Those participants

whom we identified as willing to cooperate with **Editor-Red** and incorporate its suggestions, did not seem to mind suggestions being "absurd," "crazy," and "out there" (we will refer to these as "out of sync"). These suggestions sometimes led to considerable changes to the subsequent and prior narratives; participants made decisive creative moves when they were willing to engage in this way.

Monster hunting, cryogenic sleep, a detective in 1890 Austria, and the first human to die on Mars were just some of the ideas that participants received as suggestions. Incorporating these suggestions depended on whether participants were mentally and emotionally ready to make the necessary efforts to synthesize an easy transition from the prior text and the given suggestion. It was "a much longer route" for monster hunting (P5), and didn't come "at a good time for the story" for cryogenic sleep (P1), and so these suggestions were not integrated. However, "you are a detective in 1890 Austria" made the participant think about the concept of time in their story (P1). P2, having completed almost the whole story that took place on a London farm, received a suggestion saying "you are the first human to die on Mars." P2 did comment that the description "is not very accurate" to their situation but then changed their mind: "You know, let's make it about Mars, why not?" and rewrote the story in four places to fit the premise of taking place on Mars.

We conclude from this assortment of complementary and contradictory behaviors that the incorporation of a suggestion that is "too creative" does not depend just on the content of the suggestion, but rather on the possibility of transition which is influenced by individual and situational factors. The transition towards a suggestion that is unexpected and unrelated to the input text is dependent on the readiness and motivation of a user to the requisite cognitive and/or emotional work toward a meaningful synthesis of elements. These observations align with Freiman's characterization of the writer's drafting process, involving a "state of unknowing", a "kind of faith" that something will emerge from the drafting, and ultimately how "something that perhaps lacked cohesion or structure now becomes more concrete or coherent in the making of the text" [161]. Freiman suggests this happens by the writer making cognitive, affective, linguistic, and other creative decisions through a series of drafts and changes. We also expect that cognitive work done on drafting and revising to achieve such a synthesis may also become a path to support ownership of the text and creative endeavor, in our context of AI-generated suggestions.

Cognitive reorganization and expectations of non-human writing systems

What are the underlying cognitive mechanisms by which distant suggestions are able to be meaningfully integrated into users' existing narratives? Participants of our study were actively aware of the task environment [155]: writing a story using **Editor-Red** (non-human, AI system), a rhetorical problem (write a story given a prompt), integrating **Editor-Red** suggestions (which they had preconceptions of being based on human language, rule-based, possibly random, illogical, and creative), and the text itself that is evolving and changing. Since it was possible to get suggestions multiple times and the suggestions were different every time (both in content but also in terms of the level of relevance or detail), every new suggestion created a micro-moment of interaction and adjustment.

Attending to **Editor-Red** suggestions, building up all the missing cognitive links or not immediately visible links so as to update the story sometimes involves a considerable amount of cognitive reorganization of narrative information, in the sense of reorganizing what one already knows (e.g. Piaget's equilibration [388]) or, in this case, has already written. One possible mechanism for this is self-explanation, which is an attempt to make sense of new information by explaining it to oneself [92]. Unlike self-explanation in learning, wherein the central inferential process needs to construct new knowledge at the level of "the world", here self-explanation may provide an inferential process to reorganize the narrative by finding possible connections and associations, similarity, extracting abstract properties, or making referential links (for example, as we described earlier with P4 having the precondition of a crime, seeing an elephant that seems irrelevant, and explaining the presence of the elephant by making it the object of the crime involved). Other possible mechanisms for combining distant concepts have also been described in prior literature, such as causal reasoning [274], comparison and construction [548], conceptual integration or "blending" [112, 519], and satisfying constraints like diagnosticity, plausibility, and informativeness [108].

In the case of writing with our system, the willingness of participants to engage in this process may come from user expectations, due to the non-human source of the suggestions. For example, we noted earlier that participants may expect suggestions to be random, illogical, having connection to the real world yet often being situationally "out of sync." In that way, any faults of the suggestions and difficulties that arise from those become not only something that has to be looked

past but actually act as inherent features of the system and accepted as part and parcel of this interaction. These suggestions can be seen as not a bug but an inherent feature and a necessary condition of this interaction, as humans perform integrative leaps, engaging in cognitive reorganization of narrative because they accept the premise of interacting and co-writing with a non-human system, and the implications that come with it.

How can "out of sync" suggestions be helpful for writing?

Earlier work has illustrated how completely unrelated ideas and unusual word combinations can be evocative and productive for creative writing [77, 128, 542]. In the case of causal reasoning, the surprisingness of combinations may provoke additional and exploratory processes and thereby the production of creative ideas [274]. We hypothesize that another mechanism by which semantically distant suggestions might be useful is by explicitly prompting more critical evaluations of written content, i.e. what Flower and Hayes call "evaluating" and "revising" [155]. By contrast, we might consider highly probable and user-adapted word predictions, which can be absorbed into a writing task with minimal effort (e.g. a click) to accomplish well-defined goals more efficiently (e.g. respond to a work email). We can model distant suggestions with such semantic difficulties as we observe as being useful inefficiencies which prompt critical evaluations of drafts and suggestions, metacognitive reflection about narrative development, and ultimately axes for more substantial narrative reorientation, where otherwise there would be no prompt or incentive to re-engage with and reconsider prior thoughts and writing. More work is needed to examine this possibility in detail.

8.6.4 Design recommendations

We observed that participants are capable of making leaps to integrate suggestions into their writing when presented even when the suggestions were unrelated to their current writing. However, there seems to be a general need to have these suggestions build on, refer to, or otherwise be relatable to aspects of their writing/story for many writers to have a more helpful experience. There is a need for details and descriptions of objects and important places when developing the story, and for systems to attend to the right parts of stories, which vary, when making suggestions. In our study, we observed most users rely on the system to enhance their writing when adding supporting material. When the system was not helpful in either introducing

supporting material or helping them think of new directions, frustration and lack of trust in the tool often began to arise. However, as we have explored, suggestion quality is a multidimensional property which varies individually and contextually. To make suggestions useful to participants does not always mean maximizing their immediate relevance, but rather requires supporting the *process* of suggestion integration. Here we consider what that may mean for different technical and design considerations for creative writing support tools at every level of the process.

Datasets for creative writing

Participants in our study had different experiences with the two suggestion channels (*Plot* and *Description*), despite the commonality in the modeling method. Mirroring calls from other domains for data-oriented rather than model-oriented progress in AI [119, 443], we argue that well-curated datasets oriented towards domain constructs can support diversity and relevance, two factors we identified earlier as especially salient in machine contributions to creative writing work. Larger and more diverse pretraining sets can also result in greater coherence, if matched with an appropriately parameterized model, which we find would be helpful to several users in a variety of contexts.

Language modeling

What can better models help with? As noted, more modeling power can result in increased coherence and relevance, especially as processed sequences get longer, if pretrained on appropriately large and diverse datasets, as well as fine-tuned on downstream datasets that provide creative value. These properties are desirable in many cases, as pointed out by our study participants. In parallel, models with implicitly richer knowledge bases [387] may also extend more diverse suggestions to users, finding interesting relations with aspects of their writing, and assisting them in performing contextually appropriate and creatively fulfilling integrations.

What can't better models help with? Larger models are typically slower, more difficult to fine-tune and host, and increasingly closed-source, expensive to obtain access to, and private. Additionally, we noted many instances in which the cognitive work done by participants was the operative force in making suggestions helpful and ultimately able to contribute to their writing. For these participants, writing styles,

and situations, larger language models may not necessarily help much, but would incur costs in interactivity, which were already pointed out by some participants in our current prototype. In our case, suggestions typically took 3-5 seconds after requests (given that we were running two separate fine-tuned models, extracting keywords, etc.), depending on the length of the input text; larger models may take significantly longer (one writer estimates GPT-3's Davinci model's typical speed at 147 words per minute [60]) and are very challenging to host and serve interactive requests with due to the resources needed.

Even the best possible language models have an extremely limited capacity to understand our intentions. They cannot reason about human internal cognitive processes, implicit judgments, and novel forms of creative exploration and expression that intentionally disregard convention. Better language models, better for different purposes, can support the process, but a great deal of what makes human creativity successful is outside of their purview.

Semantic influence Some participants indicated a desire to influence or control this facet of suggestions with prior information, e.g. high-level story goals, moods, feelings, and ideas. While relevance can already be expressed to language models at *sampling* time to some extent, through stochastic decoding methods and controls like *temperature*, the ability to semantically "steer" relevance towards more fruitful integrations, rather than expressing it as a numerical value, might also better support diverse writers' diverse needs. Such steering can be explicitly enabled [253, 268, 298], for example, by conditional modeling, or, in the absence of specialized approaches, even discovered by so-called "prompt engineering" which has been successfully used[7] by many for language-controlled visual art generation [383] with general-purpose vision+language models [402].

Interface design

Overall goals of interfaces Based on the behaviors observed in our study, we recommend that creative writing suggestions be designed to prompt and support cognitive processes that lead to suggestion integration and narrative engagement, rather than auto-complete style continuation. This seems to additionally support participant ownership over the outcome, as we observed in our study. There is a great

[7] https://ml.berkeley.edu/blog/posts/clip-art/

deal of cognitive effort involved in writing with external stimuli, in order to make sense of them, recognize the possibilities for their contributions to the work, and perform effective integrations.

We argue that the focus of designing new creative writing support tools with intelligent augmentation should be on supporting this cognitive effort while preserving writer autonomy, authorial discretion, and creative flow. In our interface, we do this by implicitly discouraging directly absorptive behaviors; suggestions are presented in a different graphical environment rather than overlaid on the text, and the familiar tab key invokes new suggestions rather than directly integrating them into the writing. The corresponding reduction in cognitive load for most participants (17/23) by a small amount overall may reflect both the helpfulness of external suggestions in easing the cognitive burden of blank-page style writing, as well as the additional load introduced by the additional stimuli in context.

Multimodal support for a unimodal task Additionally, visual and auditory suggestions cannot be simply inserted into a textual story, and we expect that the process of resolving these morphological differences to create meaningful semantic connections may also contribute to making creative leaps in writing stories. Our results suggest these features be made easy to turn off: this was a feature our participants used extensively to account for both individual and situational variation. Future work might examine the methods for communicating these parallel channels of information.

Evaluation criteria and methodologies

Our Expectation-Process-Outcome model, which guided our study design that combined surveys, behavioral observation, and semi-structured interviews, allowed us to capture several things: a rich representation of conceptually relevant background which participants brought into their interaction with a novel system, their interpretive reasoning through the course of the interaction, and their evaluative judgements and impressions afterwards. Additionally, through capturing prior assumptions and explanatory models, we were able to begin to obtain a fuller picture of how the interaction is framed by and adjusts expectations, as well as some effects this may have on the experience and outcomes.

We recommend that designers of complex, novel tools to support open-ended creative tasks similarly consider the conceptual priors of their users in conjunction with

evidence from their experiences, behaviors, and *a posteriori* thoughts. Through this, we might begin to better characterize the significant level of individual and situational variation, and design tools that not only practically accommodate this but actively benefit from it.

8.7 Conclusion

This research presents an extensive study of machine-in-the-loop creative writing, centered around a new interface that makes writing suggestions through sight, sound, and language. Through collecting data on participant expectations, processes, and outcomes of interacting with this system, we discussed how individual writing approaches and narrative circumstances influence the interaction. By eliciting user explanatory models of AI, human and AI creativity, and creative writing, we explored how expectations might influence and be influenced by the interaction. We additionally reported on users' responses to suggestions through the lens of *integrative leaps*, by which participants incorporate suggested ideas into their writing process by performing cognitive work to make transitions possible.

As AI-based systems increasingly engage in traditionally human creative capacities, building stronger and more adaptive human-centered foundations for human-AI creative interaction will be increasingly important. Modeling advances in the systems periphery of everyday life have made it increasingly plausible that AI can be creative, but the more challenging work is to make it plausible that it might broadly extend our creative faculties by understanding our needs differently than other human creative partners. We believe that deep and wide-ranging investigations such as those we described in this work can inform design methodologies and yield powerful and useful tools that extend our abilities.

9

FigurA11y: AI Assistance for Writing Scientific Alt Text

When we study how humans integrate AI suggestions into their writing processes, we often focus on open-ended generative tasks like storytelling (as we did in Chapter 8). We looked at how writers perform cognitive work—"integrative leaps"—to transform suggestions of varying relevance into fruitful content. What happens when we constrain this integration process with the need for accuracy and domain expertise? In many real-world writing tasks, this is precisely the setting where AI assistants must operate.

In this chapter, we examine the case of scientific alt text authoring. Given the low prevalence of alt text in the field, AI assistance could help bridge this gap by enabling authors to efficiently and effectively make their papers more accessible. Like creative writing, it requires translating ideas into clear and effective prose. Unlike fiction writing, however, these descriptions must accurately convey specific visual information to readers who rely on them for access. The "integration" work here is an important bridge between complex technical content and accessibility for real-world readers. As such, an AI suggestion must be evaluated against the author's deep knowledge of their research and context (e.g. the broader field). Scientific figures are frequently complex images with multiple parts, requiring relatively precise inferences for accurate communication. So far, such images have eluded even the best vision-language models. Full automation remains challenging since authors uniquely understand the context and interpretation surrounding their figures. As such, in this work we propose a system that supports authors in drafting and revising comprehensive alt text descriptions, and study how authors make use of this system with their own figures.

Abstract

High-quality alt text is crucial for making scientific figures accessible to blind and low-vision readers. Crafting complete, accurate alt text is challenging even for domain experts, as published figures often depict complex visual information and readers have varied informational needs. These challenges, along with high diversity in figure types and domain-specific details, also limit the usefulness of fully automated approaches. Consequently, the prevalence of high-quality alt text is very low in scientific papers today. We investigate whether and how human-AI collaborative editing systems can help address the difficulty of writing high-quality alt text for complex scientific figures. We present FigurA11y, an interactive system that generates draft alt text and provides suggestions for author revisions using a pipeline driven by extracted figure and paper metadata. We test two versions, motivated by prior work on visual accessibility and writing support. The base **Draft+Revise** version provides authors with an automatically generated draft description to revise, along with extracted figure metadata and figure-specific alt text guidelines to support the revision process. The full **Interactive Assistance** version further adds contextualized suggestions: text snippets to iteratively produce descriptions, and hypothetical user questions with possible answers to reveal potential ambiguities and resolutions. In a study of authors (N=14), we found the system assisted them in efficiently producing descriptive alt text. Generated drafts and interface elements enabled authors to quickly initiate and edit detailed descriptions. Additionally, interactive suggestions from the full system prompted more iteration and highlighted aspects for authors to consider, resulting in greater deviation from the drafts without increased average cognitive load or manual effort.

9.1 Introduction

Digital dissemination has allowed scientific authors to reach broad audiences with their work. However, one audience that continues to face barriers is blind and low-vision (BLV) readers. BLV individuals typically rely on alternative text (alt text) descriptions to access key data and concepts communicated visually in figures. Distinct from and complementary to captions, which typically provide figure context or commentary, alt text descriptions convey figure *content* including that which may be visually apparent.

Despite its vital role, alt text is often absent or of inadequate quality in scientific papers [94, 360]. One key reason for this that has been elucidated in prior work is that authors face challenges in producing high-quality descriptions [547]. Scientific figures can depict intricate concepts and relationships through numerous visual encodings, making translation to textual descriptions cumbersome. Authors must determine which aspects to describe and how to adequately convey them. Additionally, accurate and complete alt text requires deep domain knowledge and contextual insight into the figure's interpretation. This can make it challenging for non-authors and/or automated systems lacking such expertise to replace or supplement author effort. As such, insufficiently detailed or even entirely lacking descriptions remain prevalent.

Guidelines designed to assist authors in writing effective alt text[1,2] are often narrow in scope as they focus on specific figure types like line or bar plots, or tree diagrams. This makes it difficult for authors to extend their principles more broadly, such as to compound figures [547]. Similarly, though fully automated approaches are rapidly improving in quality, they are often also constrained to specific types of figures such as plots or natural images, limiting applicability to scientific communication more broadly which often involve complex diagrams and multi-part scientific figures [188, 498, 556]. Model errors also risk creating misinformative alt text if authors or publishers over-rely on automated methods. Despite advances in computer vision and natural language processing for processing and describing many types of images [200, 300, 580], scientific figures often contain nuanced details and contextual factors that might hamper the applicability of these models to producing high-quality accessible descriptions. As such, it is important to effectively engage authors, equipping them to critically review and refine auto-generated descriptions.

More specifically, there is a need for methods that generalize across authors' open-domain figures while providing tailored guidance within interactive alt text drafting workflows. To inform the design of such an interactive alt text authoring system, we first conducted a formative study with authors (N=6). This study used an initial prototype which provided authors suggestions during drafting, using large language models conditioned on figure metadata. The study revealed needs for more guidance during drafting and increased control over text generation.

[1] http://diagramcenter.org/table-of-contents-2.html
[2] http://diagramcenter.org/poet.html

Based on these findings, we developed an interactive system for alt text authoring[3] that combines human and AI capabilities, with specific features detailed below. Authors upload their paper into the system, which then automatically extracts figures along with corresponding captions, mentioning paragraphs, figure text, and an estimated data table (for plots). This establishes a knowledge base to use for AI suggestions. Subsequently, a suite of drafting and editing features ranging from pre-generated drafts to iterative description snippet generation and queries to elicit author input are provided to assist them in efficiently producing detailed alt text. We conducted a within-subjects user study to evaluate our system (N=14), which, to our knowledge, is the most realistic and general study of AI-assisted alt text authoring to date, with authors writing descriptions for their own figures across a diverse set of figures and fields of study.

Overall, our work contributes:

1. A formative study (N=6) with authors, using a technology probe which offered alt text writing suggestions. This revealed needs for guidance, control, and varied suggestions.

2. An automated pipeline to generate descriptive draft alt text for open-domain figures without requiring ground truth data. It uses an ensemble of methods to extract metadata, assembles knowledge-based prompts, and uses large language models for generation. In contrast to prior work, this is training- and data-free, fast, generalizes to arbitrary figure types, doesn't require ground truth data or scene graphs, and allows us to incorporate existing accessibility guidelines, all necessary features for real-world alt text applicability.

3. An interactive alt text authoring system which (A) scaffolds alt text production by providing extracted paper and figure context with figure type-specific accessibility guidelines to support reviewing and revising generated drafts, and (B) two additional features: *Generate at Cursor*, which interactively expands descriptions at user-directed points based on writing support approaches, and *Potential User Questions* (and suggested answers), which prompt authors to address ambiguous elements following from prior work using queries and templates.

[3]Code available at https://github.com/allenai/figura11y

4. A within-subjects user study (N=14) where authors described their own figures, mimicking real-world use. Findings show the system assisted rapid drafting and editing of descriptive alt text through different strategies based on author needs. Interactive features enhanced experience without increased cognitive load or effort on average, and enabled greater deviation from generated drafts by supporting iterative refinement.

9.2 Related Work

9.2.1 Figure Accessibility in Scientific Publishing

A systematic analysis of alt text practices across scientific disciplines has yet to be conducted. However, smaller-scale studies highlight significant issues with low alt text prevalence. For example, only 4.6% of figures had valid alt text in a sample of Accessibility and HCI papers [94], despite explicit author guidelines from venues in these fields. Even lower prevalence has been observed in other fields like biomedicine: an examination of recent papers from 16 leading biomedical and ophthalmology journals found no meaningful alt text beyond basic information [360]. While these results highlight the need for improvements, more work is required to characterize issues in figure accessibility across domains. Still, in response to the lack of quality alt text, we aim to address a broad set of scientific figures without an explicit domain or type constraint.

Beyond prevalence, the quality of alt text is also important to consider. Web accessibility guidelines suggest that alt text should convey the same information or function as visual content[4]. Scientific figures are information-dense, making full coverage of relevant information difficult to judge. Rubrics have been proposed to assess the descriptiveness and structure of alt text content. Williams et al. [547] developed a rubric to assess the overall descriptiveness of figures in HCI papers, building on prior work for other types of images [187]. Lundgard and Satyaranayanan proposed an influential four-level semantic model of descriptions for data-driven figures like plots [317]. This model decomposes the descriptions of such figures into *elemental and encoded*, *statistical and relational*, *perceptual and cognitive*, and *contextual and domain-specific* content. We factor this semantic model into our system, in order to steer language models towards generating structured, meaningful descriptions and

[4] https://www.w3.org/WAI/tutorials/images/

suggestions.

9.2.2 Author Challenges in Alt Text Writing

One well-documented reason for inadequate alt text is that authors face challenges in effectively describing figures. Interviews by Williams et al. [547] reveal that their author participants were confused about what information to include in the alt text (or, as one participant put it, "what's missing" beyond the figure caption). Their results point out that interviewed authors wanted advice on the structure and content of their descriptions, given the density of visual elements and relationships depicted in their figures.

Guidelines are a traditional mechanism by which authors have previously been supported in writing alt text. For example, SIGACCESS provides guidelines for computing publications[5], the American Chemical Society (ACS) provides guidelines for ACS authors[6], and the multidisciplinary publisher *Taylor & Francis* provides guidelines for authors submitting to their journals[7], among others. However, guidelines are often based on example figures. Authors must interpret such guidelines and adapt them to their own figures and even figure types. Additionally, the content of guidelines can be difficult to interpret. For instance, such guidelines often emphasize brevity, but this can come at the expense of accessibility, especially for complex figures as Williams et al. also note. In our system for supporting authors, we leverage guidelines to generate figure-specific drafts and suggestions.

9.2.3 Automated Image Description Generation

Early work in automated image description, often associated with computer vision, typically relied on detecting objects and relations or constructing patterned templates [148, 272]. While these initial approaches produced reasonable descriptions for constrained domains of images, they were limited in flexibility and language quality. Additionally, figures frequently exhibit higher complexity than natural images. Recent work has explored automated figure captioning [21, 218], including doing so with knowledge-augmentation in the form of mentioning paragraphs and OCR text [568]. We use a similar knowledge-augmented approach in conjunction with alt text guidelines and zero-shot large language model inference to achieve broad

[5] https://www.sigaccess.org/welcome-to-sigaccess/resources/describing-figures/
[6] https://pubs.acs.org/doi/full/10.1021/acsguide.60108
[7] https://www.tandfonline.com/pb-assets/tandf/authors/tf-alt-text-guide-1636994956097.pdf

applicability for figures without training, especially since datasets of high-quality alt text for diverse scientific figures are not readily available.

More recently, large language and multimodal models have been used to generate improved image descriptions [537, 580]. Language models can produce varied, high-quality text conditioned on input information, making them useful for this task. In multimodal models, this input can also often be visual [5, 295, 302, 458], allowing direct input of figure information. However, scaling and providing interactive access to cutting-edge multimodal models remains challenging due to computational demands and rapid changes in their capabilities. Additionally, these models' visual capabilities are still error-prone and often not evaluated on tasks as complex as describing figures in scientific research.

An emerging solution for improving vision-language reasoning is to decompose the task into vision and reasoning components through a number of different strategies [199, 493, 552, 567, 570, 580]. This can allow using separate specialized models for each part. For example, dedicated vision models can efficiently handle image information extraction as a frontend, while language models can focus on reasoning over these visual features. This has been done for general images, but also leveraged for tasks like question-answering based on plots and charts [300]. Prior work has also shown promise in generating descriptions for data visualizations from metadata alone. VisText [498] produces descriptions for plots based on data tables and scene graphs available during visualization design. Interestingly, this work's experiments found visual inputs did not improve over metadata-only methods. For open-domain figures, visual information could still be advantageous. However, these results demonstrate language models can perform well (in the plot domains covered by their models) given sufficient contextual information.

Our methods stem from a similar motivation, but we tailor them to open-domain scientific figures without original data or metadata available. Since describing figures is knowledge-intensive, we look beyond just visual features to extract contextual information from paper text and writing guidelines. Integrating this knowledge aims to assist language models in producing high-quality, tailored alt text suggestions by providing critical context beyond what is visually evident. Overall, our approach selectively combines strengths of language models, computer vision models, knowledge extraction methods, and human input to provide robust assistance for authoring accessible figure descriptions.

It is important to note, however, the rapid advances occurring in multimodal models. Future vision-language models might well provide strong automated generation capabilities. However, we believe supplementary information and human interaction will remain valuable. Extracted paper content can provide essential contextual knowledge beyond visual inputs, both for generation and revision. Human workflows also enable assessing accuracy, eliciting additional details, evaluating coverage sufficiency, and customization.

9.2.4 Alt Text Writing Support

Crowdsourcing has been identified as one viable avenue for communicating visual information to blind and low vision (BLV) users. VizWiz is an influential platform that leverages crowdsourcing to describe visual content in real-time [48]. However, extending this paradigm to scientific figures poses challenges: describing figures often requires additional domain knowledge, as well as added effort to ensure accuracy and detail. Other recent work has explored how crowdsourcing can be combined with other strategies, including automation and retrieval, to generate alt text for images on Twitter [188]. Like this work, we rely on a human-in-the-loop, specifically an author, and propose a suite of features to allow figure-specific description workflows. We introduce a collaborative AI-based system in order to distribute the workload of producing detailed alt text for complex figures. Rather than asking crowd-workers to acquire sufficient knowledge of the figure, we represent extracted knowledge as a structured prompt for a language model which can rapidly create content for the author to evaluate and incorporate.

Work targeted to author support has explored templates and queries. Morash et al. [350] explored the use of templates to elicit information from non-specialists in order to produce effective alt text. They found that this *queried image description* (QID) approach resulted in improved results compared to *free response image description*. Mack et al. [323] observed that templates helped authors write better alt text compared with automatically generated options, which were brief and regarded by authors as a gold standard, leading to reduced final quality. Text generation has made significant strides in recent years, however, which result in generated descriptions no longer being limited to brief and general content. Further, templates require per-image crafting. We generalize queries into our *Potential User Questions* feature which leverages text generation to elicit author input on possible ambiguities.

These questions are also motivated by VizWiz's approach, which treats questions and answers as a mechanism for making images non-visually accessible.

9.2.5 Language Models for Writing and Editing Support

Large language models (LLMs) have recently shown promise in providing contextual suggestions for diverse writing and revision tasks [113, 183, 289]. A common application is to open-ended tasks such as creative writing, which allow for wide-ranging suggestions useful for inspiration [345, 471, 572]. In contrast, alt text requires faithfulness to the source visual information. It has aspects in common with *expository* writing tasks [459], requiring steps such as reasoning over and synthesizing information, and facilitating composition. Our approach aims at these components in the specific case of alt text writing. Extracted information provides a knowledge base for faithful generation. Refinement interactions support accuracy verification and content enhancement. Together, these aim to leverage the capabilities of advanced LLMs to assist authors in efficiently producing high-quality, accessible figure descriptions.

9.3 Formative Study and Tool Design

9.3.1 Initial Prototype

We created a high-fidelity interactive prototype, serving as a technology probe, containing early versions of two key features designed based on reviewing prior work and proposing methods to generalize across open-domain figures: text continuations and question-answer pairs. The continuations appended generated text conditioned on figure metadata, a common strategy in writing support which allows suggestions that build on user-authored text [361, 471, 572]. The question-answer pairs were motivated by *queried image description* [350] wherein authors were provided a predetermined series of guideline-derived questions depending on figure type. With this feature, we aimed to highlight elements the author might consider describing for the figure.

9.3.2 Formative Study

We conducted a formative study using this initial prototype with paper authors (N=6) to inform the design of our main system. Participants had varying levels of experience with authoring alt text, ranging from none to over 5 years of experience. We used

a think-aloud protocol and semi-structured interviews during 45-minute remote sessions conducted via video-conferencing. Participants were provided with access to our prototype. We asked participating authors to verbalize their thought processes while trying out these features on their own figures. Session audio and screencast were recorded and transcribed. One member of the research team performed inductive thematic analysis on the data via open coding, guided by discussion with the full research team.

9.3.3 Feedback and System Redesign

Through analyzing and interpreting participant feedback, we identified key design goals for improvement:

DG1 More guidance during the drafting process, such as feedback to ensure authors provide sufficient coverage of key information in their descriptions. For example, one participant (P4) suggested the system could provide a hypothetical question like *"you didn't actually mention anything about the axes, do you want to do that?"* to prompt the author to describe missing elements.

DG2 Increased control over where and how much automatic text generation occurs within the description, e.g. targeted expansions of specific parts based on author needs. For instance, P6 suggested *"what I would really like is something... where I can put my cursor somewhere and say get continuation from here."* P3 proposed a similar interaction: *"Something that could be cool is if I could highlight something and then say generate more about this."*

Some participants also noted that the two suggestion types (continuations and Q&A pairs) could emphasize similar information, though in other cases participants found both independently useful. This highlighted an opportunity for differentiating the two to provide complementary guidance. For example, P5 noted: *"a continuation I could see including some information that I might not have thought would be relevant... the [Q&A pairs], I could also see that working in a similar way,"* while P1 noted that *"They felt useful for different things"* like the continuation helping to create an *"outline or skeleton."* Based on this feedback, we concluded that differentiating these suggestions would help offer both benefits, i.e., interactively creating outlines and highlighting missed or ambiguous content.

In response, we implemented modified versions of the original features to provide

potential user questions and on-demand text generation at user-selected points which we call *Potential User Questions* and *Generate at Cursor* respectively, described in detail in the following section. To compare with, we also implemented a simplified interface without these two suggestion features. In summary, author feedback highlighted needs for improved guidance and disambiguating suggestion types. Our redesign addressed these by providing targeted author feedback and control over text generation.

Additionally, our pilot interface positioned metadata in a menu, and used this information to prompt for suggestions. However, we observed participants referencing this metadata to get context for beginning, editing, and evaluating their descriptions and the system's suggestions. To account for this usage, we moved the metadata into the main interface so the user can easily cross-reference as needed.

9.3.4 Improving the AI Assistance

In addition to improving the features, we also sought to improve the quality of the AI assistance provided. We iterated on model choice and prompts by optimizing on a development set of scientific figures and captions.

Figure Sampling for Development Set We constructed a development set of figures and metadata to iterate on the AI assistance. We started with the SciCap [218] challenge[8] validation set, which contains figures, captions, and paragraphs for a large number of figures. Initially, we observed imbalance in figure types and research fields. We quantified this using pretrained classifiers for figure type (DocFigure [236]) and field of study from the mention paragraphs (S2-FOS[9]), finding highly skewed distributions. Figures from Physics were highly overrepresented, as well as line plots. We resampled with replacement to the third most populous categories for type and field, then dropped duplicates to obtain a broadly representative set without overly distorting the original distribution. To create a modest-sized development set, we embedded and vectorized the figures using CLIP [402], caption and mentions using SPECTER [103], and figure type as one-hot encodings. We used a facility location submodular optimization algorithm from the apricot [451] package to efficiently select a diverse subset of 30 figures. We confirmed through manual review that the figures had low content overlap, visual distinctiveness, and representation across

[8] http://scicap.ai/
[9] https://github.com/allenai/s2_fos

scientific fields. We used this set to iterate on prompts by generating descriptions for these figures and spot-checking the results.

Guidelines for Suggestion Generation In feedback from the formative study, authors noted that there were errors in some of the AI suggestions. In response, we updated the OCR model used for figure text extraction from Tesseract[10] to EasyOCR,[11] which produced more accurate textual figure representations. We also greatly expanded the set of guidelines used; initially we only implemented two sets of guidelines for plots and non-plot figures respectively. In the re-designed system, we collected an extensive set of guidelines from sources including the DIAGRAM Center[12] and SIGACCESS[13] guidelines. We adapted these guidelines to (1) remove references to specific example figures, (2) remove presentational guidelines, such as conciseness, and focus on those relating to content, and (3) organize them as a nested list indexed by figure type. From the previous version, we maintained general guidelines applicable to all figures, and additional guidelines for all plots including the first three levels of Lundgard and Satyanarayan's four-level model [317]. We included the fourth in the formative system, but removed it based on the observation that the first three levels are more often found useful by end-users, and that the fourth typically requires significantly more exogenous context to integrate, which may not be available from the extracted metadata.

Base Model Selection We compared baseline generations from GPT-4 [373] and LLaVA [302] for 5 figures sampled randomly from our development set of 30. Among the outputs, GPT-4 tended to produce more descriptive alt text with fewer hallucinations. This, coupled with the higher likelihood of LLaVA failing to generate any alt text at all, led us to choose GPT-4 as our base model. This choice can be reconsidered in the future with the emergence of more powerful language and vision-language models. Note that GPT-4 with Vision was not available for comparison at the time that this work was conducted.

Prompt Engineering We identified unhelpful patterns in model-generated suggestions through testing and observations during the formative study. To address these,

[10] https://github.com/tesseract-ocr/tesseract
[11] https://github.com/JaidedAI/EasyOCR
[12] http://diagramcenter.org/table-of-contents-2.html
[13] https://www.sigaccess.org/welcome-to-sigaccess/resources/describing-figures/

we made several prompt adjustments:

1. Added instructions like “Respond with only x” to avoid chat-like responses and keep suggestions focused on the requested task (e.g. text continuation).

2. Added an instruction and logit biases to avoid explicit references to metadata. Metadata should inform responses without being directly referenced (e.g. "the OCR-extracted text contains..."). We experimented with reiterating the instruction at the end of the instruction set, finding this further reduced such occurrences.

3. Motivated by prior work [580], we added an “uncertainty prompt” to mitigate sensitivity to metadata extraction errors. In our version, we acknowledge they may exist and encourage inferring details despite this to provide helpful suggestions.

4. Added instruction to focus on the figure visual metadata and key information, reducing suggestions derived from the text that do not describe visual aspects of the figure.

Although it is difficult to systematically evaluate the effect of such changes or their reproducibility, we include them here to describe our design process for improving AI assistance and mitigating observed issues.

We also created prompt variants for different contexts like generating initial summaries versus later continuations, adding placeholder text and instructions to improve infilling around user text. The appropriate context is inferred from the system’s state:

1. Initial High-Level Summary Prompt: Generates a high-level summary when no description exists yet.

2. Continuation + Infilling Prompt: Extends existing text by referencing the description context. For infilling, we tested multiple strategies:

    - Naive infilling (without providing post-cursor context): often resulted in duplicate content

    - In-prompt context (adding post-cursor text as a prompt metadata element): still resulted in duplicate generations

- Placeholder text at the cursor position: reduced duplication, so we selected this approach

3. Draft Prompt: Variation on the initial high-level summary prompt to generate a full description, used for pre-generating drafts.

4. A separate prompt for generating *Potential User Questions* and corresponding suggested answers.

9.4 System Design

FigurA11y consists of a backend architecture for processing and extracting figures from scientific PDFs (Section 9.4.1), figure metadata extraction (Section 9.4.2), and figure description prompting (Section 9.4.3), as well as a user interface for AI-supported figure alt text writing (Section 9.4.4).

9.4.1 Overall Pipeline Architecture

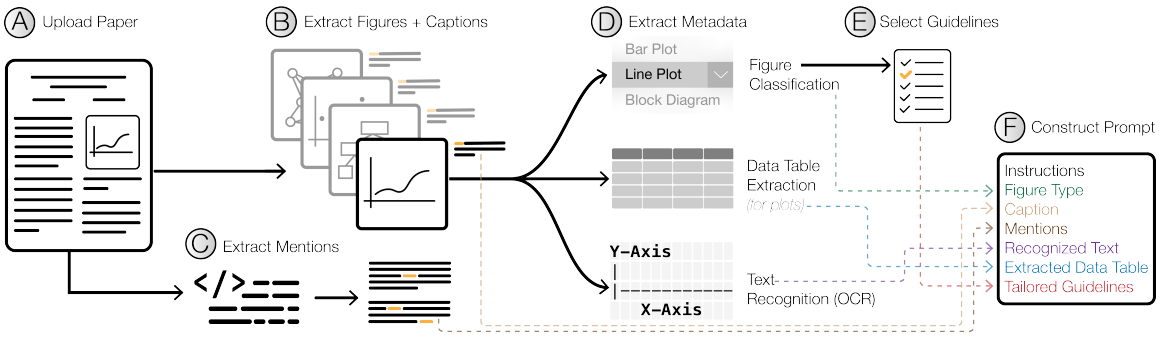

Figure 9-1: Pipeline for extracting information from figures, and using this information in a prompt to generate draft alt text and suggestions for enhancement. The author first **(A)** uploads a paper, from which **(B)** figures and their captions, and **(C)** mentions of each figure in the paper are extracted. Then, **(D)** the figure is classified, a data table is extracted if it is a plot, and the figure text is recognized. Finally, **(E)** based on the figure type, a set of guidelines are selected. **(F)** all of this information is put together with instructions into a prompt for the LLM to use in generating drafts and suggestions.

The overall system architecture consists of several steps, as depicted in Fig. 9-1. The first stage involves uploading an academic paper in PDF format. The system then extracts figures, captions, and paragraphs mentioning each figure. The figures and captions are extracted using PDFFigures2 [100], and the mentioning paragraphs

are extracted from the paper using GROBID [312] to extract the text and a regular expression to match mentions with the figure number in each caption. These methods are similar to those used in recent work on knowledge-augmented figure captioning [568], but we incorporate the caption as well since our goal is to generate alt text, in addition to information extracted from figures and hierarchical guidelines (reviewed next).

9.4.2 Metadata Extraction

Metadata is then extracted from each figure, including classifying the figure type using the pre-trained DocFigure [236] classifier (e.g. bar plot, tree diagram, etc.). We focus our methods on plots and diagrams, and so we construct an "Other" figure category to account for figure types outside of plot and diagram sub-types. For plots, the plot data table is extracted using the pre-trained UniChart [333] model by default, or the DePlot [300] model if desired. The latter is slower, but we find that it sometimes yields better results, depending on the figure. Text in the figure is extracted using EasyOCR,[14] with the layout preserved; UniChart's results [333] suggest that this can help with LLM reasoning over charts, and we adapt this to our context of open-domain scientific figures.

Finally, this extracted information is assembled into a prompt for the LLM to use in generating text, along with tailored guidelines based on figure type. The full pipeline synthesizes disparate recommendations from prior work, as discussed above, to construct a prompt with tailored instructions and hierarchically-selected guidelines depending on figure type. We next discuss the structure of this prompt.

9.4.3 Prompt Structure

We use structured prompts to effectively harness large language models (LLMs), specifically GPT-4 in our current system, for generating useful alt text suggestions. Prompts contain two main elements, described in more detail below.

Instructions

We designed several prompts with detailed instructions supporting the core interactions: content generation and potential user questions, as briefly described in the previous section.

[14] https://github.com/JaidedAI/EasyOCR

- **Initial Summary**: instruct the LLM to introduce the figure in 1-2 sentences focusing only on the most important elements and relationships shown, without additional commentary.
- **Continuation**: prompt the LLM to expand on the existing description by adding 1-4 sentences conveying missing details and relationships relevant to understanding the figure, avoiding repeating content already provided.
- **Infilling**: use the *Continuation* prompt but with added context. Includes placeholder text at the cursor location and instructs the LLM to provide distinct suggestions that fill in gaps within the existing description section.
- **Drafts**: adapt the *Initial Summary* instructions to generate a full figure description.
- **Potential User Questions**: instruct the LLM to generate pointed questions querying unclear visual elements that need explanation in the description. These come with 1-4 suggested answers each, also generated by the LLM based on figure metadata. To maintain this structured format in the generations, we use OpenAI's function calling API[15]. We define a function which accepts a question, along with 1 required answer argument and 3 optional answer arguments, to construct the question-answer sets.
- **Summarization**: provide a brief summary ($\sim$1 paragraph) of longer alt texts, to align with guidelines around conciseness and short/long alt text versions[16,17].

The complete prompts are provided in Appendix A.6.2.

Metadata

As noted in Section 9.4.2, prompts incorporate metadata extracted from the figure and paper to ground the LLM's generations. These act as visual information for the LLM to reason over, allowing us to leverage its advanced information processing capabilities without relying on newer, less robust multimodal approaches that result in less descriptive and sometimes empty outputs as we found in our prototyping (Section 9.3.4). Additionally, VisText [498] found that metadata-driven description

[15] https://openai.com/blog/function-calling-and-other-api-updates
[16] https://www.w3.org/WAI/tutorials/images/complex/,
[17] https://authors.acm.org/proceedings/production-information/describing-figures

outperformed visually-improved description in their case, for plots. Though our circumstance differs (our two model options are not directly comparable), we also find that the combination of metadata we use can produce detailed and grounded descriptions.

- **Figure type** provides high-level context.
- The **caption** often summarizes main ideas depicted and can contain useful details about visual elements.
- **Mentioning paragraphs** give further context from the paper, e.g., describing key concepts or results shown.
- **Extracted text** conveys lower-level visual details like axis labels and diagram text.
- For plots, the **extracted data table** approximates the values depicted.

9.4.4 Interface Design and Implementation

The FIGURA11Y interface was designed to provide authors with AI-assisted support throughout the alt text drafting process, while scaffolding the review and revision process by concisely presenting figure metadata. The left side of the interface displays the figure along with extracted metadata like the figure type, caption, paragraphs which mention the figure, and extracted data values for plots (see Fig. 9-2 for the **Interactive Assistance** version, for instance). These metadata components serve as prompts to inform the initial AI-generated draft and subsequent suggestions.

The right side contains the main alt text authoring field where authors can write and iteratively refine descriptions. The **Interactive Assistance**'s two augmentative features are engaged by clicking buttons in the authoring field's toolbar, or using the corresponding key commands: `TAB` for *Generate at Cursor* and `(CMD|CTRL)+/` for *Potential User Questions*. The results of the former are shown in the description field; the generated text is highlighted in red and with a differently tagged underlying HTML element. Then, the user can click on a generated snippet, and decide whether to accept or reject it. If accepted, it becomes part of the description and the special formatting is removed. If rejected, it is discarded.

In **Draft+Revise** (shown in full in Fig. 9-3), the interactive features are replaced with a simple text box with which to prompt GPT-4 as the user desires, to simulate access

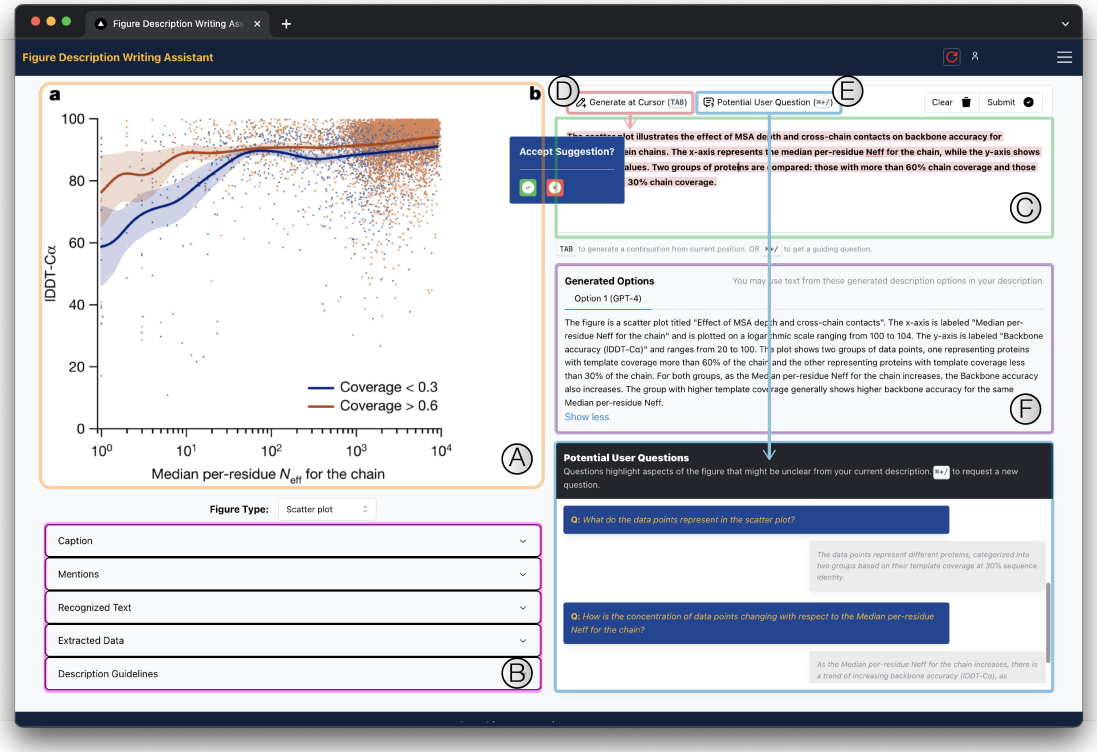

Figure 9-2: Screenshot of our **Interactive Assistance** alt text authoring assistant interface. On the left, it shows **(A)** the figure and **(B)** extracted metadata. On the right, it shows **(C)** the description authoring field, **(D)** the *Generate at Cursor* feature with generated initial text below, **(E)** the *Potential User Questions* request button and results, and **(F)** a pre-generated draft description. Example figure is taken from [240].

to an LLM as authors may have in their normal writing workflows. After drafting in either of the system versions, authors can run the summarization workflow. Additional interface features are described in Appendix A.6.4.

The full system was implemented in around 6100 lines of TypeScript and 2000 lines of Python using ReactJS, Next.js, Zustand, and Mantine for the frontend interactions, Tiptap and Prosemirror for the interactive editor specifically, Flask for the backend server, and PostgreSQL for the database.

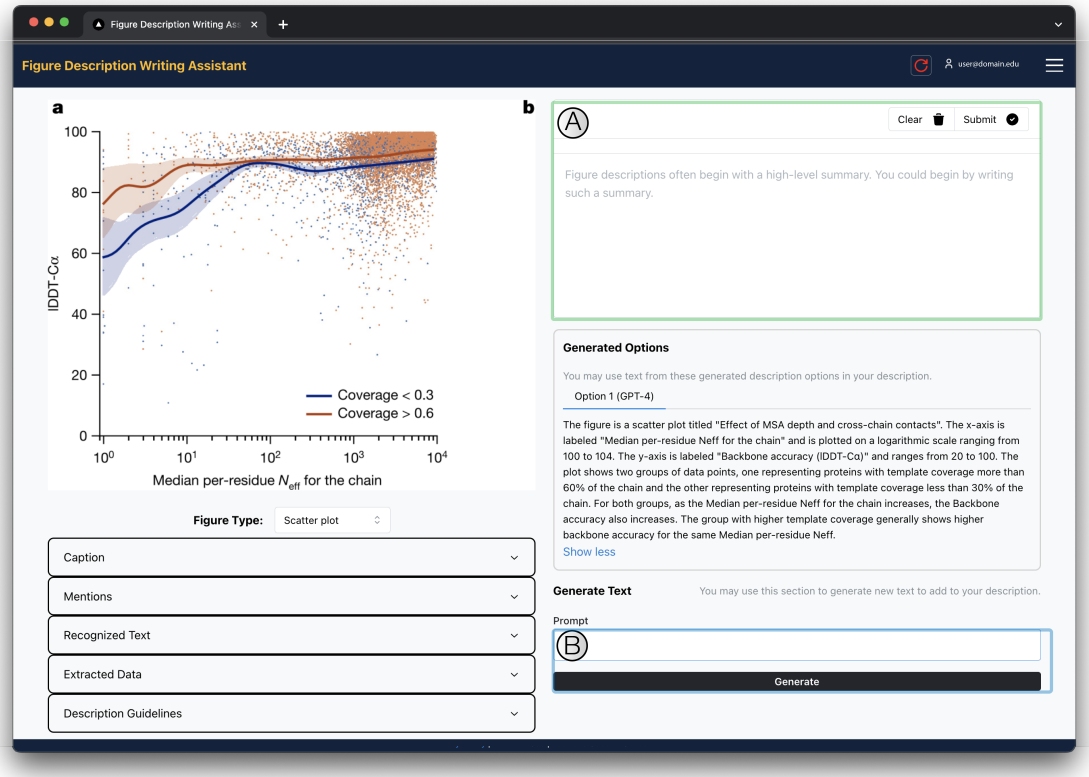

Figure 9-3: Screenshot of our **Draft+Revise** alt text authoring assistant interface, showing some of the same features as the **Interactive Assistance** version: figure and metadata on the left side; and the description authoring field and a pre-generated draft description on the right side. However, there are two differences: **(A)** the description authoring field does not contain the *Generate at Cursor* and *Potential User Questions* features, and **(B)** we provide a box to freely prompt the LLM to generate text that the author can integrate into their description. Example figure is taken from [240].

9.5 Study Design

We designed a study to evaluate the usefulness of our system for assisting authors in producing alt text. In particular, we sought to examine (1) whether authors perceive benefit from our pipeline's scaffolding and pre-generated drafts, (2) if the added interactive features in **Interactive Assistance** support authors in further enhancing descriptions beyond editing pre-generated drafts, (3) whether added features incur additional cognitive load, and (4) what strategies participants used when integrating our tool's features into their alt text authoring workflows.

Rather than using a standardized task with predetermined figures, we chose to conduct the study with authors describing figures from their own recent papers. Since our prototype aims to support alt text writing across diverse open-domain figures, it was essential that our lab study be grounded in a realistic context using authors' knowledge of their own content. Our formative results and prior work have also emphasized authors' contextual knowledge as essential for informing alt text drafting.

Beyond assessing overall usefulness, our goal was to understand how different features supported the process of creating complete and accessible descriptions. To compare feature sets, we used a within-subjects design with the two system versions discussed earlier: **Draft+Revise** and **Interactive Assistance**. The **Draft+Revise** condition allowed us to evaluate the draft-generation pipeline and overall revision-support interface. The **Interactive Assistance** condition focused on specific writing assistance interactions. Using both versions allowed us to gather comparative insights. We did not include a baseline without access to any generated text because we do not believe it is realistic to restrict author access to LLMs, given their wide use; however, we note that **Draft+Revise** is a strong baseline not previously available to alt text writers, as it uses our refined alt text draft generation pipeline.

9.5.1 Materials: Figure Selection

We invited authors recruited for the study to share two to three recent papers containing figures for which they had not yet written alt text. We extracted figures from these papers, and then selected two figures per participant (one for each system version condition).

One challenge with this design is that participants could apply the guidelines and suggestions from the first condition to the subsequent condition, if the figures are sufficiently similar. To avoid this, we aimed to select different figure types within participants when possible, typically one chart and one diagram. In cases where participants did not have both types available (e.g., results presented in tables instead of charts, as is common in some domains), we aimed to select substantially different instances (e.g., different plot types, or diagrams that were visually very distinct and did not represent overlapping information).

A second concern was figure complexity. Since figures have a different prior com-

plexity for description tasks (e.g., by being compound, or having many variables or components), varied complexity could produce biased results. Since there is no validated metric for the complexity of scientific figures, we aimed to minimize the impact of this in two ways. First, we randomized the assignment of figures to conditions within participants. This ensured that figure complexity does not systematically factor into the difference between conditions. Second, given our small participant pool, we sought to further reduce this bias. We heuristically selected figures with comparable numbers of visual elements (prior to random assignment) and, if this was difficult to determine, overall subjective complexity. This was to minimize large mismatches in complexity between a participant's two figures, subject to the availability of figures from participants' submitted papers.

We pre-loaded figures into our system to save participants time and effort during the study compared with the full workflow of paper upload and figure selection. We wanted to focus the tasks on writing the alt text itself. Participants were given URLs with figure IDs, which pre-populated the interface with the figure information.

9.5.2 Study Procedure

We conducted this study remotely via video-conferencing. Participants were assigned to one of two counterbalanced groups determining the order of writing with the two system versions. Group 1 used the **Draft+Revise** version first, followed by **Interactive Assistance**. Group 2 used the reverse order. For all tasks, participants were instructed to write descriptions that were as descriptive as possible, rather than aim for conciseness. This allowed for participants to take a more consistent approach towards maximizing information content to make accessible, rather than employing intuitive strategies for conciseness, and also to avoid challenges authors face deciding whether to include a piece of information [547]; we believe that given diverse alt text reader preferences [317] that reader customization should happen at a later stage. At the end of the workflow we provided a semi-automated step to allow authors to create a more concise version.

The study procedure consisted of four main components. First, participants were given a brief 5 minute introduction to alt text and shown examples of effective alt text for a tree diagram and scatter plot from the DIAGRAM Center guidelines. They also received an overview of the study tasks and timeline. Second, there were two 10 minute alt text writing sessions, one for each system version. We determined

this time through piloting and observation during our formative study. Participants were allowed to conclude each writing session early, if desired (e.g., if they felt their description had saturated available information to describe). Prior to each one, the experimenter provided a brief, structured walk-through of the features available in the interface. Each session was followed by a 5 minute survey gathering feedback. After the second survey, participants completed an additional 5-10 minute comparison survey. For the first few sessions and those ending with sufficient time remaining, we also conducted a semi-structured follow up interview probing participants' overall impressions, the usefulness of different features, and the strategies they employed beyond what we observed. In these interviews, we asked participants to walk us through their process writing alt text with each system, to offer feedback, and additional questions based on their interactions and comments. This multi-stage procedure allowed us to observe system use, gather both immediate and retrospective feedback, and have an open-ended discussion.

9.5.3 Recruitment and Participants

We recruited participants using the authors' academic social networks, snowball sampling, and institutional mailing lists. Our study included a total of 14 participants: 9 women, 4 men, and 1 non-binary participant. Their ages ranged from 18 to 44 years old, with most (10 participants) aged 25–34. In terms of roles, there were 7 graduate students/research assistants, 3 postdoctoral researchers, 2 assistant professors, 1 lawyer and researcher, and 1 scientific assistant. The participants' fields of study were diverse, including 5 in formal sciences like computer science and math, 3 in applied sciences like engineering and medicine, 3 in human-computer interaction or design, 2 in social sciences, and 1 in information sciences. The participants also varied in their amount of prior research experience, with 4 having published 1-5 works, 4 having published 5-10 works, 2 having published 10-20 works, and 4 having published over 20 works. Most participants indicated that the majority or all of their prior published works contained figures. However, many had limited experience writing alt text for these figures, with 5 having written no alt text previously and 7 having written alt text for 50% or less of figures. In terms of familiarity with alt text guidelines, 6 were somewhat familiar and 2 were very familiar, while 6 were not very or not at all familiar. When asked about AI writing assistants, 8 had tried them before, 4 used them regularly, and 2 were aware of them but had not used them.

9.5.4 Data Collection, Evaluation Methodology, and Measures

Questionnaires

Participants completed the following:

1. Cognitive Load and Usability (completed after each system variant):
    - NASA TLX dimensions: mental demand, temporal demand, effort, frustration. We also included own performance, but factor it out in our analysis to differentiate self-assessed performance from experienced cognitive load.
    - A usability or system acceptance scale based on recent work on AI assistance [260].
2. Comparative Preference: A single preference rating on a divergent scale ranging from 1 (**Draft+Revise**, referred to as *Without Suggestions*) to 7 (**Interactive Assistance**, referred to as *With Suggestions*).
3. Open-Ended Questions: A set of questions covering topics such as in which situations the system variants were helpful or unhelpful, and suggestions for improvement.

Description Measures

We computed metrics to compare the final descriptions against the generated draft. In particular, we sought to capture the degree to which participants' descriptions diverged from these drafts. We assessed this using a range of metrics like the Levenshtein edit distance [291] and Zlib-based normalized compression distance (NCD) [99], using implementations from the `textdistance` package[18]. We also used cosine similarity of embeddings produced by the `all-distilroberta-v1` [414] pretrained language model from the `sentence-transformers` package[19] for a less length-sensitive and more semantic view.

Logs

In addition to logging participants' descriptions, we logged key presses (split into "Input" (additions) and "Backspace or Delete" (deletions), as well as whenever text

[18] https://github.com/life4/textdistance
[19] https://www.sbert.net/

was pasted from the clipboard (e.g. copied from the draft or suggested answer for a Potential User Question). Examining keylogs allows us to assess task effort and compare against reported cognitive load, to assess whether the added features in the full **Interactive Assistance** system induced or saved additional effort. We also examine *traces* of the interaction through these logs over time, to illustrate different strategies used by participants to produce alt text descriptions with the features available in both systems.

Screen Recordings and Transcripts

The study sessions were screen-recorded to capture participants' on-screen interactions. We also recorded audio and transcripts of the participants during interviews. These were later examined to compare against usage logs, and to keep track of observations made during the sessions.

Challenges for Evaluating Quality

We considered using a descriptiveness metric from prior work [547] to evaluate the level of detail of alt text descriptions. However, the descriptiveness measure was defined based on the range of human-written figure descriptions, with a substantial part of the scale dedicated to low descriptiveness or not descriptive alt texts. The pre-generated alt text drafted by large language models used to seed our system variants introduced a distributional shift from human-written alt text. These generated descriptions tend to be sufficiently long and detailed, such that the descriptiveness metric is no longer effective for distinguishing between pre-generated and human-edited versions of these alt texts.

During our system redesign, we piloted an annotation task with two base models generating draft descriptions: GPT-4 [373] and LLaVA [302]. We asked three individuals with undergraduate training in physical and life sciences to annotate descriptions generated for 5 figures from our development set: 2 each with GPT-4 and LLaVA, and one for each model with and without description guidelines. We adapted annotation guidelines based on the previously defined levels for descriptiveness [547], while introducing half-step levels (9 total levels) to capture finer-grained differences. We found that there was low correlation between pairs of annotators (Spearman's rho 0.246-0.462), challenging the use of this metric in our high-descriptiveness regime. Instead, we evaluate description detail through metrics like divergence from drafts

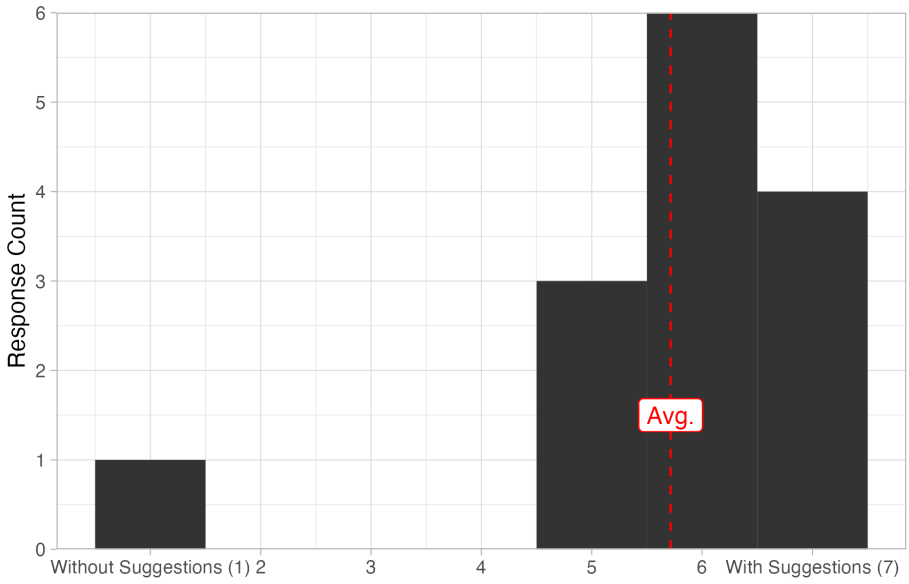

Figure 9-4: Overall participant preference between the system versions. Results favor the **Interactive Assistance** version.

and length, and leave establishing robust descriptiveness metrics to future work.

9.6 Results

Overall, the results indicate that participating authors preferred the **Interactive Assistance** system version over the **Draft+Revise** version. The **Interactive Assistance** version helped users produce longer, more detailed alt text that diverged more from the initial AI-generated drafts on average. Participants appreciated the pre-generated drafts in both systems, but found features like *Potential User Questions* and *Generate at Cursor* useful for highlighting additional details and supporting incremental drafting in **Interactive Assistance**.

9.6.1 User Preferences and Responses

Participants generally preferred the **Interactive Assistance** interface as shown in Fig. 9-4, with 13 participants indicating preference for the **Interactive Assistance** tool to varying degrees. The one participant who preferred **Draft+Revise** noted that they found the workflow of editing the pre-generated description to be less effortful. We tested that these ratings deviated from the neutral level (4) with a one sample

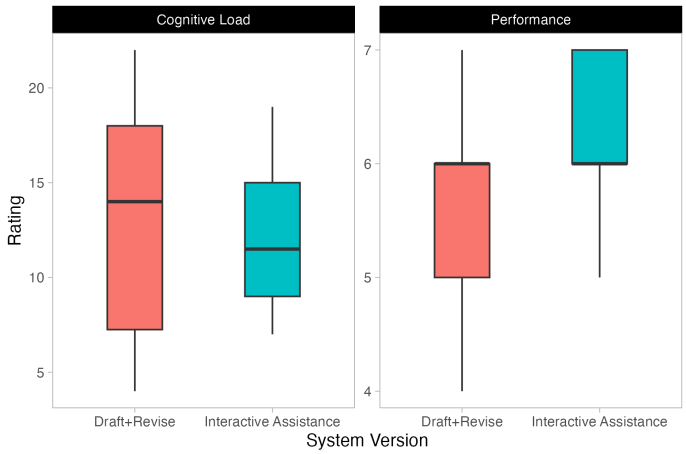

Figure 9-5: Partial raw NASA TLX results, summing the demand scores (left; with the *Physical Demand* item removed), and factoring out the *Performance* item (right). The score distributions are comparable between the two system versions, overall.

t-test, which showed a statistically significant result with a large effect ($t(13) = 4.16$, $p < 0.01$, Cohen's $d = 1.1$).

Participants who preferred the **Interactive Assistance** offered a number of reasons for this, including:

- *Potential User Questions* highlighting elements that might have been missed.
- *Generate at Cursor* allowing incremental drafting.
- *Generate at Cursor* anticipating user needs or replacing user effort in context.

Several participants found the initial pre-generated draft (available in both conditions) useful, with some even indicating that the usefulness of this option diminished the value of the *Generate at Cursor* feature.

Finally, participants identified some usability issues and potential changes to improve experience when working with their own figures. These ranged from behaviors in edge cases (e.g., rapidly double triggering suggestions produced unexpected behavior) to interface features that would assist in smoother review (e.g., visibility of multiple types of metadata at the same time).

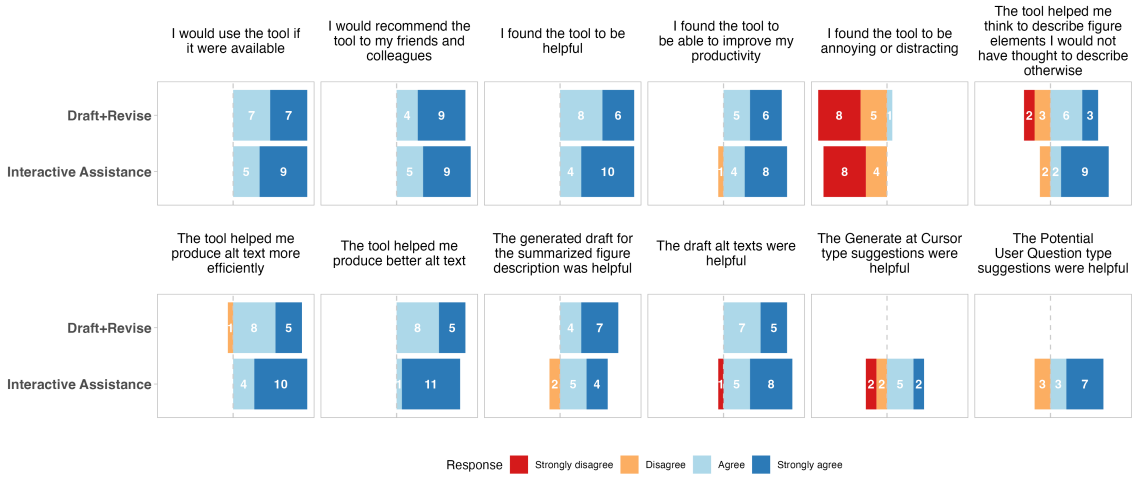

Figure 9-6: Usability and utility ratings of both versions of the system.

9.6.2 Workload, Usability, and Utility

Both system versions show comparable cognitive load (Fig. 9-5), despite the added interactive features in **Interactive Assistance**. We report the four-item raw NASA TLX score representing cognitive load without the *own performance* and *physical effort* items, and the factored-out item representing participants' assessment of their own task performance.

Usability and utility questions indicated an overall preference for the **Interactive Assistance** version, but generally favorable results for the base **Draft+Revise** version as well (shown in Fig. 9-6). The first four items relating to general tool usability and acceptance showed comparable results for both system versions. However, more participants strongly agreed that **Interactive Assistance** improved efficiency compared to **Draft+Revise**. This was also true for quality ("better alt text"); however, one fewer participant agreed overall for **Interactive Assistance** despite 11 strongly agreeing. **Interactive Assistance** was also reported to more effectively prompt participants to describe elements they may have otherwise missed, a core goal of the added *Potential User Questions* feature. The *Potential User Questions* in **Interactive Assistance** received positive feedback, and the *Generate at Cursor* feedback was mixed but biased positive as well.

While the pre-generated draft feature was identical in both systems, it was rated as slightly more helpful in **Interactive Assistance**. This could suggest it was used differently in conjunction with the interactive features. Overall, the user feedback

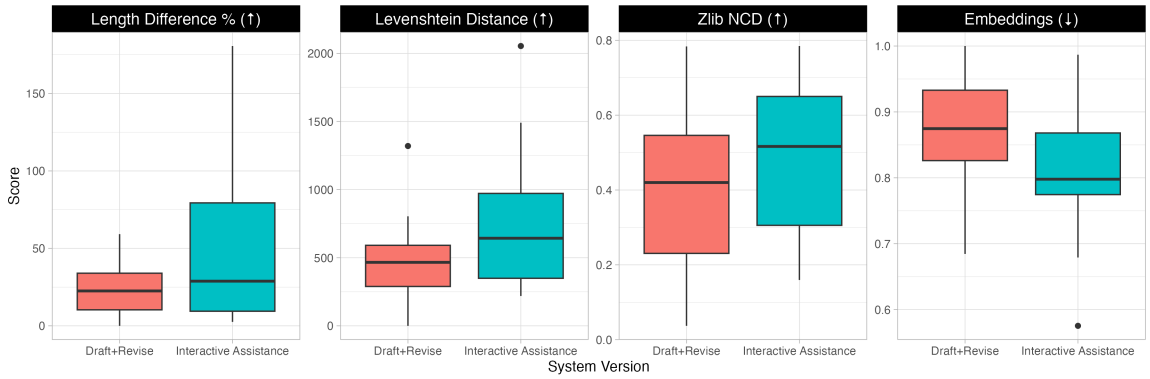

Figure 9-7: Measures of divergence between the pre-generated alt text drafts and authors' final alt text. Overall, descriptions in the **Interactive Assistance** condition deviated substantially more from pre-generated drafts across methods (note that the *Embeddings* scores are cosine similarity, and as such are inverted compared with the other metrics; higher similarity indicates *lower* divergence).

indicates broad acceptance for the core draft generation, with added value from the interactive assistance features in **Interactive Assistance**.

9.6.3 Change in Final Descriptions

One of our core design goals was to encourage greater detail and reflection from authors when writing alt text. We compute several metrics to quantify the textual divergence between the pre-generated draft alt text and authors' final alt text (i.e., how much authors edited the generated alt text) across conditions as a proxy for detail and reflection (Fig. 9-7).

Between the draft and final alt texts, we computed the absolute percentage length difference, the Levenshtein edit distance [291], the normalized compression distance (NCD) [99], and the cosine similarity of language model embeddings as computed by Sentence-BERT, `all-distilroberta-v1` [414]. Across all metrics on average, authors deviated more from the initial AI-generated draft when provided interactive assistance. This suggests that the added interactive features (*Generate at Cursor* and *Potential User Questions*) provided more opportunity for authors to revise and customize the alt text.

First, we observed descriptions in the **Interactive Assistance** condition to be longer on average than those in **Draft+Revise** (1348 vs. 1075 characters on average). On average, the **Interactive Assistance** condition saw significantly greater changes

in length (mean of ∼52% change) compared to **Draft+Revise** (mean of ∼23% change). Individual differences from generated descriptions ranged up to 150% in the **Interactive Assistance** condition. Though informative, length alone does not fully capture textual changes. The Levenshtein edit distance [291] count how many insertions, deletions, and substitutions are needed to transform one string into another. Edit distance also revealed significantly more alterations in the **Interactive Assistance** condition. However, as Levenshtein distance can be influenced by the previously reported length differences, we report two additional metrics. Normalized compression distance (NCD) [99] measures how much compressing two strings together differs from compressing them separately. Unlike Levenshtein distance, NCD is less sensitive to length differences. Additionally, cosine similarity of language model embeddings captures semantic similarity beyond length. With both these metrics, we again see greater divergence from the pre-generated drafts in the **Interactive Assistance** condition (higher on average for the former, and lower on average for cosine similarity in the latter).

9.6.4 Semi-Structured Interviews

We interviewed 7 participants (half of the total 14), selecting the first 7 whose sessions left sufficient time remaining (typically 5-10 minutes) after the interactions and surveys within the total 1-hour time slots. The variation in time available mainly had to do with time spent on surveys' free-response items, if participants arrived late to the session start, and any connectivity issues, rather than time spent drafting (though some participants did finish particular descriptions before 10 minutes). Specifically, this included P1, P2, P3, P4, P5, P10, and P12. One researcher reviewed these transcripts using a hybrid approach: deductively, to add nuance to the primary findings in the surveys and observed behaviors, and also inductively, to discover factors not covered by other feedback.

The interviews highlighted the value of different system features for prompting potentially missed details and incremental drafting. Participants appreciated the pre-generated drafts, and most also found the *Potential User Questions* and *Generate at Cursor* features useful.

Potential User Questions helped authors reflect on missing elements

Authors reported ways in which the *Potential User Questions* highlighted elements they might have missed otherwise. P1 noted *"they were asking questions of images I wouldn't necessarily think of because I've seen most of those images hundreds of times,"* and similarly P5 commented: *"The potential user question function was super helpful because it was asking some questions that I never thought of and pointing out certain points that I might miss."* P12 more broadly noted that *"it made the captions so much richer and so much fuller and better."*

Generate at Cursor encouraged thinking and supported iteration

P4, who made use of the *Generate at Cursor* suggestions to produce a very long and detailed description of a complex figure, highlighted how they found these suggestions useful: *"it could be a bit overwhelming when you're looking at the full generated text and then figure out how you're gonna tweak this... You could generate more chunks based on what you're writing as well. So it's very collaborative,"* and specifically pointed out *"the chunks are nice in the sense that they encourage you to use your own brain as you're writing, and then use that as an aid."* P2 expressed a similar idea, that *"the cursor feature... makes you think more,"* however disliked this aspect and preferred the **Draft+Revise** workflow. P3 similarly commented that *"it takes a little more time but it gives you much deeper breadth to the text."* P5 highlighted the value for rewriting: *"it's helpful for especially just editing a certain portion of the text rather than rewriting the whole entire text."*

Pre-generated drafts were a useful starting point

P1 noted how the generated drafts brought their attention to the difference between captions and alt text: *"I think comparing [the generated description] to [those] that were already written in my papers is adjusting to what alt text would look like versus just a regular image caption."* P10 remarked that the generated draft was sufficiently helpful that the *Generate at Cursor* didn't add much beyond it, *"I tried a little bit the add text at cursor... I just felt like for mine, at least, I either liked parts of the generated text better or it just wasn't really adding anything more to what I already had there."* They also commented more specifically that *"it pulled in pieces of data that I just wouldn't have."* P12 noted *"I like being able to copy the whole paragraph and edit as I needed,"* and that *"it was pretty great and then it didn't really need that much editing."*

P2 reported, when asked to compare this process to their prior experience writing alt text: *"To be honest, the alt text descriptions generated... are much more exhaustive and it's covering all the major parts."* P2 also noted having to correct an error in the generated description, but explained that this was easy to do: *"it generated text for bar chart... it somehow detected the other category which was not there, but it was very easy to do it."*

Interface features helped authors review descriptions

P10 commented that they *"hadn't looked before [for] guidelines for alt text, but it was nice to have that as a reference."* P2 noted that though they did not directly use the guidelines in their interaction with the specific figures, they *"usually have a lot of different types of graphs and I used to struggle with finding the guidelines and then again I had to open the tab and Google and search about it,"* and so could see the automatic guideline selection and presentation being useful. P2 also noted that *"the mentions were good"*, but that they did not find value from reviewing the OCR-extracted figure text and extracted data table. We also observed from participants' screens that several participants referred to the caption, mentions, and guidelines to check against their alt text draft.

Authors perceived value for authoring tasks beyond alt text

Some participants identified how the tool might be useful for broader contexts in their academic writing. P12 remarked that *"[potential user questions] made me think more about the paper, and things that I might want to include in the discussion section or limitations."* P1 commented on the *Generate at Cursor* feature's initial high-level summary that *"I think those short summaries could be really helpful in writing my presentation script, to have something to describe the images on the screen especially [during an] oral presentation. I don't want to go into too much detail or depth."* P10 noted *"I also may use it [for] generating my captions as well, because I noticed my captions are really lackluster."*

9.6.5 Participant-identified limitations

Participants flagged instances of incorrect generations, including figure classification errors, mis-recognized characters in the OCR and these leaking into the description (e.g. "1" for "I"), or mistaking values or value ranges. Though some errors were

identified as *"nothing I couldn't easily correct," "very minor," "not a big issue,"* or needing *"very little effort,"* P1 noted that *"It might be unhelpful if the tool generates false captions for something that isn't in the image and the author doesn't read it over,"* emphasizing the importance of author revision. For example, P8 noted that *"the tool had missed the most important category [in the figure] (in my opinion, since that category was central to the paper's argument)."* Additionally, participants made several usability recommendations that we plan to adopt, such as supporting parallel review of multiple figure metadata attributes to compare to the description.

9.6.6 Log Analysis

We also investigated event traces for insights into how participants engaged with our features. We found that authors expended comparable manual effort between conditions, measured in additions and deletions. Examining individual interaction traces showed how participants employed different strategies in using the features; ranging from reviewing and submitting a pre-generated draft to incrementally building a draft using snippets interleaved with manual writing. This demonstrates that participants who found the tool's features useful may have used them in diverse ways, adapted to their needs and figure context.

Logs Sample Analyzed

Due to database sync issues, event logs from the first five participants were incomplete (specifically, from the base **Draft+Revise** system). As such, results in this section proceed with the remaining 9 participants. Since participants were scheduled by availability, and the group order was alternated between subsequently scheduled participants, we do not expect this to systematically bias our results in any way. As a robustness check, we alternately dropped the first and last participants in this list to create balanced sets of 8 (4 in each group) and the subsequent results did not substantially change.

Aggregate Counts of Events

As a first analysis of participants' logged interactions, we examine aggregates by type (deletions, additions, and text-pasting). Median counts of deletions ("Backspace or Delete"; 60 vs. 56) and additions ("Input"; 450 vs. 399) are comparable across the two system variants (we observe very slightly higher medians and moderately

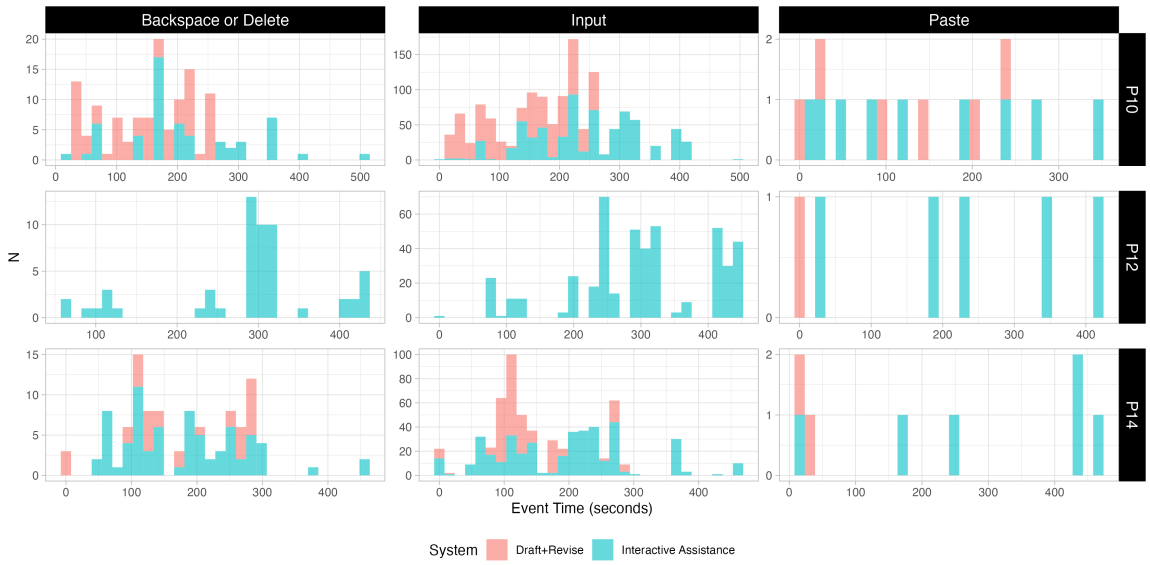

Figure 9-8: Traces of three participants' interaction by event type, highlighting how participants used different strategies to produce final descriptions.

higher dispersion for **Draft+Revise**). The narrower range of key-presses in the **Interactive Assistance** version could indicate that the added features allow a more efficient writing process in some cases. However, the lowest addition and deletion count for **Draft+Revise** is 0, lower than for **Interactive Assistance**, as P12 did not make any edits in this condition, but was satisfied with the pre-generated draft after reviewing for some time. The median count of paste events is higher in **Interactive Assistance** (5 vs. 2); this might account for the added pasting from suggested answers to *Potential User Questions*, in addition to pasting from the pre-generated draft and from elsewhere within the figure metadata or participants' working descriptions.

Event Traces

To obtain a more fine-grained view into participants' interaction and writing strategies, we examined event logs by participant as histograms over time (starting from the first in-session event). Three examples of this are shown in Fig. 9-8, to highlight the differences in how the tools' affordances supported the participants' alt text authoring.

P10, shown in the first row, incrementally built up their description by pasting text from the draft at various points, interleaving this with their own writing. They used the *Potential User Questions* to test *"how it was coming across,"* and found this useful

despite not directly pasting in suggested answers.

P12 only performed one action directly towards producing the description in the **Draft+Revise** condition; they pasted in the pre-generated description, and then reviewed and accepted it. This is apparent in the second row, where the **Interactive Assistance** condition shows substantial presence of deletions, additions, and paste events over the session, but the **Draft+Revise** condition shows only one paste event at the beginning and no other events logged. The screencast recording of P12 shows that they used the open-prompt box, and even tried to obtain a similar effect to the *Potential User Questions* by asking which aspects were unclear from the description (this participant was in group 2 and had already interacted with **Interactive Assistance**). They ultimately decided that, from the resulting questions, *"none of this is helpful."* This process took almost 5 minutes, including time spent reviewing figure metadata.

P14, on the other hand, pasted parts of the generated draft at the beginning in both conditions, but then in **Interactive Assistance** proceeds to paste additional text. Some of this came from the draft (particularly before 300 seconds), and then subsequently from suggested answers to the *Potential User Questions*. Towards the end of this description, P14 pasted two separate answers to the same question into their description in sequence, as they contained complementary details relating to the function of the same visual cues shown in the figure.

In summary, participants used the available features in individual ways reflecting their needs and preferences to craft detailed figure descriptions. Strategies we observed varied even more widely, including patterns like pasting generated drafts and then extensively editing them. The examples illustrate the diversity of strategies employed to balance writing, integrating suggestions, and revising, with support from the system. We include all participants' individual event traces in Appendix A.6.3.

9.7 Discussion

The present work demonstrates how a human-AI collaborative workflow can support authors in making their figures accessible through producing descriptive alt text. Our results show that automatically generated drafts and an interface supporting revision accelerated the authoring process. Additional interactive writing support features, including on-demand text generation (*Generate at Cursor*) and information-seeking

queries (*Potential User Questions*), further helped most authors by progressively building comprehensive descriptions and highlighting points they may have otherwise missed. An analysis of system usage shows authors leveraged these features extensively and in diverse ways depending on their figures and preferences. In the interactive condition, authors produced longer alt text diverging more from the initial drafts, despite similar cognitive load and key-press counts on average. Overall, the system mitigated key challenges authors face in crafting complete figure descriptions. This human-AI collaborative approach highlights the opportunities for combining human contextual knowledge and AI capabilities in making scientific communication more inclusive. However, realizing the full potential of such collaborative authoring systems for accessibility requires addressing issues like generalization and robustness across real-world figures and alt text authoring contexts, integration into diverse author workflows, and responsible deployment, which we explore in this section.

9.7.1 Rise of Multimodal Models

Advances in multimodal language models, which incorporate vision and language, point toward expansive future capabilities for automated alt text generation. Our approach relied on metadata extracted from figures and papers to provide contextual grounding for language models, since today's large language models substantially outpace widely available multimodal models in terms of their generation capabilities and can better incorporate large amounts of metadata representing knowledge about figures. However, the ability to process complex figures directly could reduce dependence on potentially error-prone metadata extraction pipelines while incorporating the right kinds of contextual knowledge for support.

While this could enable purely automated description systems, risks accompany such approaches. Recent work has shown how current state-of-the-art multimodal models can make errors when processing complex figures such as scientific figures [223, 228]. Without human validation, model errors or biases could more easily propagate. Maintaining author discretion may prove wise, even as automated methods become more capable. Furthermore, descriptive tasks require not just visual recognition, but reasoning, inference, and judgment. The wisdom accumulated in authors and fields, who can respond to changing contexts, might allow tailoring descriptions for clarity and relevance. Thus, while future multimodal models may better parse figures, the role of human guidance and customization is unlikely to dissolve. Specialized metadata

extraction models could also enhance such models' zero-shot capabilities. Visual control could, however, be useful; automatically decomposing complex compound figures into components for iterative description is a promising approach we did not explore.

9.7.2 Realizing Gains for Alt Text Consumers

A critical question for future work is whether the increased alt text length and apparent descriptiveness from the system translates to improved comprehension for blind and low vision readers. Evaluating alt text quality remains an open challenge, as illustrated by prior work showing divergent reader preferences. Even assessing descriptiveness in the presence of generated drafts may prove difficult, evidenced by low agreement in our annotation pilot. While we aimed to make it easier and faster for authors to produce detailed alt text, realizing accessibility gains requires considering the perspectives of and impact on readers. Follow-up work on evaluation methodology and studies which evaluate the impact of human-AI collaboratively written descriptions on figure comprehension could help to quantify this impact.

Future work should investigate how to incentivize adoption. Though our study aimed to mimic natural workflows, factors like time constraints, competing demands, and incentive structures also inevitably shape real-world use. Even if the system can help improve alt text completeness, lagging integration risks limiting its impact. Overall, while initial evidence is promising, confirming and extending the benefits requires both rigorous accessibility-focused evaluation and understanding practical barriers to mainstream integration.

9.7.3 Transforming Descriptions to Match Individual Needs

While comprehensive alt text can benefit accessibility, readers have diverse preferences [317] and may desire descriptions of varied lengths tailored to individual needs. Our approach focused on highly descriptive alt texts by design, so that this text can serve as a base to produce personalized derivative texts. As abstractive summarization techniques continue advancing [194, 566, 574], in addition to dialog and other interactive language processing approaches, future systems could apply these methods to accommodate diverse preferences and needs. For example, a concise 1–2 sentence overview could assist quickly grasping key ideas, while retaining the option to query for more information, or expand to more detailed versions for nuanced

understanding. Appropriately customizing alt text poses challenges beyond generic document summarization, requiring preservation of visually salient information such as trends in depicted data. However, customization also holds promise to reconcile the objectives of maximizing completeness for authors while matching diversity in user preferences.

9.7.4 Ethical Considerations

A key ethical consideration is the risk of imposing additional burdens on marginalized communities. Blind and low vision readers already often face exclusion from scientific communication due to the low prevalence of alt text, in addition to other challenges. Providing them erroneous and verbose descriptions without thoughtful human involvement could create further challenges. Though relying on language models is core to the approach in this work, it also risks introducing hallucinations, errors, and biases. Our approach emphasizes author involvement to mitigate these issues, but incentives and workflows must ensure careful review if deployed at scale. The goal should be lightening authors' workload without absolving responsibility. Overall, we must weigh accessibility gains against potential harm, and ensure technical progress on aiding authors in describing figures aligns with the goals of assistive technology.

9.8 Limitations

While we evaluated our system on a diverse and realistic set of figures, the study still involved a limited number of author participants (N=14) describing a small set of their own figures (2 per participant). Evaluating the approach on a larger scale with more figures per author would provide stronger evidence. Relatedly, our participants covered a range of fields, but some areas like life sciences were still underrepresented despite our best recruitment efforts. Testing robustness across even more diverse figures and author backgrounds is an important next step towards deployment.

Additionally, our study instructions asked authors to maximize descriptiveness. A different motivation such as information density (maximizing amount of information conveyed in the shortest amount of text) could change how the system is used and the resulting alt texts. The interface features we designed for the initial goal may not generalize to other aspects of alt text that authors or readers may prefer to optimize for in certain settings.

The automated pipeline also occasionally produced errors (like incorrect figure classification or OCR errors) which propagated to the alt text drafts. Though authors could correct these errors (and pointed out such instances), robustness is critical for real-world utility. We did not systematically characterize authors' ability to resolve errors in final versions of their descriptions, but such an evaluation could also help gauge real-world effects of errors in drafts. Enhancing these components, or integrating uncertainty estimates to guide authors, could improve draft quality and adoption.

Finally, though we demonstrate that our system has the potential to improve alt text writing for scientific figures, the disconnect between assistive writing interfaces such as ours and the scientific publication process limits the true utility of our tool. While authors may be able to produce better alt text using FigurA11y, the processes around integrating this alt text into their publications and making the alt text easily accessible to those who need it are still cumbersome. We acknowledge this limitation and push for better and more intuitive processes around scientific paper accessibility that will make it easier and motivate more authors to include alt text in their publications.

9.9 Conclusion and Future Work

We present FigurA11y, a human-AI collaborative approach to improve the accessibility of scientific figures through descriptive alt text. By combining a pipeline for automatically generated drafts with an interactive authoring interface that makes contextualized suggestions, our system helped authors efficiently craft detailed descriptions of their own figures. Interactive suggestions further assisted authors by highlighting aspects they may have missed describing, enabling iterative refinement of descriptions, and supporting longer descriptions which diverged more from pre-generated drafts without increasing cognitive load or taking more effort on average. Future work can extend this approach by pursuing strategies like incorporating visual information directly, improving robustness of parts of the pipeline, and integrating with real-world author workflows and incentives, to maximize the positive impact on the accessibility of scholarly communication.

10

AI for Musical Discovery

What follows is a position piece on desiderata for future Generative AI-based musical systems, co-authored with Manaswi Mishra and Tod Machover. In it, we draw on literature from music, AI, and several adjacent fields to shape an optimistic perspective on what generative AI could contribute to human creativity, learning, and community in music. Working backwards from this, we consider the kinds of capabilities necessary to achieve the goals we outline, and how these capabilities align with and extend those currently under active development and deployment. Different from all preceding chapters, this one reports no empirical results, but outlines a perspective (enriched by much of the work described beforehand) to inform future contributions in this area.

10.1 Abstract

What role should Generative AI play in music? Long before recent advances, similar questions have been pondered without definitive answers. We argue that the true potential of Generative AI lies in cultivating *musical discovery*, expanding our individual and collective musical horizons. We outline a vision for systems which nurture human creativity, learning, and community. To contend with the richness of music in such contexts, we believe machines will need a kind of *musical common sense* comprising structural, emotional, and sociocultural factors. Such capabilities characterize human intuitive musicality, but go beyond what current techniques or datasets address. We discuss possible models and strategies for developing new discovery-focused musical tools, drawing on past and ongoing work in our research group ranging from the individual to the community scale. We present this chapter as an invitation to collectively explore the exciting frontier of AI for musical discovery.

10.2 Introduction

Music has been a wondrous laboratory for creativity, learning, and community throughout human history. Despite this enduring influence, music's form is anything but static: each era and culture develops distinct musical forms shaped by their values, socio-political contexts, intricate structural logics, and personal narratives. When technology is thoughtfully leveraged, it can profoundly magnify music's reach and widen creative participation. Advances in recording, computing, and networking over the last century have underscored this potential.

More recently, Generative AI has made strides across various sectors, including in music generation. Yet, as these systems advance, we must deliberate on the objectives behind their musical applications. Rather than merely imitating past conventions, how might AI push boundaries and reveal new insights? What novel interfaces could enable more people to develop musical abilities? Broadly, how can we apply these technologies *to enrich human music-making*? With care, we believe AI can inspire and amplify creativity rather than constrain it.

This chapter argues that the primary aim for Generative AI in music should be to nurture human creativity, learning, and community across all skill levels. We propose *musical discovery* as a guiding concept—encompassing not just novel artifacts, but fresh perspectives that deepen understanding and broaden participation. Advancing this vision requires interdisciplinary efforts, from technical innovations to nuanced applications. If developed collaboratively under this humanistic lens, Generative AI has immense potential to inspire new musical ideas that profoundly expand the universal human pursuit of discovery.

10.3 On Human Musical Discovery

A number of psychological theories account for musicians' personal motivation to discover new ideas, beyond those which one assimilates early in musical development. For example, Csikszentmihalyi's concept of *flow* [110, 111] emphasizes the importance of a challenge, which novelty can introduce. Even more diverse are the strategies used to find and pursue novel ideas. Beyond external influences, musicians routinely engage in solo exploration by improvisation, studio craft involving

technological tools, and experimentation in composition[1]. Collaborations between musicians stimulate new ideas by encouraging them to integrate contrasting backgrounds and concepts, leading to surprising combinations and translating existing ideas into new musical domains. Presentational aspects of music, like performing to a crowd, encourage feedback and social dialogue involving a broader array of community members, prompting iteration and integration of new perspectives. Educators also introduce new ideas to musicians, scaffolding their discovery. Ultimately, each discovery, however small, can unlock new expressive modes, enhance technical skills, deepen understanding of musical traditions, and simply invite delight at expanded possibilities.

At an historical level, musicians experiment with fundamental musical elements and concepts. In the Western world, for instance, this can include aspects like form, harmony, rhythm, and instrumentation [76, 450]. Music has also been theorized to evolve alongside broader sociocultural forces [20, 447] and empirically shown to transform through social transmission [12]. The impact of historical-scale musical discovery is multifaceted. It can lead to the creation of entirely new musical styles, enriching the tapestry of human expression. Ultimately, musical discovery expands our understanding of the art form and its potential, while using music as a test-bed for exploring ideas and moving to action [276].

10.4 The State of AI in Music

Early generative approaches focussed on modeling music as a sequence of discrete symbols of musical events, represented as notes or MIDI [106, 212, 379], an approach that has carried into contemporary research [221]. The last decade of advancements in modeling long sequences has enabled us to model music as a sequence of raw audio samples [371]—capturing the detailed nuances of music like timbre, human performance, production, and recording artifacts. Further advancements in higher audio quality, long term structure, and consistency have led to commercial generative AI music services[2] gaining traction. Despite this, a major limitation continues to be the missing agency and control over generated musical outputs.

[1] Jones [238] discusses solo exploration and development through various means. For example, "Avoid Paralysis From Analysis" details how Jones overcomes inertia, such as by engaging in musical practice with abandon.

[2] Examples include *AIVA*, *Infinite Album*, *Endel*, and *Boomy*.

Fueled by an aggregation of large-scale datasets for music [1, 43, 156, 179, 559], we are now faced with a myriad of *foundation models* [57] for music [3, 107]. Such models show an impressive ability for pastiche but are highly restrictive in their limited musical diversity, textual conditioning, poor extrapolations, and missing provenance. The desire for interpretable and controllable models can be supported by a number of research developments [71, 137, 557], and bespoke generative AI experiments by musicians [117, 546]. In summary, AI music generation is both historically rich and rapidly evolving, with impressive progress in symbolic and raw audio generation, foundation models, and interpretable approaches. However, limitations remain in agency, control, diversity, and provenance. Addressing these limitations will be crucial for unlocking the full potential of AI in music.

10.5 Developing Musical "Common Sense" and Long-Term AI Progress

Despite impressive pattern recognition and generation, modern AI systems still lack the "common sense" understanding of the world that comes naturally to humans. This is evident across domains like language, vision, robotics, and music. In AI research, "common sense" refers to the ability to reason intuitively about everyday situations depending upon implicit knowledge about how the world works [97, 115], including aspects like intuitive physics and psychology [277].

In music, we argue this involves recognizing and manipulating the intricate structures, semantics, and aesthetics that form the fabric of musical expression. Such musical intuition is difficult to capture through explicit datasets or training objectives. Rather, many aspects of music emerge through implicit learning processes [433]. We call it "common sense" because it reflects shared assumptions and sensibilities within real-world musical expression, acquired through a complex interplay of biological, psychological, and cultural processes. Developing this level of comprehension remains a grand challenge for AI in music, and indeed music has been argued to provide deep challenges for AI development more broadly [432].

Much has been written about human *musicality* [216, 512], a complex and even contested [65] notion that can be seen as addressing implicit musical abilities, similar to what we call musical common sense. In humans, the notion of musicality is entangled with questions of talent vs. skill, ability and aptitude, cultural universality,

and species-specificity. We aim to distinguish our notion of musical common sense from musicality. One reason for this is that in order to encourage technical progress in research, we must align sufficiently on the capabilities we hope to develop. The other is that we do not aim to replicate all facets of human musical intelligence in AI systems. Rather, we hope to cultivate the aspects of musical comprehension that allow AI to most effectively enhance human creativity, through musical discovery.

We propose the following categories of capabilities as *layers* of musical common sense; parts that, though inevitably incomplete in capturing all of human musical behavior, enable developing clearer goalposts for future progress.

Structural Attributes This involves recognizing and manipulating the fundamental patterns, idioms, and theoretical constructs that form the building blocks of music in different stylistic, cultural, and social contexts. Expert musicians fluently apply such conceptual understanding when communicating ideas and intentions. Structural knowledge also aids music educators in conveying concepts to students, both to transmit knowledge of the past and to offer building blocks that students can use to generalize and extend past ideas.

For example, in the context of Jazz, this could include chords and extensions, harmonic substitutions, canonical rhythms, higher-level notions like progressions (e.g. "Rhythm Changes"), sections, and standards. This conceptual understanding could allow an AI system to support a jazz musician in various ways. During practice, the system could generate variations and reharmonizations on chord changes to standard tunes to help expand the musician's harmonic vocabulary. In live performance, it could listen and respond with expected or challenging accompaniment. For analysis, the system could identify key patterns and structures in improvised solos to elucidate techniques. For Jazz composers, common forms often rely on knowledge of both repertoire and harmonic concepts, such as contrafacts and reharmonizations.

Imagine, for instance, a musically knowledgeable multimodal foundation model. A novice might query such a model with a vague but intuitive textual description or auditory example of a musical idea, and the model would respond by identifying relevant theoretical constructs, retrieving examples from the literature and synthesizing new ones, and offering application ideas in order to help the learner build a meaningful mental model of the underlying concept. For an intermediate student, the system could generate reharmonizations and stylistic variations on a standard.

This provides material to practice improvisation in novel and diverse contexts. For an expert performer, the system allows specifying ideas in precise musical language and iterating to rapidly explore new ideas. For example, a saxophonist could explore reharmonization concepts for a standard under different ensemble configurations by generating examples building on their intuition. A meaningful exchange in such a scenario is predicated on shared structural comprehension; the model must be able to represent and manipulate the expert's ideas accurately and fluidly. In each case, this layer of musical common sense allows the model to build on what the musician knows and can convey to eventually reach new territory for educational or creative goals. As musicologist Paul Berliner writes, "There is... a lifetime of preparation and knowledge behind every idea that an improviser performs." [42]

It is essential, however, to maintain humility about the fluidity and subjectivity of such notions. Musical knowledge resists over-codification, as conventions evolve dynamically across cultures and eras based on myriad factors, and are personalized based on individual experience, references, and context. What is considered standard in one generation may be cast aside in the next, and structural models of musical information often only crystallize in retrospect (e.g. through significant musicological efforts). For instance, the well-known sonata form exhibits considerable heterogeneity [436]. We must also acknowledge the inherent limitations in formally encoding creative human practices like music.

As such, we should be wary of over-reliance on explicit idioms in building and evaluating generative AI for musical discovery, and instead seek to perceive and participate in music's ever-changing landscape with openness and nuance. The priority should be conveying possibilities in the musician's own terms, not imposing assumptions. Consider the "Beginner's Mind" or *shoshin*, an idea with its roots in Zen Buddhism. Suzuki writes "In the beginner's mind there are many possibilities; in the expert's mind there are few." [494]

Emotional Context Disparate theories account for how emotion is expressed through, perceived in, and induced by music [241]. While academic discourse on the topic continues, musicians effortlessly intuit music-emotion relations. Composers and songwriters learn associations between musical devices and emotional states within their style and culture, often without explicitly reasoning about these associations. Performers even make subtle adjustments to phrasing, articulation, and expression

to evoke varied affective responses.

Machines are usually taught to connect music and emotion through explicit tasks like music emotion recognition (MER) [227], or implicitly through aligning music with affectively valent textual [133] or visual [492] correlates. However, these methods are unlikely to fully capture the nuance involved in musical emotion. MER often depends on datasets and prediction targets derived from simplistic taxonomies [201]. Implicit learning from textual associations may instill biases from datasets, due to limitations in how emotion is often discussed and the need for more information than language alone for representing rich emotional concepts [390]. Such techniques lack grounding in the human experiences, embodiment, and enculturation that gives rise to musical emotional fluency, and may implicitly encode biases in music-emotion connections.

Progress could require models that learn holistic musical emotion understanding through real-world immersion, or simulations and other experiential strategies. As one application example, imagine a system assisting a film composer in exploring ideas for a score. Generating plausible and compelling ideas requires an implicit understanding of the precise emotional arcs, aligned to scenes. A system could suggest musical ideas with knowledge of this context, and even ideas that thoughtfully deviate from it (for instance, to foreshadow future events in earlier scenes [55, 497, 531]), supported by a rich model of emotional context. This capacity for emotional insight is key to AI that can meaningfully collaborate in human musical communication, and stretch creators in new affective directions.

Interaction Dynamics Human musicians communicate through an unspoken language of musical cues [249, 250]. In classical ensembles, quiet signaling enables almost inhuman feats of coordination and results in the synergistic performances we are accustomed to as audiences, from ad-hoc duos to conducted orchestras. In jazz, players cue solos, accompaniment, and transitions seamlessly, displaying complex decision-making in real-time. Sensitivity to such social signals facilitates participation in music.

However, current AI systems lack awareness of such nuanced musical interaction. As Browning and LeCun note, “social customs and rituals can convey all kinds of skills to the next generation through imitation.” [62] To collaborate meaningfully with creators, AI must appreciate the social dynamics of music.

Progress in this area may require interactive environments where systems learn subtleties experientially. Evaluation metrics should assess musical-social intelligence beyond technical ability. Musical collaboration relies on tacit knowledge, and so such social competence is critical for AI that aims to enhance creativity through interaction on human terms, rather than replace it through automation.

Adaptivity and Personalized Behavior In prolonged musical interactions, AI assistants must learn to adapt contributions to complement individual creators. User-adaptivity is a classic goal in computing systems [67]. In language modeling, this goal has been bridged with modern foundation models by leveraging techniques like in-context learning and prompt engineering. For instance, OpenAI's ChatGPT interface allows setting "Custom Instructions"[3] that allow long-term consistency, and users may prompt within sessions to bias behavior towards personal desires as they change over time or respond to exogenous factors.

However, as Glassman recently described, human-AI interaction involves complex loops of intent formation, expression, inference, action, verification, and updating [186]; in light of this, adaptation to users and goals from simple strategies like prompts may not be straightforward. Additionally, Picard proposed years ago that learning user subjectivity requires establishing shared "common sense" specific to the user, and then observation and learning over time [389].

For music, we propose that personalization involves technical capabilities like recognizing preferred rhythms, motifs, emotional tones, and other artistic factors, but also entails detecting strengths, weaknesses, tendencies, and gaps. The goal over time is creative growth through personalized scaffolding; whether expanding the user's skillset, their output, or simply keeping track of their musicianship as it changes.

This requires architectures that accumulate rich user models, responsive to both immediate and longitudinal patterns in individual creative expression; akin to what Bickmore and Picard [46] once described as *relational agents* in a more general setting. In this way, AI systems can complement, challenge, and empower human creators while retaining their unique voices.

Cultural Sensitivity Music poses a profound challenge for cultural understanding in AI. Musical conventions and aesthetics vary dramatically across the world's cultures,

[3] openai.com/index/custom-instructions-for-chatgpt

which each carry unique symbolic meaning and social significance. Riedl discusses the goal of *machine enculturation* [421], describing this as "the teaching of sociocultural values to machines," and proposes a way to accomplish this: through stories, which often implicitly encode values and tacit sociocultural knowledge. Finding strategies that similarly implicitly convey such values in the context of music, and supplement narratives like stories, could be helpful towards this goal.

Another important aspect is training data. Training data often encodes implicit biases [56, 66], so achieving culturally sensitive AI requires intentional efforts to improve representation; for instance, implementing ethical data sourcing and sampling strategies, involving community members for evaluation and feedback, and using technical measures to reduce imbalances where possible. Ultimately, datasets are insufficient without participation from people to instill nuanced comprehension. We expect progress to come through partnerships with cultural communities, where human guidance and validation steers systems away from bias and towards genuine sensitivity.

Moreover, granting cultural communities authority over their musical representation is imperative to avoid misinterpretation and appropriation by AI systems. By incorporating these strategies, AI can progress towards genuine cultural sensitivity, understanding cultures as complex, evolving entities rather than static sets of traits.

10.6 Extrapolating Beyond Today's Sounds

While generative models have achieved impressive results emulating existing styles, moving beyond today's musical horizons presents acute challenges. Definitionally, today's models are trained on yesterday's data; this is what makes them so fluent at recreating the past. Yet relying on imitation alone risks stagnation. How then can we grow new sounds?

10.6.1 Embracing Uncertainty

Modern generative AI models have extraordinary imitative abilities, yet often err in intriguing ways. This unpredictability has parallels to the long tradition of artists finding inspiration in chance processes, such as in the aleatoric music of John Cage [52]. However, the uncertainty of large language models is not arbitrary randomness, but rather can be seen as unexpected interpolations and combinations within their

domain of imitation: explorations of the latent space learned from their training data. Recent work has shown how diffusion models can encode aspects of human musical expectation and surprisal [332]. This suggests that model errors and uncertainties have aesthetic potential if creators can scaffold and direct them in a meaningful way. However, they are currently serendipitous side effects of imitative processes, rather than being scaffolded with meaningful interactions. One possible path forward is to leverage abstract reasoning processes, such as those at play in large language models, to systematically recognize and leverage model errors and scaffold how human creators tap into them as resources for discovery and creativity. In doing so, we propose that Generative AI provides an opportunity to advance the artistic legacy of revealing creativity hidden within unpredictability.

10.6.2 Transformational Creativity

Boden famously proposed three forms of creativity: combinatorial (or combinational), exploratory, and transformational [50]. Combinatorial creativity involves novel syntheses of familiar ideas. Exploratory refers to generating novel ideas within an established conceptual space. Transformational creativity, however, fundamentally reshapes a domain's possibilities. Though this is an ambiguous notion, Boden cites Schoenberg's ideas about atonality as a musical example.

Transformational creativity is rare and revolutionary—it is the long tail of creative acts. Even so, it is vital for the future of music; this is how we catalyze periodic upheavals of musical thinking and yield influential new movements, while in turn using music as a catalyst to inspire hope and optimism that positive pathways and solutions to any situation—no matter how intractable—can always be found. With generative systems becoming increasingly capable at combinatorial and exploratory tasks, there are opportunities to also support this most ambitious form of human creativity[4].

Presently, it is hard to see a path to models independently achieving this transformational type of creativity. However, a promising way forward is human-AI collaboration. In this context, our machines need not recast music independently but instead amplify human creativity into unfamiliar and radical new domains. We hope for systems that can scale up cycles of co-creation and feedback to accelerate refinement of

[4]Amabile [8] presents social and motivational factors which influence creativity, which provides one possible basis for Generative AI-based interventions that do the same.

transformational ideas, as well as aggregate cross-disciplinary knowledge to make unconventional connections across domains. Evaluations of this should prioritize long-term contribution, wherein AI tools enhance imagination to help sustain music's endless evolution.

Thoughtfully integrating AI has potential to accelerate human musical discovery in multiple ways. At times, embracing uncertainty can spark novel ideas within established conceptual spaces, and help us rapidly explore them. Unexpected permutations can reveal overlooked possibilities, encouraging us to take a closer look. Periodically, transformational leaps enable us to explore uncharted territory. All of these forms of discovery are vital for music: the former two nourish thriving ecosystems, while the latter propels enduring reinvention and growth. With human creativity amplified but not displaced by machine collaboration, music can evolve without losing touch with human experience. AI can assist discovery, but music's capacity to speak across eras originates in our shared experiences.

10.7 Developing New Tools for Human Creativity and Discovery

While past creative tools provide useful starting points and evocative models for promoting musical discovery, fully realizing Generative AI's transformative potential also requires new perspectives. Rather than simply replicating long-standing assumptions and interfaces, we must rethink human-machine interaction to prevent established biases from implicitly constraining the potential of future systems. Here, grounded in past and present research in our group, we outline our vision for future generative tools that encourage musical discovery across a set of exciting applications.

10.7.1 Augmented Ideation

Musical ideation is profoundly shaped by context, from lone composition to ensemble improvisation. These environments present distinct opportunities for AI augmentation while posing challenges requiring thoughtful sensitivity. For example, composers ideate through cycles of exploration and refinement, necessitating adaptive systems that toggle between divergent idea generation and focused iteration. Meanwhile, improvisers often ideate fluidly from real-time stimuli, implying tools for rapid variation and response.

Our group has long cultivated systems to enhance musical ideation, for instance using computational methods to interface with large audio datasets [468]. However, modern Generative AI offers a fundamentally distinct design material for fueling creativity through its capacity to synthesize novel outputs that derive from existing musical datasets, for instance with text-based semantic guidance. Recently, we developed a sound generation method that introduces semantic guidance to the modular synthesizer paradigm, a historic set of tools that has fueled musical ideation for decades. This method allows users to generate sounds from prompts, but then adjust these sounds and freely explore using a small set of interpretable knobs [89, 91], in contrast to black-box sound generation methods.

Designing interactive systems also allows nuanced and reciprocal influence. Our group has built an AI ideation system that taps into individual users' voices to brainstorm and create compositional material [347]. For a recent concert, we developed and deployed a real-time Generative AI system based on a set of RAVE [71] models. This system translated and varied performer gestures into provocative new timbres, provoking them to form a dialogue with altered versions of their own ideas. This real-time call-and-response resulted in stimulating duets that neither party could have produced alone.

10.7.2 Augmented Presentation

Historically, music has been commodified and marketed as static, definitive products—fixed recordings and compositions intended for passive consumption. However, our group has previously proposed a more flexible paradigm for musical experiences centered around fluid musical "sound worlds" that users can manipulate and extend indefinitely [214, 322]. *Artificial.fm* is another proof-of-concept system which demonstrates an "AI Radio", allowing collaborative steering of AI-generated music outputs using participatory curation [193].

Generative AI could prove integral to realizing this vision of mutable musical ecosystems that break from traditional attachments to predefined songs and recordings. However, this is a non-trivial extension of current paradigms for music generation: composers must retain the ability to endow generative models of their music with certain invariant qualities that establish their aesthetic values. Even so, generative techniques offer promising means to manifest adaptable, personalized sonic experiences that transcend static compositions. Beyond encouraging re-discovery of existing

music, we have also held an interest in how adaptive music can be used for affect improvement [286, 487]. Broadly, we are interested in harnessing these methods to give people agency in the music that they share and experience, as well as to introduce surprise and delight in hearing well-liked music that reveals new secrets at each listening.

10.7.3 Creative and Adaptive Learning

Influential pedagogical theories like constructionism highlight learning through creating meaningful artifacts, often facilitated by technology [381]. Vygotsky's zone of proximal development model [532] describes a scaffolded learning process, where guidance nudges students just beyond current competencies. These frameworks underscore the potential of Generative AI to contribute to transforming learning beyond passive transmission and towards creative invention. Learners can translate conceptual ideas into personally relevant works to internalize new knowledge.

Prior systems like *Hyperscore* [147], developed in our lab, exemplify this creative learning by enabling students to draft motifs and develop compositions with coarse-grained sketching behavior and intelligent harmonic controls. It is essential that future tools—such as the extended Hyperscore environment that our group is currently designing for the new Johnson Education Center at the Dallas Symphony Orchestra—similarly maintain learner agency and engagement to maximize their learning and growth. When preserving this, the immense power of generative techniques to actualize ideas can profoundly enrich learning across skill levels. Students stand to gain deep understanding and identity-forming creative skills as they steer personalized journeys and shape multifarious variations grown from their own seed ideas.

10.7.4 Scaling Participation and Collaboration

Beyond empowering individuals, generative AI could also transform music's social fabric by facilitating creativity within and amongst communities. Recent endeavors like our group's *City Symphonies* invite residents to contribute to musical portraits of urban areas through diverse submissions aggregated into grand-scale experiences. We have developed a range of technologies to support community input into collaborative works [321, 526, 527], but generative techniques present new opportunities for such designs; they could enable community members to contribute and combine a

wide variety of expressive ideas, with even greater facility and power than previous tools provided. To nurture communal creativity, systems must maintain individual voices, and enable both personal exploration and constructive dialogue between different contributions. We seek thoughtfully designed technologies that help foster deep belonging and equitable exchange between community members in creative collaboration, and are currently developing such tools for the *Wellbeing of the World: A Global Symphony* project, scheduled to premiere in 2025.

10.7.5 Identifying Limits

While Generative AI promises rich possibilities for musical discovery, we must also identify boundaries: certain profoundly human qualities and lived experiences of music may remain beyond capture. This is, of course, true even more broadly than music: we must probe the conceptual limits of new technologies, meaningfully speculate on their potential consequences, and consider what we need to preserve when bringing automation into human endeavors. Our group explores these tensions through the rich medium of Opera, which brings together artistic and technological means to imagine and interrogate such issues [2, 235]. Opera can help us tell important, humanistic stories that provoke and ground conversations about future technologies. We explore AI's cultural tensions through operas integrating stories and real systems. These operas enact both dreams and limitations in order to crystallize priorities at the heart of our research—catalyzing creative discovery through machines built first and foremost for expanding human potential rather than simply accelerating industrial progress.

10.8 Conclusion

The discovery of new musical ideas, for individuals and across communities, has long progressed through an intricate exchange between human creativity and technological innovation. Generative AI now stands to carry this legacy forward—but truly nurturing musical creativity relies on developing transformative new systems guided by this synergistic interaction. Our goal in this chapter has been to propose *musical discovery* as an orienting principle, outline key *musical common sense* capabilities—structural, affective, and sociocultural—that are vital to meaningfully enable this, and showcase possible models for the design of *new tools* for musical discovery. Despite the impressive accomplishments of present-day musical generative models,

we believe the path ahead is rich with challenges that will necessitate insightful solutions both technical and artistic, broad collaboration, and lively community dialogue. Our task is now to formalize these challenges, propose and manifest solutions, and collectively progress systems while ensuring that they expand, rather than constrain, the horizons of human musical endeavor.

Part IV

Putting it All Together

11

Discussion

The work in this thesis has explored how intelligent systems can meaningfully extend human capabilities, particularly in or motivated by creative work, through the lens of several application domains and methodological approaches. Through engaging in this series of investigations spanning computational modeling, system building, and empirical studies of human-AI interaction, I have found myself reflecting on themes that seem to recur and resonate across it. I have endeavored to describe some of these themes below.

Working with Imperfect but Useful Approximations

One recurring finding across multiple projects is that imperfect or approximate intermediate representations can serve as powerful bridges between human and machine capabilities. Chapter 2 works with images and estimated depth maps. While these are clearly crude approximations of true scene geometry, they provide sufficient structure to enable meaningful acoustic simulation at the level needed for many creative tasks. Chapter 5 dealt with a fairly specific synthesizer architecture, certainly not suited to the diversity of sounds one might think to generate. Even so, this limited representational capacity lends itself to abstraction: a property that might often be desired in sound design contexts. In Chapter 8, suggestions that were technically "incorrect" or "irrelevant" still proved valuable to many writers by sparking new ideas by prompting reinterpretation and *integrative leaps*.

One possible interpretation of this is that the primary goal of human-AI systems can extend beyond perfect fidelity or accuracy, and embrace representations that are "usefully imperfect". These maintain enough structured information (e.g. visual or textual coherence) to enable meaningful augmentation, while expanding the room for human interpretation and agency.

Indirection of User Intent as a Design Strategy

Some approaches used in this thesis involve introducing indirection between user input and system output, i.e. they output a different representation than what the user might intuitively expect, but in doing so create additional opportunities for creative follow-ups from users. The synthesizer-based approach to text-to-audio generation in Chapter 5 achieves creative sound design through producing results in an interpretable parameter space rather than direct waveform generation, which would be perhaps the most direct way to get auditory results. Without this indirection, however, the possibility space for what the user can do with the output is significantly more limited. The multimodal writing system in Chapter 8 provides suggestions through parallel channels of text, image, and sound rather than just direct text completion, and provides multiple suggestions. If it functioned as an auto-complete, it may not have been able to support (or uncover, to begin with) the diversity of needs and strategies arising in the user writing tasks.

Such indirection can be useful at a design level: it creates spaces for human interpretation and creativity while working within current technical limitations. It suggests that rather than always pursuing end-to-end pipelines, thoughtfully designed detours may often better serve human-AI collaboration.

Value of Training with "Unrealistic" Distributions

A more speculative pattern we observe across multiple technical contributions in this thesis that training on data distributions with carefully controlled divergence from target (downstream task) distributions can yield benefits. The audio doppelgänger work in Chapter 6 demonstrated that training on synthetic sounds—which systematically differed from real audio distributions—could still produce robust and useful representations. Similarly, the dubbed movie approach in Chapter 3 showed benefits from controlled variation in the audiovisual relationship; the visuals remained constant while the audio diverged in predictable ways that are nonetheless not observed in downstream datasets. Both works also deal with a counterfactual-like augmentation strategy: in the former, we produce random sound pairs that vary in terms of their underlying parameters. In the latter, we consume dubbed videos, which encode a "what if" scenario in terms of speech differences. Both try to approximate phenomena that we often don't directly observe in the world, but conceptually occur

in the form of counterfactuals, and were studied through controlled variation. When constructing training data and procedures, this work suggests it may be prudent to consider exploiting strategies that don't directly resemble existing datasets, but reflect assumptions that meaningfully structure learning.

Cognitive Integration Work as a Feature

The empirical study in Chapter 8 reveals that users might perform substantial cognitive work to bridge between system capabilities and their goals. This integration work may be useful for maintaining agency and ensuring high-quality outputs that align with the user and contextual needs, rather than being a limitation to be circumvented by better automation. In particular, the writing study showed how this cognitive effort appeared to facilitate novel story directions and greater perceived ownership. Similarly, in Chapter 9, authors' revision work was important for accuracy and completeness, informed by their contextual awareness of their own research (from which the figures were taken), and the field in which they operate. In a way, Chapter 5 hints at this notion as well: the abstraction inherent in the synthetic results may require a little cognitive work to map to desired sounds (consider listeners' greater uncertainty, despite their reasonable accuracy in recognizing the categories). However, this effort appears valuable for sound design outcomes (users also perceived the sounds as being more artistically interpretive of the concepts). This suggests that system designs should actively support and scaffold this integration work, in conjunction with developing better capabilities.

The Art and Science of Parameter Space Design

One central challenge in building effective human-AI systems, highlighted by the work presented thus far in this thesis, lies in the design of parameter spaces. These parameter spaces often constitute both controls for model behavior, and specifications for human input. This thesis demonstrates this across multiple contexts: sound synthesis parameters enabling creative audio generation or perceptual sound understanding, visual features guiding acoustic modeling, and design space parameters structuring experimental exploration (to come in the chapter to follow).

A careful look at these cases reveals several principles that appear important for effective parameter space design:

1. Parameter spaces must serve dual objectives. First, they must support efficient computational manipulation and learning. Second, they should enable meaningful human interpretation and control. Without the former, machines may struggle to operate them sufficiently for machine intervention to be useful. Without the latter, the parameter spaces become more useful for automation than augmentation Chapter 5 shows this explicitly: synthesis parameters simultaneously support optimization for matching text descriptions while providing interpretable controls that sound designers can manipulate. Chapter 6 goes beyond this, showing how the same parameter space can be used to support efficient and effective audio representation learning. The success of this approach suggests that finding such dual-purpose parameters may be tractable, and a useful alternative to building separate human/machine representations.

2. Parameter design can follow multiple paths. In particular:

   (a) Mining historical domain knowledge: The synthesizer work leverages decades of sound design expertise encoded in synthesis parameters. This is useful because: (1) these parameters are already validated through extensive human use, (2) they capture meaningful dimensions of variation, and (3) they have clear relationships to perceptible outcomes.

   (b) Theoretical derivation: Another strategy is to construct parameters through systematic analysis of design spaces. This requires explicitly enumerating design dimensions, grounded in a cogent theoretical framework. In the following chapter (Chapter 12), we will present an example of this.

   (c) Empirical derivation: Here, traditional designer strategies such as need finding may come into play. Grounding a parameter space in an empirical investigation (e.g. a need-finding study with a target population) could be a useful way to discover meaningful factors of variation, and thus meaningful parameters for this population to use.

   (d) Data-driven discovery: Future work might also consider discovering possible parameters from data or from models trained on data, for example using techniques from interpretability wherein we probe abstract computational representations to make meaning of them. This direction has not yet been explored in this thesis, but model steering has already proven useful in the language modeling context, as a precedent.

3. Effectively evaluating parameter spaces may require applying multiple, complementary criteria:

   (a) The *technical utility* of the parameters, i.e. how they support effective learning and/or optimization.

   (b) Their *human interpretability*, i.e. how well they facilitate conceptual understanding and purposeful manipulation by users.

   (c) The *creative affordances* revealed by them. Effective manipulation is a necessary, but not sufficient, condition; manipulating these parameters must also yield some reward for humans.

   (d) Their *contextual robustness*; the degree to which they maintain their utility across different contexts of use.

   In a way, the synthesizer parameters in Chapters 5 and 6, backed by the efficient implementation in Chapter 4, shows technical utility for optimization and learning, human interpretability and creative affordances by virtue of our results and their historical roots, and contextual robustness by being useful for both human-centered sound generation and audio representation learning. Still, developing strong evaluative frameworks for parameter spaces may be the best way to ensure that we continue to develop better configurations.

Of course, parameter spaces may never be *complete*. As Jaron Lanier notes in *You Are Not a Gadget* [283], "[the] definition of a digital object is based on assumptions of what aspects of it will turn out to be important." When we deal with essentially truncated representations of complex real-world phenomena, we may have to, as Lanier notes, treat them with "special caution." Still, even simplified representations can grow into more wholistic conceptual models. Lanier gives the example of MIDI as a highly simplified representation of music. Though this is true, MIDI has recently evolved to accommodate many more nuances through MIDI Polyphonic Expression (MPE) and MIDI 2.0, getting past the conceptual "lock-in" of earlier versions.

The Knowledge Integration Problem

Consider a designer tasked with building an AI system, capitalizing on a recent technical advance, to support architects in exploring new building designs. They

might look to prior work, finding papers about systems that help writers craft stories, musicians compose pieces, or visual artists generate concepts. Each describes valuable lessons about how users responded to different forms of algorithmic intervention. Still, how should they translate these insights to architecture? Which aspects of the writing assistant's suggestion timing would carry over? Would the strategies that helped musicians maintain creative agency work similarly for architects? Further, what should they do when suggestions conflict between the different domains? Since these works may not have been done with explicit knowledge of each other, the designer's attempts to reconcile their findings may be the first.

In this discussion so far, we have engaged in intuitive and analogical reasoning to build bridges across this line of work. To me, this has required years of immersion in these research areas, and the broader domain they are sampled from. In the field more broadly, such integration may occur through dialog between researchers working on different subproblems. Yet, the question of how the subproblem-level insights should be systematically integrated to form cogent theories with high predictive utility is not clear.

Indeed, the challenge runs a little deeper than just difficulty translating findings, especially when we consider the effort involved in the design and implementation of an AI-based interactive system. Without systematic ways to vary and study design and interaction strategies across contexts, we risk rebuilding similar systems repeatedly while missing opportunities to understand what about them truly generalizes. Each new project starts, in part, from scratch, implicitly rediscovering principles that might have been more efficiently established through careful experimental design.

This ad-hoc approach to knowledge building becomes increasingly untenable as the underlying technical capabilities advance rapidly. The design space grows ever larger, and making progress one prototype at a time leaves too much territory unexplored. Arguably, it leads to knowledge that is even more "provisional, contingent, and aspirational" [174] when working with generative models than with the less expressive computational substrates engaged in traditional research-through-design [582]. Motivated by this observation, I introduce a new conceptual framework for studying human-generative AI interaction in the next chapter. Though this discussion has sought to hypothesize about some common principles, I argue realizing their value requires new tools for studying them rigorously.

12

Meta-Prototypes: Towards Integrating Design and Experimentation in Human-Generative AI Interaction

Abstract

Traditional machine learning models such as classifiers lend themselves to systematic study through simulation, enabling precise manipulation of model behavior and integration with controlled experimental designs. Simulation is, however, infeasible for high-dimensional, interactive generative models, which demand in-situ analysis due to their dynamic and context-dependent behavior. As a result, design knowledge for generative AI-powered systems advances through ad hoc, prototype-by-prototype iterations, sampled implicitly from a broader, often unstructured, design space. Then, to aggregate design knowledge, we rely on scientific publication and dialog. This chapter proposes the concept of *meta-prototypes*: parametrically defined families of interactive systems that enable systematic exploration of design spaces. Extending from the paradigm of integrative experiments, we propose methods for parameterizing, instantiating, and experimentally testing such meta-prototypes. We discuss how this approach can shift us from fragmented design knowledge to more robust and predictive theories of human-generative AI interaction.

12.1 Introduction

Consider two scenarios. First, a materials scientist is trying to develop a new polymer with high tensile strength and biodegradability. They have a large database of existing polymers and their properties, but struggle to identify promising combinations for

experimentation. Second, a UX designer is redesigning a complex dashboard for a financial application. They need to balance information density with ease of use, while adhering to accessibility and usability guidelines and a company's design system.

Despite their differences, both scenarios share some common structure. They include a large search space of possibilities (polymer combinations, design layouts), multiple—possibly competing—constraints (material properties, usability factors), the need to generate and evaluate numerous candidate solutions quickly, and the potential for unexpected connections or solutions to arise. Both could benefit from algorithmic assistance, for example by summarizing key data points, suggesting novel connections, or generating alternative solutions; things modern Generative AI tools are used for.

Currently, we approach these as separate problems. We might prototype and develop a system for one or the other task, and learn something useful about how best to support it in the process. However, this ignores their shared structure, resulting in design knowledge with limited generalizability and perhaps even a limited shelf life; what happens when the technical substrate, e.g. the generative model in use, changes dramatically? Some circumstances are even more similar: imagine designing visual content vs. text for an ad campaign. They differ in modality, which suggests the use of different computational tools and thus different interactive systems. This is akin to studying caffeine's effects on writers and strategists in isolation, missing the underlying cognitive mechanisms that span domains and allow us to build stronger, more predictive theories. Consider questions like: *When should algorithms intervene in human creative processes? How will algorithmic interventions influence outcomes? Will such systems expand a creator's conceptual space, or narrow it towards more predictable outcomes? What are the long-term impacts on users? What differences are there between well-defined vs. open-ended tasks?*

12.2 The Knowledge Integration Problem

The core issue here is our inability to systematically vary and study algorithmic intervention strategies across different creative contexts. This limitation propagates through the entire field of human-generative AI interaction, and can result in redundant efforts across prototypes and missed opportunities for cross-pollination of ideas. Indeed, designing AI systems for creative work has historically been an ad-hoc endeavor, leading to fragmented insights and a lack of integrative theories. Human-AI

interaction in creative contexts often proceeds one prototype at a time, similar to how social sciences have been observed to progress "one experiment at a time" [6, 359].

The task of designing and building AI systems for creative work, a common use case for interactive generative models, is particularly under-specified due to the absence of robust consensus in its scientific study, which is marked by a diversity of definitions and theories [37, 130, 331, 397, 399]. This theoretical fragmentation complicates the design of AI systems for creative work, since there is not a common and canonical foundation for it and human-AI *co-creativity* introduces additional considerations. Recent advances in AI, particularly in generative models, have created new possibilities for human-AI creative partnerships, but our understanding of how to effectively integrate these technologies into human creative processes remains limited [404]. The field of Creativity Support Tools (CSTs) has made progress in understanding how technology can augment human creativity [464]. However, the unique challenges posed by modern generative AI tools require a new investigative framework, due to the systems' ability to generate and modify creative artifacts directly.

To address this limitation, we introduce the notion of *meta-prototypes*: flexible, modular human-AI interactive systems designed to systematically explore a broad design space. These meta-prototypes will enable rigorous investigation of the cognitive and computational factors that facilitate productive human-AI creative partnerships. By systematically varying intervention strategies and contexts, we can study factors like when and how AI should intervene to expand a creator's conceptual horizons versus consolidate their ideas, producing design knowledge for next-generation creativity support tools.

12.3 Meta-Prototypes

12.3.1 Some Definitions

First, it's worth formalizing what we mean by *prototype* in this context. Houde and Hill [217] regard prototypes as "any representation of a design idea." Here, we define a more restrictive form of prototype. As noted earlier, interactive systems based on high-dimensional interactive generative models, such as language models and text-to-image generators, are difficult to simulate in lower-fidelity settings. Even simple designs may need to actually use such technologies, if the interaction is of

interest beyond outputs alone. We consider the case of such prototypes, which have a few key features we formalize below.

Definition 1 (Prototype): A *prototype* is a tuple $P = (I, A, C, \Lambda)$ where:

- $I \in \mathcal{I}$ represents the *interaction model*, i.e. the features of the interactive system such as modalities, mixed-initiative components, etc. which are relevant to the task. These are sampled from a broader space of all possible interaction features, i.e. $\mathcal{I}$.
- $A \in \mathcal{A}$ represents the *algorithmic model*, including which generative AI model(s) and weights are used, what their inference parameters are (e.g. decoder temperature), and other computational aspects. Likewise, $\mathcal{A}$ denotes a broader space of possible algorithmic aspects.
- $C \in \mathcal{C}$ represents the *context*, which includes task-specific or environmental assumptions relevant to the system's operation, such as the type of artifact a user is authoring, the social or institutional context the artifact is embedded in, or the duration of the interaction. $\mathcal{C}$ here is the set of all possible contexts, not just those relevant in this one prototype's case.
- $\Lambda \in \mathcal{L}$ represents *nuisance parameters*, including extraneous design features or confounders (e.g., banner color, and other non-essential interface features) drawn from the broader possible space of such features $\mathcal{L}$.

The set, or "universe" [6], of all possible prototypes is denoted by $\mathcal{P} = \mathcal{I} \times \mathcal{A} \times \mathcal{C} \times \mathcal{L}$.

Definition 2 (Meta-Prototype): Then, a *meta-prototype* is a parametric family of prototypes, defined as a tuple $M = (I_{\Theta_I}, A_{\Theta_A}, C_{\Theta_C}, \Lambda)$ where:

- $I_{\Theta_I} \subseteq \mathcal{I}$ is a parametric family of interaction models, parameterized by $\theta_I \in \Theta_I$.
- $A_{\Theta_A} \subseteq \mathcal{A}$ is a parametric family of algorithmic models, parameterized by $\theta_A \in \Theta_A$.
- $C_{\Theta_C} \subseteq \mathcal{C}$ is a parametric family of contexts, parameterized by $\theta_C \in \Theta_C$.
- $\Lambda \subseteq \mathcal{L}$ are nuisance parameters held constant across all variations of I, A, C.

Each instantiation of a meta-prototype corresponds to a specific prototype: $P_\theta = (I_{\theta_I}, A_{\theta_A}, C_{\theta_C}, \Lambda)$ for some $\theta = (\theta_I, \theta_A, \theta_C, \theta_\Lambda) \in \Theta_I \times \Theta_A \times \Theta_C \times \Lambda$. The set of all possible prototypes generated by a meta-prototype is therefore given by $\mathcal{P}_M = \{P_\theta \mid \theta \in \Theta_I \times \Theta_A \times \Theta_C \times \Lambda\}$. This allows for systematic exploration of a wide design space through parametric variations.

Proposition 1 (Equivalence of Prototypes and Meta-Prototypes): Let $M = (I_{\Theta_I}, A_{\Theta_A}, C_{\Theta_C}, \Lambda)$ be a meta-prototype, where:

- I_{Θ_I} denotes a family of instances parameterized over the space Θ_I,
- A_{Θ_A} denotes a family of actions parameterized over the space Θ_A,
- C_{Θ_C} denotes a family of constraints parameterized over the space Θ_C,
- Λ is a fixed set of nuisance components, common to all prototypes derived from M.

Then, for any prototype $P = (I, A, C, \Lambda)$ with the same Λ as M, there exists a parameter tuple $\theta = (\theta_I, \theta_A, \theta_C)$, where $\theta_I \in \Theta_I$, $\theta_A \in \Theta_A$, $\theta_C \in \Theta_C$ such that $P = (I_{\theta_I},\ A_{\theta_A},\ C_{\theta_C},\ \Lambda)$. In other words, every such prototype P can be obtained by instantiating the meta-prototype M with specific parameters from its parameter spaces. We also define the mapping $s : \Theta_I \times \Theta_A \times \Theta_C \to \mathcal{P}_\Lambda$, where $\mathcal{P}_\Lambda$ is the set of all prototypes sharing the fixed Λ, by $s(\theta) = (I_{\theta_I},\ A_{\theta_A},\ C_{\theta_C},\ \Lambda)$. Then, s is a surjective function, i.e. $\forall P \in \mathcal{P}_\Lambda, \exists\, \theta \in \Theta_I \times \Theta_A \times \Theta_C$ such that $s(\theta) = P$.

This establishes that any prototype $P \in \mathcal{P}_\Lambda$ can be viewed as an instantiation of a meta-prototype M (under the appropriate parameterization). The meta-prototype M acts as a template, and the prototypes P are specific instances derived from M through the sampling of these parameters.

12.3.2 Design Spaces

Lim et al. conceive of prototypes as *filters*, explaining how they help designers "traverse design spaces" [297], arguing that they structure design decisions by isolating certain design dimensions. This view holds for traditional prototypes, since in selecting what to design with or control for, the designer necessarily leaves much else out. However, traditional prototyping processes necessitate iterative, low-dimensional

exploration of this design space, because the attention of the designer is on translating the design space dimensions into the final prototype artifact.

Here, we extend this notion and formalize it using the *meta-prototypes* definitions. In particular, designing a meta-prototype involves constructing an explicit design space by selecting the meta-prototype parameters $(\theta_I, \theta_A, \theta_C)$. The source of these parameters typically encodes some combination of prior work in the target domain, the designer's intuition, and potential empirical sources such as need-finding studies. However, design spaces are routinely extracted from such sources explicitly in prior work.

As a case study, let us consider the specific instance of *writing assistants*, a ubiquitous technology increasingly dependent on large, generative language models. Recent work by Lee et al. [290] constructs a design space in this domain through a systematic literature review. In their taxonomy, the *Technology* considers parameters we could conceptualize under the Θ_A parameter, i.e. those pertaining to algorithmic components of prototypes for writing assistants. These include features like data source (e.g. experts, crowdworkers), model type (rule-based, deep neural network, foundation model), and model access to external resources like tool use and data stores. If we consider these three parameters, they become dimensions of Θ_A, and the values they take on form the hypothetical parameter vectors to be searched over.

Another aspect they consider is *Interaction*. Here, they consider features relevant to Θ_I, the interaction model, for example how the model output is differentiated from the user output (such as through formatting or location), how system output is triggered (user- or system-initiated), the layout of the writing UI, and other such factors. Analogously, implementing a set of parameters and possible values here would construct a meta-prototype design space Θ_I from which specific versions can be instantiated. Finally, Θ_C parameters, referring to contextual factors, relate to what Lee et al. term *Task* and *User* dimensions. These include elements like the purpose of a writing task (e.g. expository vs. descriptive), and user capabilities (e.g. writing expertise).

12.3.3 Prototype Instantiation

The process of instantiating concrete prototypes from a meta-prototype is important, but complex. This step bridges the gap between abstract design concepts and

testable systems. We need a mechanism that can consistently and reliably transform our parameter selections into usable prototypes. To instantiate a prototype from a selection of parameter values, we define a rendering function f that takes in parameters and outputs a prototype: $f : \Theta_I \times \Theta_A \times \Theta_C \times \Theta_\Lambda \to \mathcal{P}$ where $f(\theta_I, \theta_A, \theta_C, \theta_\Lambda) = P_\theta = (I_{\theta_I}, A_{\theta_A}, C_{\theta_C}, \Lambda)$.

For this framework to be effective, we need to ensure that all possible combinations of parameter values can be automatically rendered, so we might uncover potentially unexpected interactions between different design choices. For this, the design of the rendering function f must meet several necessary conditions:

1. **Completeness**. For each parameter Θ_i, there must exist a complete set of rendering instructions that cover all possible values. This guarantees that, for any combination of valid parameter values, the state of the system is not undefined.

2. **Independence**. The specification of each parameter must be independent of the others. This independence ensures that changing one parameter doesn't unexpectedly alter the effects of others, preserving the integrity of our design space exploration.

3. **Composability**. The rendering function must be able to compose the individual parameter renderings into a coherent prototype. This implies the existence of a composition function g such that: $f(\theta_I, \theta_A, \theta_C) = g(h_I(\theta_I), h_A(\theta_A), h_C(\theta_C))$ where h_i are the individual rendering procedures for each parameter.

4. **Finiteness**. Each parameter Θ_i must be finite, or sampled in a finite way for practical implementation. Note this does not necessitate discreteness; bounded but continuous parameters can be sampled from.

To illustrate these concepts, let's continue our discussion of the writing assistant meta-prototype. This example demonstrates how the abstract concepts of rendering functions and parameter spaces translate into practical design choices:

- Θ_I = {suggestion layout}
- Θ_A = {model, decoding temperature}
- Θ_C = {expertise level}

The rendering function f could be implemented as follows:

- $h_I(\theta_I)$: Renders the interactive components:
  - "suggestion layout" ($\in$ {side-by-side, inline}): Creates a split-screen interface or embeds assistance within the main text area. Implemented as a conditionally rendered UI component.
- $h_A(\theta_A)$: Initializes the algorithmic components:
  - "model" ($\in$ {GPT-3.5-Turbo, Claude 3.5 Sonnet}): Calls a different language model, as options specified.
  - "decoding temperature" ($\in [0, 2]$): Adjusts the language model's decoding temperature, for the selected model.
- $h_C(\theta_C)$: Adjusts factors relevant to the interaction context:
  - "expertise level" ($\in$ {novice, expert}): Changes a line in the language model prompt to adapt to level of expertise in the writing task.

The composition function g would then combine these rendered components as needed. In its simplest form, it simply maps the parameters to their target implementations. In a more complex instance, it can impose some post-hoc dependence between the parameters to align their combinations in the system. For example, it may be that decoding temperatures behave differently per model, and the designer has a prior notion of how these should be modified on a per-model basis. The composition function g allows such conditions to be imposed, without changing the design specification and sampling approach for the original parameters

This setup allows for automatic rendering of all possible combinations of the variable parameters, each resulting in a unique prototype. Without the decoding temperature, there are $2^3 = 8$ possible prototypes. However, the decoding temperature is a bounded continuous variable. Therefore, to construct a prototype, the temperature must be sampled from some distribution. A simple strategy would be to sample temperature $\sim \mathcal{U}(0, 1)$. However, this may not be the most efficient way to proceed, since it assumes all temperatures are and will remain equally valuable to explore. A more purposeful sampling strategy might adapt instead, depending on accumulated evidence.

12.3.4 Sampling Strategy

As Almaatouq et al. note [6], uniform and random sampling from design spaces are both desirable because they are unbiased. However, they are also inefficient. In the above meta-prototype example, we considered only a small number of variables taking on a small number of values. Sampling uniformly or randomly would correspond to 8 discrete experimental conditions, with a covariate for the temperature. In more complex designs, for example even 10 binary parameters, the set of possible combinations is $2^{10} = 1024$. Uniformly sampling from this space is intractable, as it would necessitate a large sample of users to yield generalizable results. Randomly sampling could also be wasteful. Instead, *adaptive* sampling methods allow us to focus our exploration on the most promising or informative regions of the parameter space, based on the results of previous experiments. One powerful approach for adaptive sampling is Bayesian optimization. This method treats the problem of finding optimal parameter configurations as a sequential decision-making process, where each decision is informed by all previous observations. Another class of adaptive sampling strategies comes from the multi-armed bandit (MAB) literature. These methods balance exploration (trying new parameter configurations) with exploitation (focusing on known high-performing regions). For example, Thompson sampling [503] maintains a probability distribution over the optimal parameter configuration, and samples from this distribution to decide which prototype parameters to test next.

An important factor to consider here is that adaptive sampling strategies require well-defined evaluation criteria to guide the sampling process effectively. These criteria are essential for determining which prototypes to explore next and for assessing the overall performance of different configurations. A few potential options for this could be:

1. *Time on task:* one goal could be facilitating efficient completion of a task, as judged by the user themselves. As such, adaptive sampling strategies could enable sampling from the region of the parameter space that minimizes this quantity.

2. *Engagement:* By contrast, user engagement time or repeated use can indicate valuable regions of the design space too.

3. *Quality:* if a quality metric can be defined, then this can be used to guide the sampling process. Note that this does not necessarily need to be automated.

Crowd-sourced ratings of outputs can be a scalable and asynchronous evaluation signal that enables sampling decisions to be driven by a notion of quality.

4. *Usability:* participants can be asked to complete a System Usability Scale [239] or other such instrument, through which usability-maximizing prototype versions become feasible to explore.

5. *Cognitive load:* similarly, using a simple instrument like the NASA TLX [206], or proxies for task effort taken from user interaction logs, can be useful in finding regions of the parameter space that allow cognitively easier performance.

6. *Combinations:* (possibly weighted) combinations of these factors can be used together, to satisfy multiple design criteria.

Such criteria can then be used as an objective or reward function for the sampling strategy.

12.3.5 Implementation Approach

So far, we have kept our formulation of meta-prototypes abstract to provide a foundation for a potential implementation. Implementations of this basic specification can take many forms, to suit the needs of researchers working with different tools, domains, and populations. Here, we provide an example implementation of the writing assistant meta-prototype we discussed previously, to anchor our discussion in a realistic path to implementation.

I_Θ. Since our simple interaction model variable only considers the suggestion layout, this can be implemented using a conditional rendering[1] approach. This effectively implements a branching structure, dependent upon some UI property. Here, the property only takes two possible values, so deriving the UI from it is straightforward (just an if statement). For more complex UIs, other such principles of declarative UI programming, wherein θ_I is treated as the state, are possible to implement. For example, rendering lists[2] can be used to operationalize an "option count" variable (e.g. showing a different number of generated options to a user for them to consider integrating).

[1] https://react.dev/learn/conditional-rendering
[2] https://react.dev/learn/rendering-lists

A_Θ. The algorithmic model variable in our meta-prototype encompasses two key aspects: the choice of language model and the decoding temperature. Implementation of this variable involves setting up an abstraction layer that can interact with different language model APIs uniformly. This can be achieved using a factory pattern [165], where a single interface is used to create objects (in this case, API clients) for different language models. Such a pattern would also be applicable if the API were user-implemented with a non-API-hosted model. The decoding temperature can then be directly passed as a parameter to the chosen model's API call. These are simple examples, but abstractions of the kind needed here are often well-supported by existing software design patterns. The exercise of the prototyper becomes one of choosing the appropriate abstraction, and implementing accordingly.

C_Θ. The context variable in our example considers the user's expertise level, which affects how the system interacts with the user. This can be implemented using a strategy pattern [165], where different strategies (e.g., NoviceStrategy, ExpertStrategy) encapsulate the logic for adapting the system's behavior based on the user's expertise. These strategies could textually prescribe various aspects of the system's intended behavior, such as the complexity of language used in suggestions, the level of detail in explanations, or the frequency of interventions. The chosen strategy could then be used to modify the prompts sent to the language model. This approach allows for a clear separation of concerns and makes it straightforward to add new expertise levels or other contextual factors in the future. Note that adaptation strategies may vary; here, they are cleanly separable from the other variables. In other cases, they may need to be entangled in aspects of I_Θ or A_Θ. An important decision must be made here about whether to conceptually model them as part of these aspects, or of user context, with the difference potentially affecting the process by which inferences are drawn from the results and how these inferences are interpreted.

12.4 Towards More Predictive Theories of Human-Generative AI Interaction

Human-generative AI interaction involves complex, rapidly changing socio-technical systems. We believe that the meta-prototypes approach, applied thoughtfully, can help to build more stable and generalizable design knowledge by making prototypes *commensurable*.

12.4.1 Quantifying Interaction Dynamics

One of the most promising aspects of the meta-prototype framework is its potential to quantify the dynamics of human-AI interaction in unprecedented detail. By systematically varying parameters across the interaction model, algorithmic model, and context, we can begin to build quantitative models of how these factors influence user behavior and outcomes. For instance, consider a meta-prototype study of a writing assistant that varies the frequency and randomness of the language model's suggestions. By collecting fine-grained data on user actions, writing outcomes, and subjective experiences across these conditions, we could develop predictive models of how suggestion frequency interacts with user expertise to influence writing speed and quality. We could also characterize the relationship between AI randomness and user creativity, potentially revealing synergistic or inhibitory effects. These quantitative models, while inevitably simplified, could provide a foundation for more precise and testable theories of human-AI co-creativity.

12.4.2 Identifying Generalizable Patterns

As we accumulate data from meta-prototype studies across different domains of human-AI interaction, we may begin to identify generalizable patterns that hold across various contexts. These patterns could form the basis of more general theories of human-AI interaction. For example, we might discover:

1. Threshold or scaling effects: There may be critical thresholds in AI capability or interface design beyond which user behavior changes dramatically. Identifying these thresholds could lead to theories about phase transitions or scaling laws in human-AI partnerships.

2. Interaction archetypes: Certain combinations of user characteristics, task types, and AI capabilities might consistently lead to specific interaction patterns. These archetypes could inform a taxonomy of human-AI interaction styles.

3. Learning trajectories: By studying how users adapt to a family of different AI systems over time, we might uncover common trajectories of skill development and AI reliance. This could lead to predictive models of long-term human-AI collaboration dynamics.

Developing generalizable notions about these factors would allow designers to make principled choices for future systems.

12.4.3 Bridging Levels of Analysis

Rogers [431] considers ambitious those theories in HCI research which bridge different levels of analysis. For example, theories might bridge low-level cognitive processes with high-level social dynamics. The meta-prototype framework, with its ability to systematically vary factors at multiple levels, offers a unique opportunity to develop integrative theories that span these levels in the context of human-generative AI interaction problems. For instance, a meta-prototype study could simultaneously manipulate low-level interface features (e.g., suggestion highlighting), mid-level interaction strategies (e.g., proactive vs. reactive AI assistance), and high-level contextual factors (e.g., collaborative vs. competitive task framing). By analyzing how these factors interact, we might seek to develop multi-level theories that explain how cognitive, interpersonal, and social factors combine to shape human-AI interaction.

12.4.4 Studying Counterfactuals

Perhaps most directly, the systematic nature of meta-prototype exploration enables a form of counterfactual reasoning that is often difficult in traditional HCI research beyond (potentially contrived [15]) control interface variants. By mapping out a broad design space, we can start to answer "what if" questions about alternative design choices or technological capabilities. For example, what if language models could provide perfect writing improvement suggestions but no creative suggestions? Does an observed human-AI interaction limit hold in the presence of a sufficiently more capable model? This counterfactual reasoning capacity might allow us to more systematically answer questions and test hypotheses of this nature.

12.5 Potential Limitations

The meta-prototype framework introduced in this chapter aims to propose a more systematic approach to exploring the design space of human-AI interaction systems. In this discussion, we critically examine the implications and potential pitfalls of this approach.

12.5.1 The Illusion of Completeness: Navigating Infinite Design Spaces

One of the primary motivations for the meta-prototype framework is the desire to more comprehensively explore the vast design space of human-AI interaction systems. However, we must be cautious about claiming any level of "completeness" in this exploration. As Gero and Kumar [180] note in their work on design spaces:

> Creative design may occur when new design variables are introduced in the process of designing. Thus, in creative design, the designer operates within a changing state space of possible designs; a state space which increased in size with the introduction of each new variable. [180]

Consider our example of a writing assistant meta-prototype. While we've parameterized aspects like suggestion layout, model choice, and user expertise, countless other variables remain unaccounted for, such as the specific wording and tone of AI suggestions, their timing and frequency, and other task-relevant aspects of the visual interface design such as whether suggestions are revealed at once or streamed character-by-character simulating typing, akin to current era chatbots. Each of these factors could potentially influence the effectiveness of the system; it quickly becomes clear that capturing all possible variables in a meta-prototype framework is likely impossible. This raises a crucial question: How do we decide which parameters to include in our meta-prototype, and which to leave out?

The challenge here is that by formalizing certain parameters, we may inadvertently neglect others that could be important. We risk heuristically substituting the complex, multifaceted problem of designing effective human-AI interactions with a more tractable problem of optimizing over a limited set of parameters. As such, it is important to revisit the design spaces we assume.

12.5.2 Exploration vs. Exploitation in Adaptive Sampling

On one hand, adaptive sampling allows us to focus our resources on promising areas of the design space. On the other hand, it may lead us to prematurely converge on local optima, missing potentially superior designs that lie in unexplored regions. This tension is particularly acute in the context of human-AI interaction, where the "reward landscape" of designs may be highly non-linear and context-dependent. Another challenge is optimizing for longer-term outcomes. Consider the case where

we optimize for task completion time, but potentially sacrifice long-term learning and skill development.

Each of these choices embeds value judgments about what constitutes a "good" human-AI interaction, judgments that may vary across users, tasks, and cultural contexts. The meta-prototype framework, in its current form, does not provide guidance on how to navigate these value-laden decisions when making assumptions that shape the sampling decisions. As such, one important step could be maintaining uncertainty rather than only seeking optimal designs. Methods discussed in this chapter, such as Bayesian optimization, do just this; allowing us to probabilistically sample prototypes from promising regions, rather than converging to a locally optimal prototype. Importantly, the implicit sampling already underway in prototype-by-prototype interactions has no guarantee of searching underexplored regions either, and this problem is a more general one.

12.5.3 Integrating with Other Methods

Designers use a diversity of methods. So far, we have assumed that quantitative methods are appropriate to the context at hand, and shown how such methods can be augmented through the use of meta-prototypes. However, future work might consider how to extend this framework to other prominent design methods, like:

1. Qualitative methods: Incorporating qualitative approaches can provide deeper insight into the "how" and "why" behind user behaviors observed in meta-prototype studies. As King et al. suggest [258], qualitative and quantitative methods share an underlying logic and can be complementary. As a starting point, pre-study interviews could inform parameter space definition, while active learning techniques might help surface instances, users, or phenomena that would benefit from a deeper post-study qualitative analysis.

2. Participatory design: Extending the meta-prototype approach to allow participants to actively explore and shape the design space could uncover overlooked design dimensions. This aligns, in principle, with the notion of participatory design [482]. User-defined parameters and interactive exploration sessions could allow participatory design patterns. Methods like Markov Chain Monte Carlo with people [445] or Gibbs Sampling with people [205] could be helpful to make inferences in such contexts.

3. Complexity science and systems theory: Human-AI interactions are complex adaptive systems, characterized by emergent behaviors and non-linear relationships. Integrating with methods from complexity science could help identify emergent patterns of interaction [39] in these systems.

12.6 Conclusion

The meta-prototype framework introduced in this chapter represents a novel approach to systematically exploring the design space of human-AI interaction systems. By formalizing the concept of meta-prototypes and their associated design spaces, we offer researchers a tool for navigating the vast space of possible AI-assisted systems in a more principled manner. As a design proposal, much work is needed to actualize its potential, such as developing standardized tools and libraries for implementing meta-prototypes across different domains. We present this initial proposal as an invitation to the HCI community, to critically and collectively explore ways to build more systematic, theoretically grounded approaches to human-AI interaction design. While the meta-prototype framework does not solve all the problems this goal faces, we believe it offers a valuable step towards a more predictive science of human-AI interaction.

A

Supplementary Material

A.1 Supplement for Chapter 2

As supplementary material, we present and review a number of input/output examples across several categories with distinct properties[1]. A summary of these results is shown in Table A.1.

Finally, to gain a qualitative view of intra-scene and adjacent-scene consistency, we plot our test set input images according to the corresponding output audio characteristics by a visualization shown in Figure A-11. We produce multiband T_{60} estimations from all output IRs, and then used t-SNE [319] to reduce the data dimensionality to two dimensions. We then solve a linear assignment problem to transform this into a grid representation. Several instances of within-scene clusters are visible, as well as closeness of related scenes. This suggests that while our method does make errors (outliers are also visible), it learns to treat similar scenes similarly while capturing variation.

[1]Link to audiovisual examples page: https://web.media.mit.edu/ nsingh1/image2reverb/

Topic	**Figure #**	**Images**
Famous and iconic places	A-1	6
Musical environments	A-2	6
Artistic renderings	A-3	6
DALL•E-generated spaces	A-4	6
Limitations (i.e. challenging examples)	A-5	4
Animated scenes	A-6	6
Virtual backgrounds	A-7	6
Historical places	A-8	5
Video games	A-9	4
Common and identifiable scenes	A-10	6
Total		55

Table A.1: Additional Results.

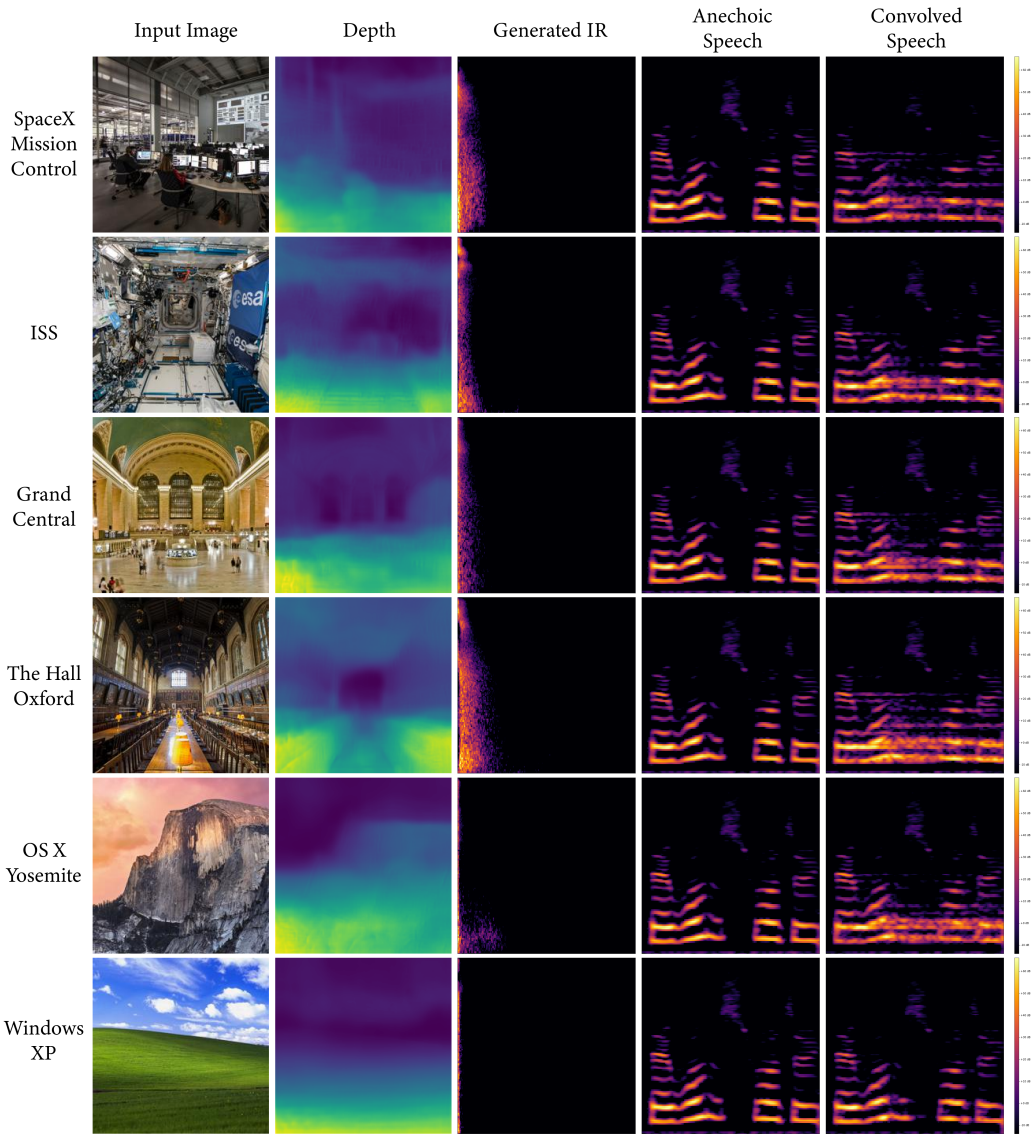

Figure A-1: Famous and iconic spaces. Columns show input images, depth maps, generated IRs, and a dry anechoic speech signal before and after the generated IR was applied to the signal via convolution. The input images come from spaces that may be impractical or impossible to record in. The indoor spaces here show longer impulse responses compared to the outdoor scenes which is typically observed and expected in real-world settings. Larger indoor spaces also tend to exhibit greater T_{60} times with longer impulse responses which we see here, though the ISS image has a longer impulse response than we expect.

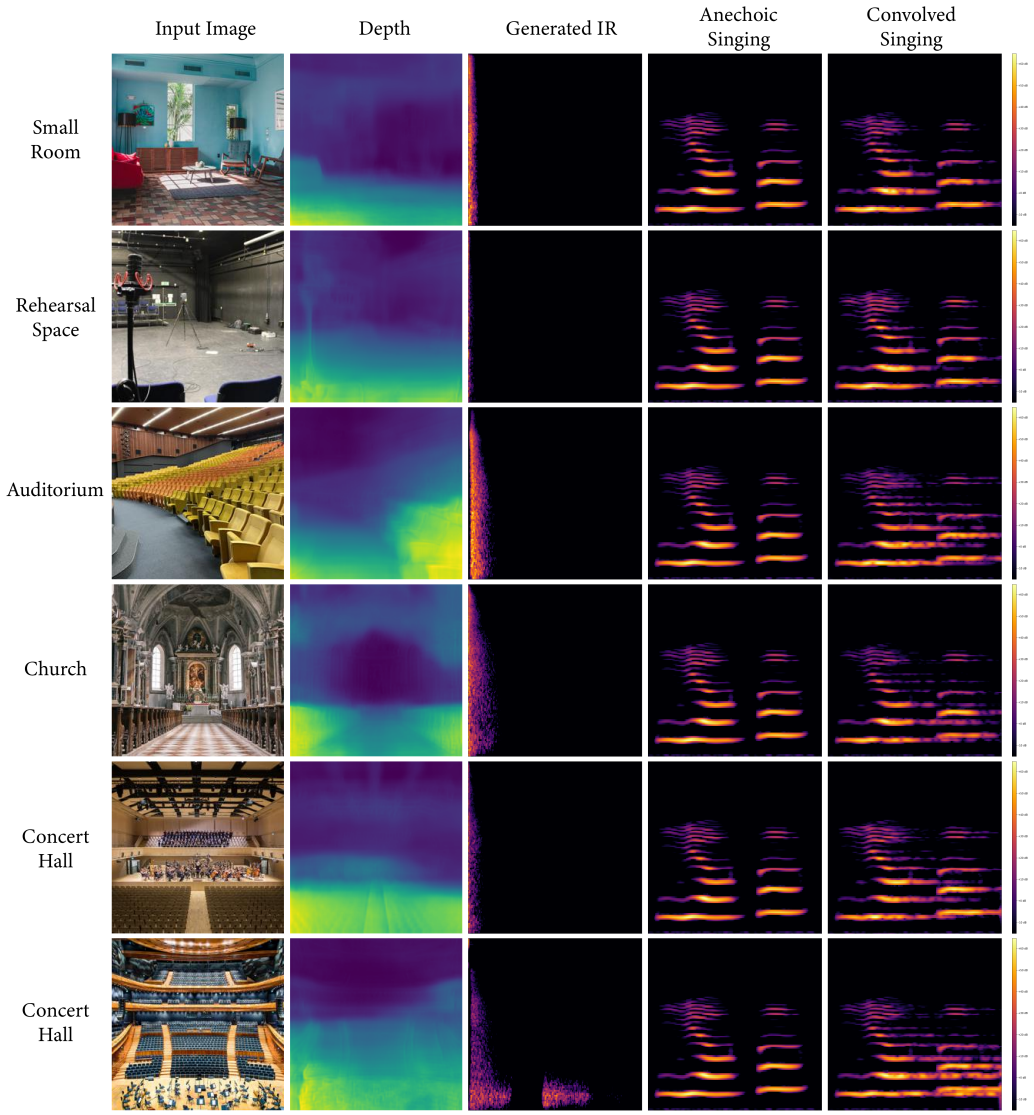

Figure A-2: Music. Columns show input images, depth maps, generated IRs, and an anechoic vocal singing signal before and after the generated IR was applied to the signal via convolution. The input images come from spaces relevant to music including a typical small room, an acoustically treated rehearsal space, an auditorium, a church, and 2 large concert halls. Generally, larger spaces tend to exhibit longer decay times in the output, however some examples such as the concert halls with visible acoustic treatment appear to have a shorter decay than more reverberant spaces like the church or auditorium with more reflective surfaces. The final concert hall shows an atypical impulse response with a visible discontinuity in the IR tail. This is not commonly observed among our model outputs, but illustrates the nature of artifacts which can occasionally occur.

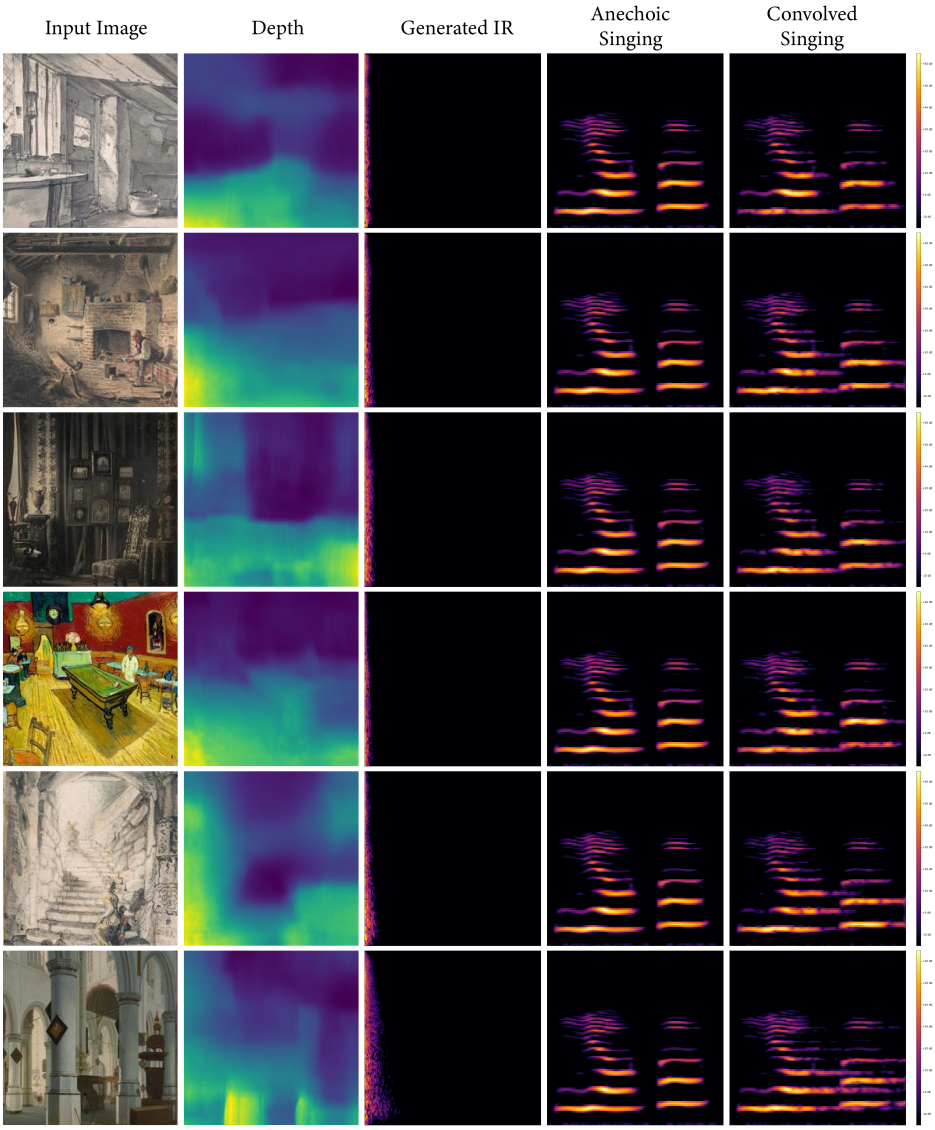

Figure A-3: Art. Columns show input images, depth maps, generated IRs, and an anechoic operatic singing signal before and after the generated IR was applied to the signal via convolution. Images here are drawings, paintings and a vintage art photograph ca. 1850. Artistic depictions of spaces were not included in our training dataset. In many cases, plausible impulse responses are generated from such input images. In general, larger depicted spaces, like the church in the bottom row, exhibit longer decay times as is observed with standard 2D photographs.

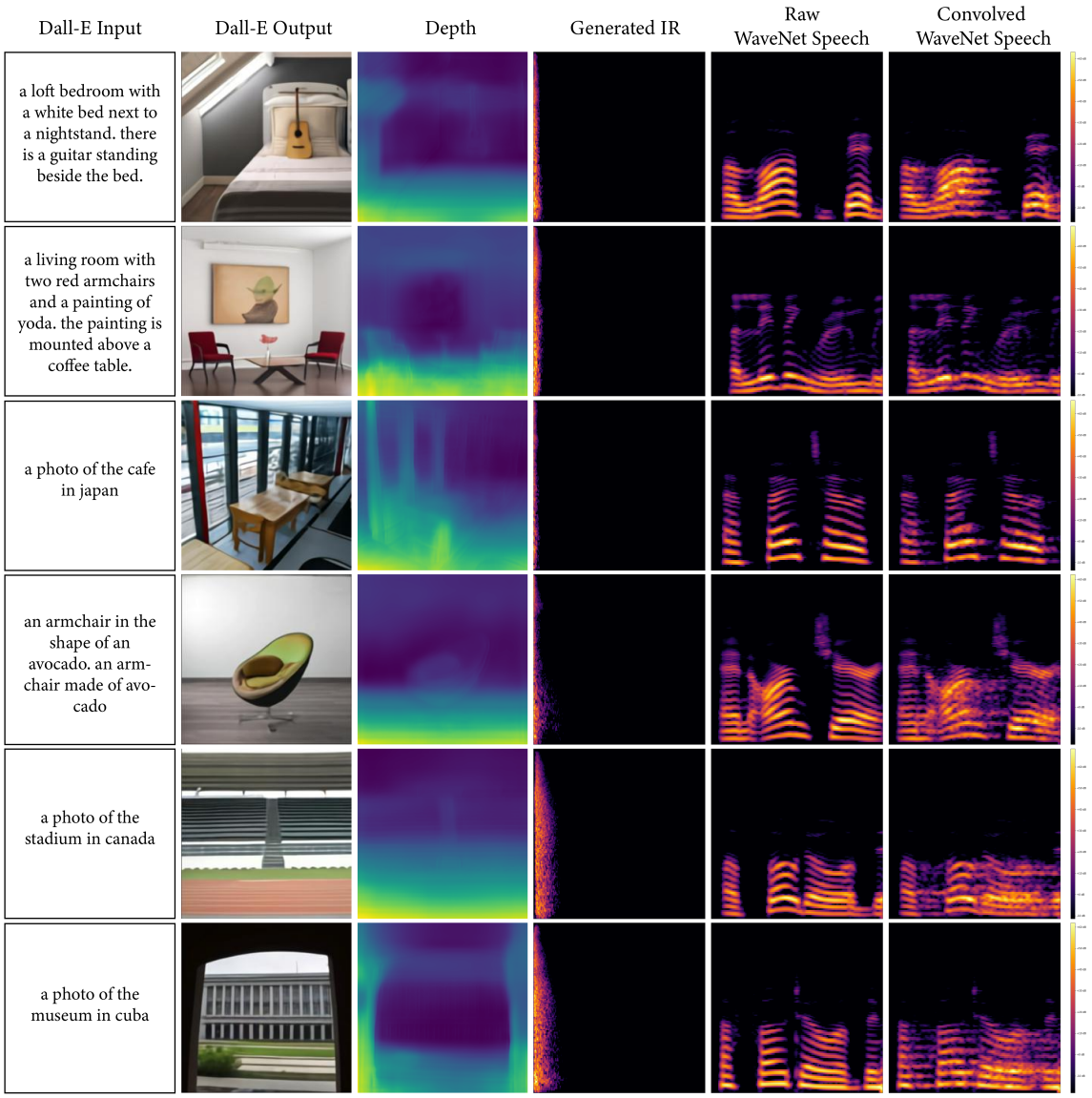

Figure A-4: DALL·E. Images generated from text by DALL·E [406] used here as input images. The same corresponding input text was synthesized via text-to-speech as our signal of interest and convolved with the generated IR. This reflects synthetic speech in a synthetic environment, indicating a path for synthesizing realistic IRs from text. It also shows how our model might work with other state-of-the-art generative media models to produce more consistent and realistic results in different domains.

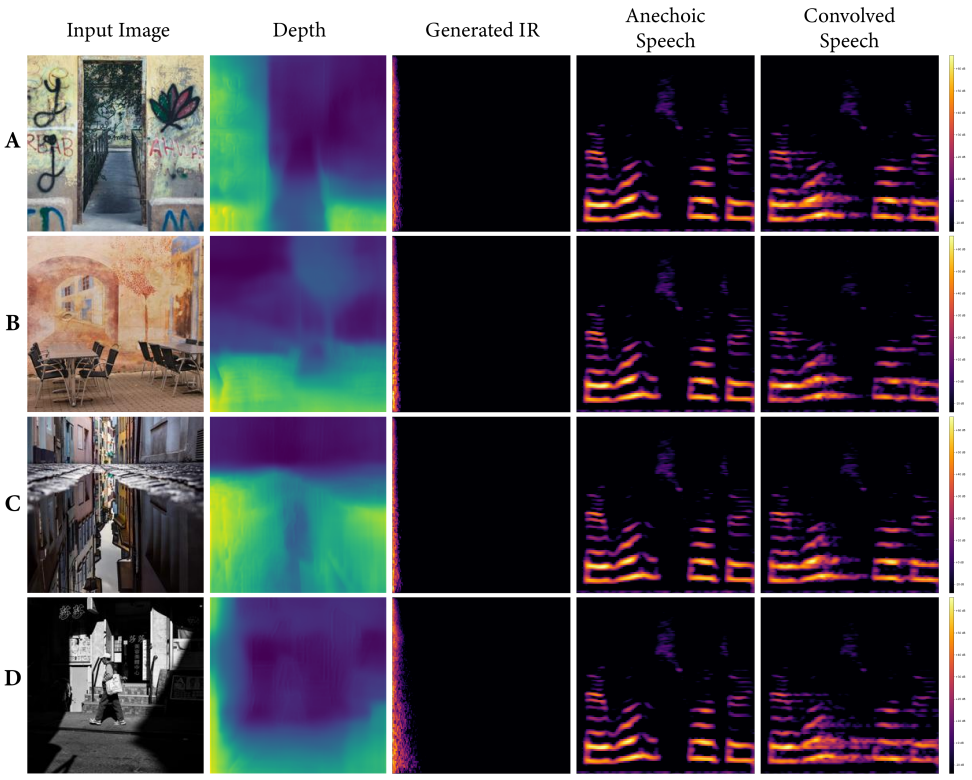

Figure A-5: Challenging images. Input images containing street murals, reflections, and shadows demonstrating cases where depth is inaccurately estimated. (A) A painted doorway giving the illusion of depth. (B) A wall with a mural of a street and tree where the depth of the wall is inaccurately estimated. (C) A low-angle photo of a reflective puddle. (D) An outdoor street image with strong shadows which results in a depth map and generated IR more similar to a room than an outdoor space. These more extreme scenarios are chosen to clearly illustrate the limitations of our approach.

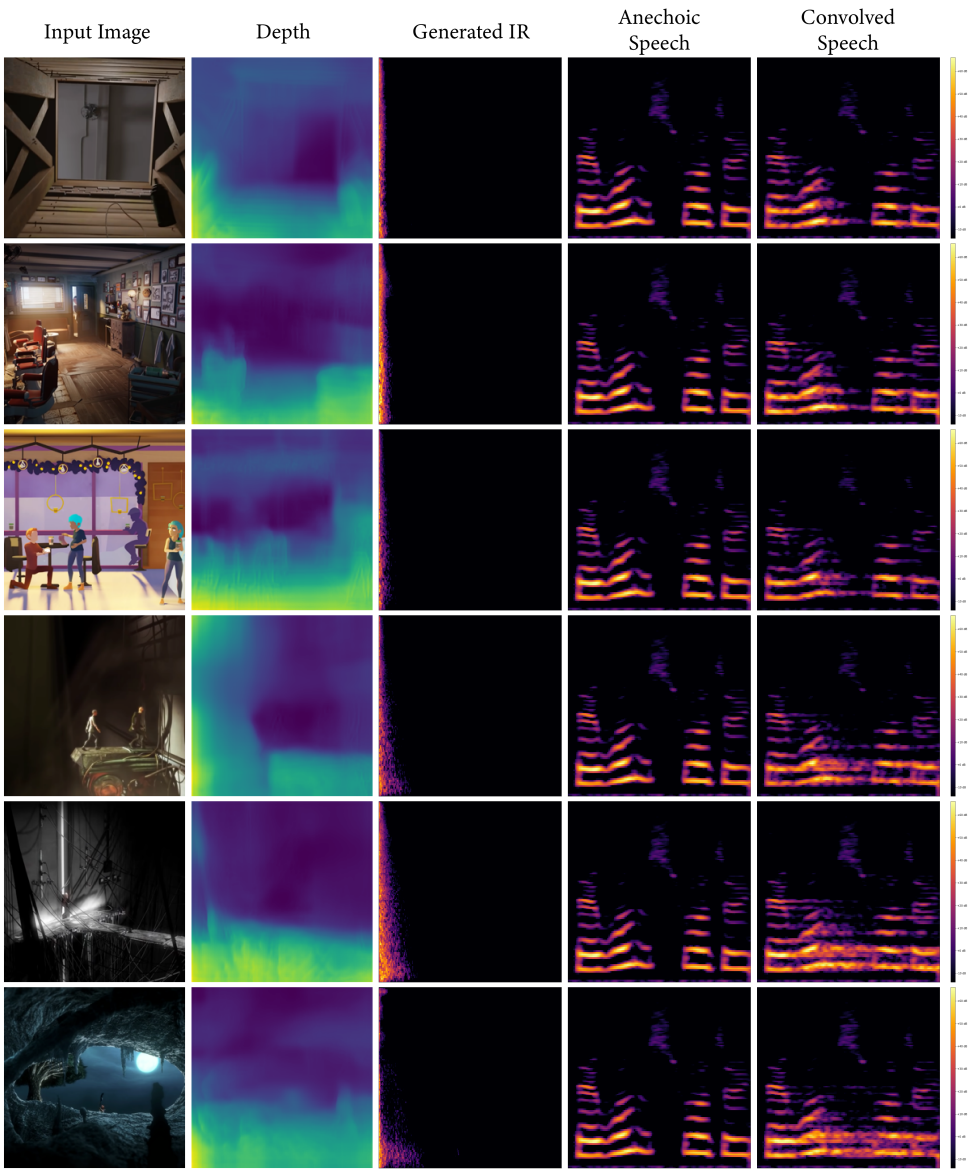

Figure A-6: Animated films. Scenes from Blender open animation films used as input images (speech convolved with generated IRs). Columns show input image, calculated depth map, spectrogram of generated IR, an anechoic passage reading sample, and the same passage with the generated IR applied via convolution. In general, we find that our model plausibly estimates the reverberant characteristics of these spaces. For example, the wooden small space is very brief. The barbershop appears longer due to some artefacts, but the broadband decay is relatively quick as can be heard in the audio. Seemingly larger spaces again correspond to longer IRs. This is a case of Real2Sim transfer, where we can approximate IRs directly that sound as measured IRs, but in virtual environments where this measurement is not possible.

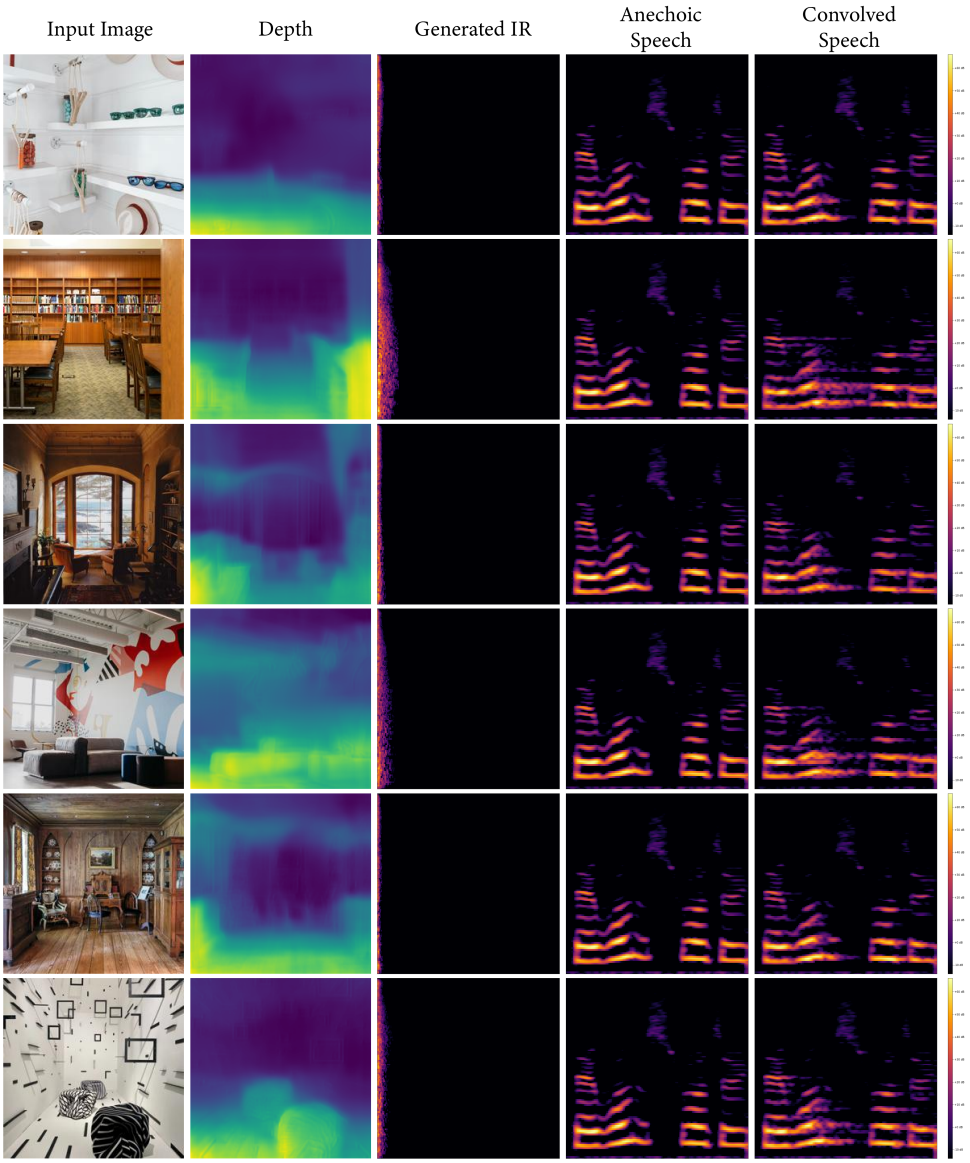

Figure A-7: Virtual backgrounds. Images which may serve as virtual backgrounds used as input images to our model. These reflect spaces that may be used for videoconferencing or other online meetings. Realistic IRs may be generated and used in these contexts to increase the sense of being in a shared space with others.

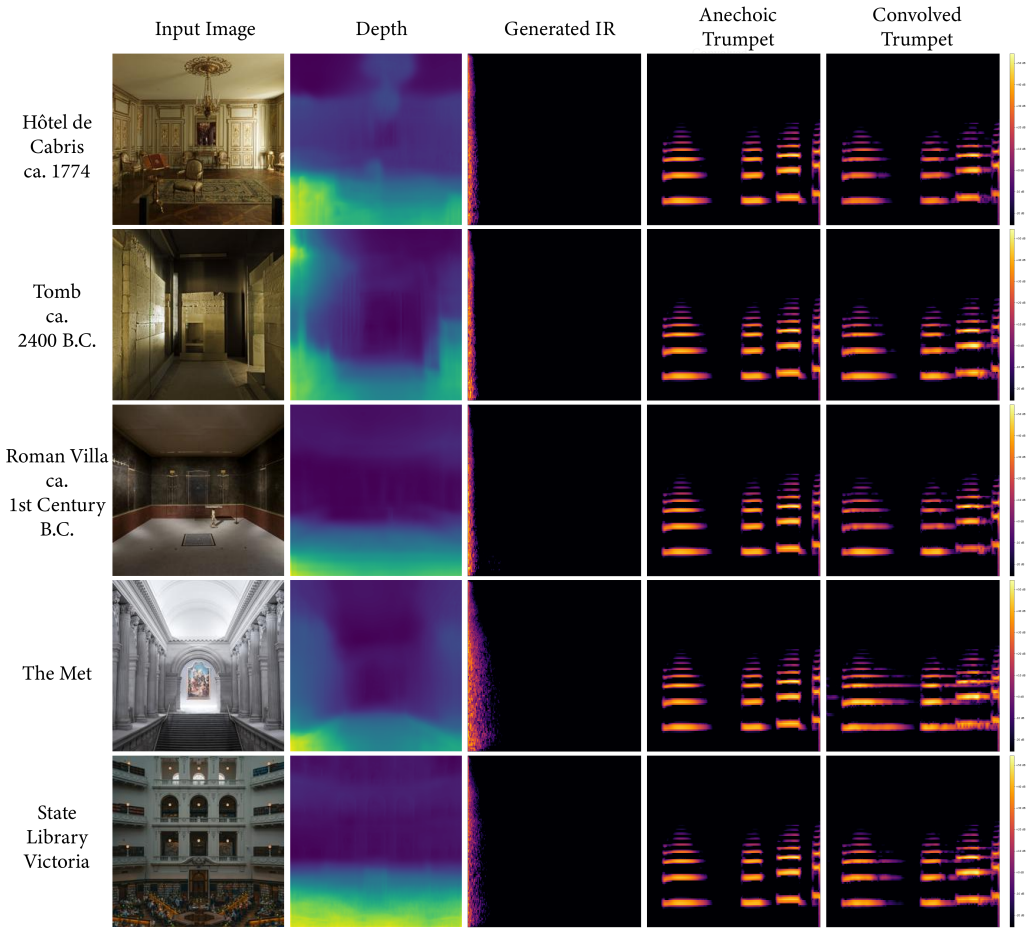

Figure A-8: Historical and notable places. Additional examples of unusual and historical spaces which may be difficult or impossible to obtain IRs from.

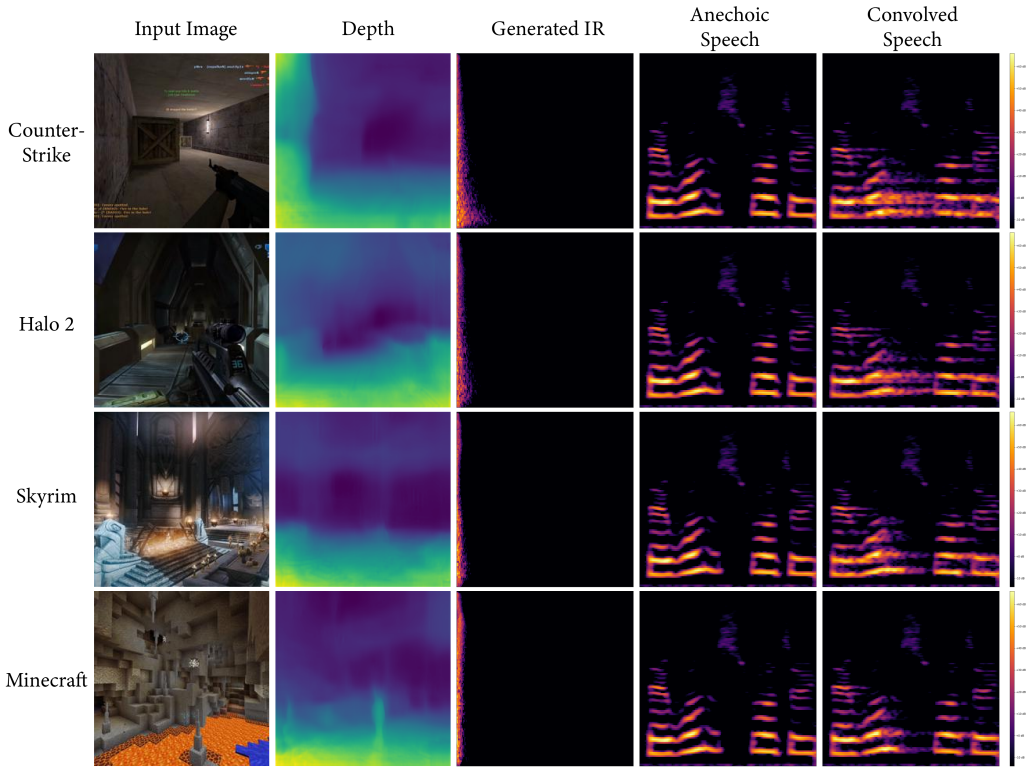

Figure A-9: Video games. Impulse responses generated and applied via convolution from screenshots of four 3D video games. Video games are one example of a virtual space that might benefit from easily generated impulse responses. While the medium sized room from Counter-Strike and the large hallway from Halo 2 may be plausible IRs, the large hall shown in the Skyrim screenshot and the cavern in the Minecraft example do not have correspondingly long reverberant tails as would be expected showing possible examples of where the scale of the space was not accurately estimated. 3D rendered images were not included in our dataset but are a ripe area of future work which might greatly increase the performance of our model on both real scenes and virtual scenes such as these video game examples.

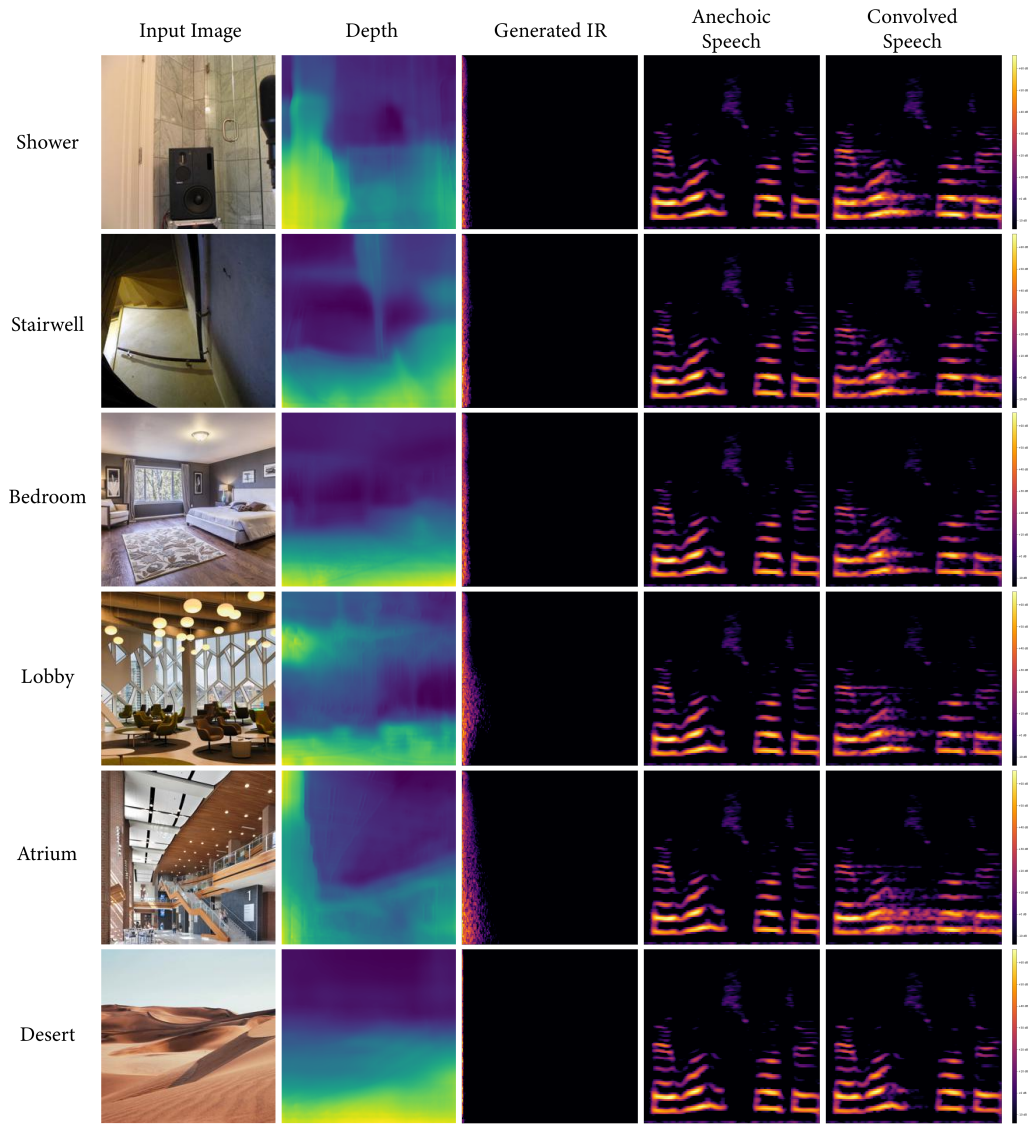

Figure A-10: Common and identifiable scenes. Input images and the resulting IRs are shown and convolved with an anechoic speech signal. Input images here reflect spaces that are regularly encountered in everyday life yet may not often be recorded in. These types of scenes are useful for audio post-production as they may be commonly found in movies and television shows. Small and outdoor scenes are observed to have very brief IRs while in comparison, the larger building interior has a much longer output IR as expected.

Figure A-11: Manifold-based visualization of our test set. We compute multi-band T_{60} estimates for output audio IRs for each image, and then perform nonlinear dimensionality reduction with t-SNE to obtain two-dimensional feature vectors for each example. We produce a grid by solving a linear assignment problem, as is commonly done to visualize large image datasets. Our visualization shows local clusters of same and similar scenes in many cases, but also some variation within scenes. In some outdoor settings, this variation grows considerably large, resulting in increased scattering. In other cases, we observe closeness between different views of the same scene and similar scenes.

A.2 Supplement for Chapter 3

A.2.1 Pretraining Details

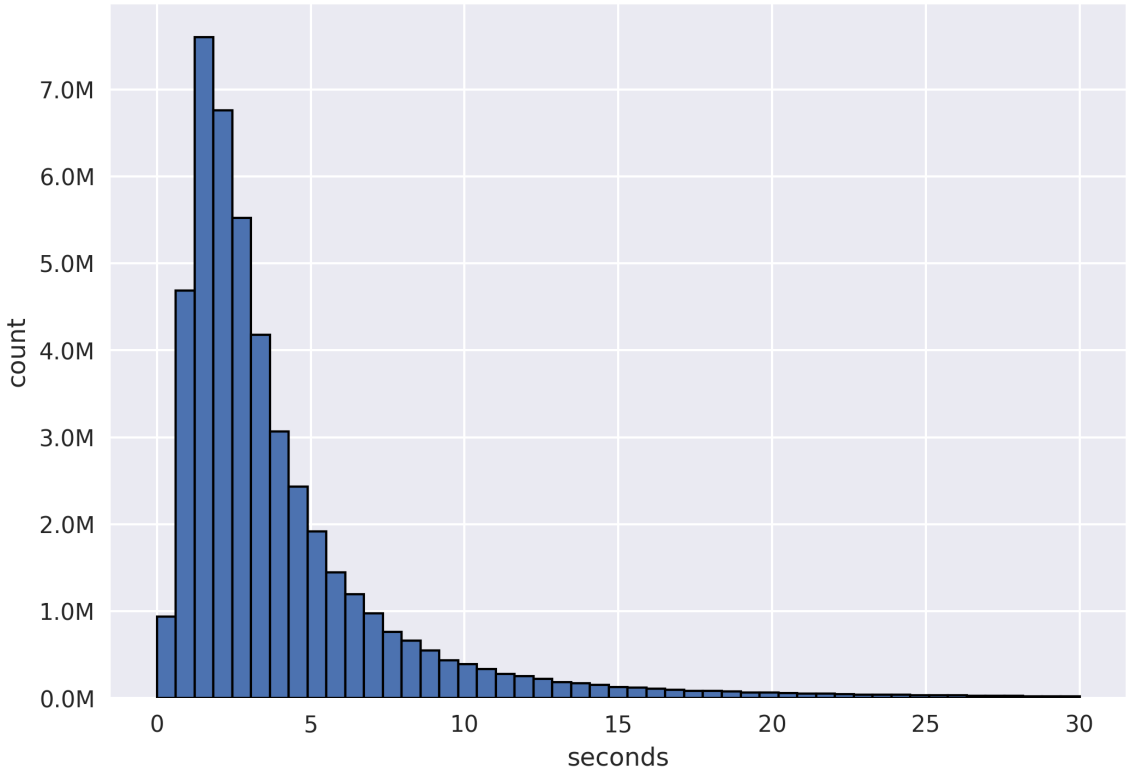

Figure A-12: Distribution of shot lengths observed in our dataset.

Data Preprocessing We temporally segment long-form content into shots (camera changes). Fig. A-12 shows the distribution of shot lengths. We ignore shots that are shorter than 3 and longer than 12 seconds. The former constraint is to make sure the snippet is long enough for our models, while the latter is to improve training throughput. The total number of shots in each pretraining setting is shown in Table 3.1 under the column #data. When creating a minibatch during pretraining, we ensure that $\frac{1}{8}$ of each batch comes from the same long-form content source (e.g. the same movie) to create hard negatives. The process of generating quadruple training instances (v_p, a_p, v_s, a_s) is as follows:

1. Given a title, randomly pick a shot.

2. Temporal jitter: randomly select two 3-second temporal windows. These two

snippets, derived from the same shot, are our primary and secondary instances. For the secondary instance, the language of the audio is different from the one in the primary instance, if an alternate audio track (i.e. dub) is available.

3. For each pair of audio and video:
   Video: Resample to 25 fps, uniformly sample 16 frames, randomly scale the shorter side of video within the range of 256-320, then perform a random crop of 224x224.
   Audio: Resample to 48kHz, convert audio to mel-spectrogram (n_fft=1024, hop_length=501, num_mels=96), convert to the decibel scale, and apply time and frequency masking with maximum value of 50 percent of the corresponding axis.

Model and Pretraining Hyperparameters. The MLP projection heads have an output dimensionality of 512. The latent embeddings (z) are L2 normalized prior to computing the loss. The temperature factor τ in the objective function is set to 0.07. We use the AdamW optimizer [313] with a learning rate of 3e-4, and weigh decay of 5e-2. We train for 12 epochs on 32 NVIDIA A100 GPUs, with a batch size of 64 per GPU, using a half-cosine learning rate annealing which kicks off after 2 warm-up epochs.

A.2.2 Additional Experiments

Results on Action Recognition

We report results on **UCF101** [479] and **HMDB51** [271], well-known benchmarks, to assess the video-only performance of our models, shown in Table A.2. Performance between our model variants is comparable, showing that the dub-augmented training does not necessarily decrease video-only performance. Additionally, we compare to recent state-of-the-art results which, like us, do not use fine-tuning. Note that these results use linear probes, vs. our MLP probes which were derived from a grid search over probing strategies. Nevertheless, the fact that we significantly beat these results without fine-tuning (>12% absolute) demonstrates the value of our learned representations.

VGGSound Results

We report results on **VGGSound** [82], an audiovisual benchmark on which we focus on audio results, shown in Table A.3. Once again, performance between our model

Model	UCF101 [479]	HMDB51 [271]
B.3	88.90	69.35
B.4	88.20	68.91
B.5	87.99	69.43
FIMA [581]	76.40	47.30
FAME [125]	72.20	42.20

Table A.2: Performance of video models on UCF101 [479] and HMDB51 [271] datasets, comparing with recent results that *do not* involve fine-tuning.

Model	VGGSound
B.3	43.49
B.4	41.95
B.5	42.96
LAION-CLAP [559]	46.20
BLAT [562]	42.90

Table A.3: Performance of audio models on the VGGSound [82] dataset, comparing with recent results that *do not* involve fine-tuning on the downstream dataset. The LAION-CLAP result reported uses keyword-to-caption augmentation.

variants is comparable, and our results are competitive with recent state-of-the-art results which don't use fine-tuning.

Controlled Dataset and Models

In this section, we discuss the methods and results from a smaller-scale, more controlled set of experiments. The pretraining dataset consists of 748 movies, about 1300 video-hours of content. Each movie contains a video track, as well as four audio tracks: English (**EN**) as the primary language, and three dubbed versions, Spanish (**ES**), French (**FR**), and Japanese (**JA**), all languages for which we find dubs are relatively commonly available. Having multiple dub options allows us to investigate trade-offs between secondary languages, and whether "multilingual" models might further strengthen performance.

The video model is a medium X3D [152], which is an efficient ResNet-based model. Our audio model is an Acoustic ResNet50 [560], which takes audio spectrograms as input. Both models output 1024-dimensional representations per clip. We share

backbone weights (i.e. Acoustic ResNet50) across audio variants with primary and secondary (dubbed) languages. We do not share MLP weights for primary vs. secondary audio, to allow for more flexibility. As in our primary experiments, we mainly train these models *cross-modally*, i.e. we compute the contrastive cost between modalities.

We train these models on 4 A100 GPUs for 10 epochs with a batch size of 26 per GPU. We use a negative sampling parameter k (samples drawn from the same movie as the positive clip), which we set to 12 per GPU. We use the AdamW optimizer [313] with β=(0.9, 0.999), a learning rate of 0.001, weight decay of 0.05, and a cosine learning rate schedule with a half-epoch warmup.

In all, we compare the following model variants in these smaller-scale, more controlled, experiments:

1. ***Monolingual*** (**EN**): In this baseline, we consider models trained with two differently-augmented primary (English) audio treated as "primary" and "secondary" (a_p=**EN**; a_s=**EN**) audio respectively. This is to account for any possible effect of two augmentations per seen sample, as occurs for the dub-augmented cases, although it does not modify the data distribution. This is a SimCLR-based setup, with two audio paths each contrasted with video.

2. ***Bilingual*** (**ES**, **FR**, **JA**): We introduce one secondary audio at a time to explore the dub-augmented training hypothesis (a_p=**EN**; a_s= **ES** OR **FR** OR **JA**).

3. ***Multilingual*** (+**EFJ**): Here, we effectively randomly select a secondary audio from the given list (Spanish, French, and Japanese) per batch (a_p=**EN**; $a_s \in_R$ {**ES**, **FR**, **JA**}). The order of samples is randomized, so in practice we simply circle through the list round-robin. We aim to explore whether there are additional benefits or drawbacks to having more than one secondary audio.

4. ***No-Speech*** (**SEP**): We establish another baseline where the speech is separated and we only train on video + non-speech audio. This allows us to examine whether simply removing the speech is enough for a performance gain on non-speech-focused tasks. We use the pretrained Hybrid Demucs v3 model [120] to separate the vocal from the rest, mixing the other stems back together. There is no secondary audio here (a_p=**EN** $_{\text{SEP}}$). Note that this variant is trained with 44.1kHz audio, as this is the input and output sample rate for the Demucs models. Although Demucs is trained for music separation, we find that it works

well on speech in practice on our dataset. We use the default (`mdx_extra_q`) pretrained model.

5. ***Audio-Only*** (Monolingual: **AUD**, and Multilingual: **AUD**$_{+\text{EFJ}}$): Finally, we examine two audio-only models. The data is similar to the ***monolingual*** and ***multilingual*** setups, except without video. The objective function is now *within-modal*, between the two audio clips. The monolingual version represents standard audio contrastive training with two augmented copies. These models cannot work on visual or audiovisual tasks, but here we seek to evaluate whether and how much dub-augmented training contributes improvements in the absence of video.

Evaluation *Evaluation Tasks.* Beyond the HEAR [517] tasks used in our main experiments, we include results from additional audio tasks to this controlled setup to gain a more complete picture in the controlled setup. First, we add audio tasks from HARES [536]; specifically, TUT18 [344] for acoustic scene recognition, Fluent Speech Commands [316] for speech command recognition, and VoxForge [324] for language identification, complementing existing HEAR tasks. As in the appendix for our main results, we include the video-only action recognition tasks HMDB51 [271] and UCF101 [479]. Finally, we add an *audiovisual* task (VGGSound [82]) to facilitate a better comparison with **SEP**, since this baseline sees no speech altogether. We hypothesize that **SEP** will be a strong performer in some cases, but that dub-augmented models will be stronger in general as they preserve the audiovisual relationship between speech actions visually occurring and sounding.

For the visual and audiovisual tasks, we train the linear probes for 200 epochs using Stochastic Gradient Descent and a learning rate of 0.2 following a cosine schedule. We train on 2 A10 GPUs with a total batch size of 1024. For HEAR tasks, we use the provided API's strategy and the 48kHz data. For HARES tasks, we follow the authors' specifications [536]: in general, with 400K training steps and a learning rate schedule consisting of 5K linear warmup steps and a cosine decay for the rest (max. learning rate of 0.0002, with the Adam [259] optimizer). We train on 2 GPUs with a total batch size of 64. In all relevant cases, we duplicate mono audio to the second channel to form a pseudo-stereo input to match our model's architecture.

	Task	M	Baselines (SimCLR) AV	SEP	A	Dub-Augmented ES	FR	JA	EFJ	A_{+EFJ}
Snd/Scn	ESC-50 [393]	A	.527	.570	.220	.580	.575	**.590**	.587	.550
	FSD50K [157]	A	.296	.307	.109	**.317**	.313	.311	.313	.277
	TUT18 [344]	A	.853	.857	.682	**.884**	.881	.849	.867	.801
	VocalImitation [255]	A	.042	.051	.022	.045	.047	.045	.050	**.055**
	VGGSound [82]	AV	.303	.287	—	**.323**	.314	.314	.311	—
NonSem	CREMA-D [75]	A	.514	.489	.354	.528	.540	.520	**.548**	.530
	GTZAN Mus/Sp [520]	A	.954	.931	.866	.946	.891	.931	**.969**	.954
	LibriCount [486]	A	.654	.608	.505	.671	**.706**	.681	.676	.678

Table A.4: **Controlled experiments evaluation results.** All metrics are top-1 accuracy, except FSD50K [157] and VocalImitation [255] (Mean Average Precision). Results in **bold** indicate the highest score, and in gray indicate the lowest. The task types are **Snd/Scn** = Sound/Scene Classification and **NonSem** = Non-Semantic Speech.

Results In total, we trained 8 different model variants and evaluated them on 15 different tasks. Table A.4 shows our main tasks on which we hypothesized improvement (N=8), grouped by modality and task type.

Does dub-augmented pretraining help? For all tasks in Table A.4, one or more dub-augmented models outperform the monolingual **EN** model. In 6/8 tasks, *all* dub-augmented variants outperform **EN**, except for the two easiest tasks (TUT18 and GTZAN). We hypothesized this outcome for the sound and scene classification tasks, where we consistently observe substantial gains, as well as the non-semantic speech tasks.

Is the improvement due only to de-emphasizing speech? We examine the source-separated version to address this question, since it offers the extreme case where the speech is removed altogether (as much as possible). The source-separated variant presents a strong baseline on the sound/scene classification tasks, despite mostly being outperformed by one or more dub-augmented models. We expect this is due to re-focusing on non-speech elements. However, despite strong performance in these cases, this variant has drawbacks. First, it results in lower performance than all other models on VGGSound (audiovisual classification) and both visual tasks (shown in the trade-off results in Table A.5). We suspect this is because there is a clear discrepancy between the auditory and visual channels in the source-separated version, i.e. speech. When a person is speaking, and there is little or no speech content in the auditory stream accompanying the visual, this may act as a confounder for coordinating the

two representations. Note that *People* is a large category in VGGSound[2].

Second, **SEP** significantly underperforms on non-semantic speech tasks and (in Table A.5) language identification, with the exception of GTZAN which we find is an easier task in general. This intuitively makes sense: this variant does not see speech, effectively, and performs lower than the monolingual variant as well. These results illustrate a trade-off: source-separation as a preprocessing method, in addition to being very computationally expensive and weakening the self-supervision assumption (by dependence on a third-party supervised model), results in poor performance on paralinguistic tasks, which require attention to aspects of speech beyond language.

Are more languages better? Given the strength of dub-augmented training, we ask whether introducing more languages into the mix improves performance further. Our results don't indicate this to be the case, but note that in Table A.4, the **EFJ** model is least commonly the lowest-performing dub-augmented variant (1/8 tasks). Additionally, the multilingual variant performs well on 2/3 non-semantic speech tasks. Even though paralinguistic features can vary by language, commonalities exist that may be useful and many practical scenarios could benefit from diverse examples. The robustness of the multilingual model suggests that it could be a reasonable default choice assuming little knowledge about the downstream tasks, and we use a similar multilingual approach in our larger scale experiments in the chapter.

Is dub-augmentation beneficial even without video? The $A_{+\mathbf{EFJ}}$ variant always outperforms the **AUD** model (including on all audio tasks we examine later for trade-offs, shown in Table A.5). **AUD** is the weakest performer on all relevant tasks, indicating the benefits of cross-modal training. Additionally, on some tasks, the multilingual variant comes close to or even outperforms (as in on VocalImitation) the cross-modal variants. Of course, this variant cannot work on visual or multimodal tasks, and still largely underperforms the multimodal dub-augmented models, but it demonstrates the significant value of even unimodal dub-augmented training.

Exploring Trade-Offs Results on the 7 tasks in Table A.5 help us evaluate possible trade-offs in the smaller-scale and controlled setup, to complement the previous results.

Can dub-augmented models still recognize language? The dub-augmented variants

[2] www.robots.ox.ac.uk/ vgg/data/vggsound

generally perform similarly or slightly worse on VoxLingua [524] but appear to do better on VoxForge [324], both language identification tasks. The latter is a large-scale user-submitted dataset, which may have different auditory characteristics from the former as a result. Taking these results together, we expect that the dub-augmented models are able to retain information useful for language identification in their pre-MLP features. It is possible that more general auditory features, which do not encode speech semantics, are still discriminative in these tasks.

Are they discriminative between spoken words? As in our results from the chapter, we do not observe major degradations on linguistic tasks. This suggests that the features learned by our dub-augmented models preserve speech-related information that can be used to, for instance, recognize words or commands. However, the source-separated models' features appear useful for these tasks, which suggests that non-speech features and more general representations of the sound signals may be helpful. We further investigate this below, where our results show that the background noise in one of these datasets (Fluent Speech Commands [316]) may provide useful signal for performance.

Is performance on video-only tasks impacted? On the visual action recognition tasks, the results from the dub-augmented variants appear similar to the baseline. The baseline performs slightly better on HMDB51 and slightly worse on UCF101. This suggests that the overall video-only performance of the model may not be significantly affected by dub-augmented pretraining, similar to what is shown in Table A.2 for our main model variants.

A.2.3 Examples of Synthetic Counterfactual Pairs

Fig. A-13 highlights clips from a synthetically generated version of the LVU dataset [551], which we refer to as **LVU-M**, as noted in the chapter. Similar to Fig. 3-4, the spectrograms show variation and commonalities between alternate audio tracks of the same clip. The examples, arbitrarily selected, show both consistency with the visual (e.g. voices, general timing, etc.) and divergence from it due to artifacts, lack of full acoustic context (e.g. reverberation), and other current limitations of the proposed pipeline. We only show the middle 10 seconds of these clips, to allow easy inspection.

	Task	M	Baselines (SimCLR) AV	SEP	A	Dub-Augmented ES	FR	JA	EFJ	A_{+EFJ}
SemSp	FlSpComm [316]	A	.379	.400	.263	.391	**.410**	.402	.373	.368
	SpComm5h [544]	A	.298	**.372**	.144	.362	.344	.325	.300	.231
	SpCommFull [544]	A	.471	.489	.162	.477	**.537**	.530	.491	.298
Lang	VoxForge [324]	A	.546	.516	.504	.580	.584	**.592**	.571	.543
	VoxLingua10 [524]	A	**.251**	.226	.111	.229	.237	.246	.227	.201
Act	HMDB51 [271]	V	**.341**	.319	–	.330	.324	.322	.333	–
	UCF101 [479]	V	.531	.496	–	.540	.523	.538	**.542**	–

Table A.5: **Controlled experiments potential trade-offs: Does dub-augmentation negatively impact performance on linguistic or vision-only tasks?** The tasks in this table include **Semantic Speech** (FlSpComm [316], SpComm5h [544], and SpCommFull [544]) and **Language ID** (VoxForge [324] and VoxLingua10 [524]), and 2 **Action Recognition** video-only tasks (HMDB51 [271] and UCF101 [479]). The results vary and often reflect relatively small differences in either direction, suggesting overall that performance is not majorly affected on language-focused and vision-only tasks.

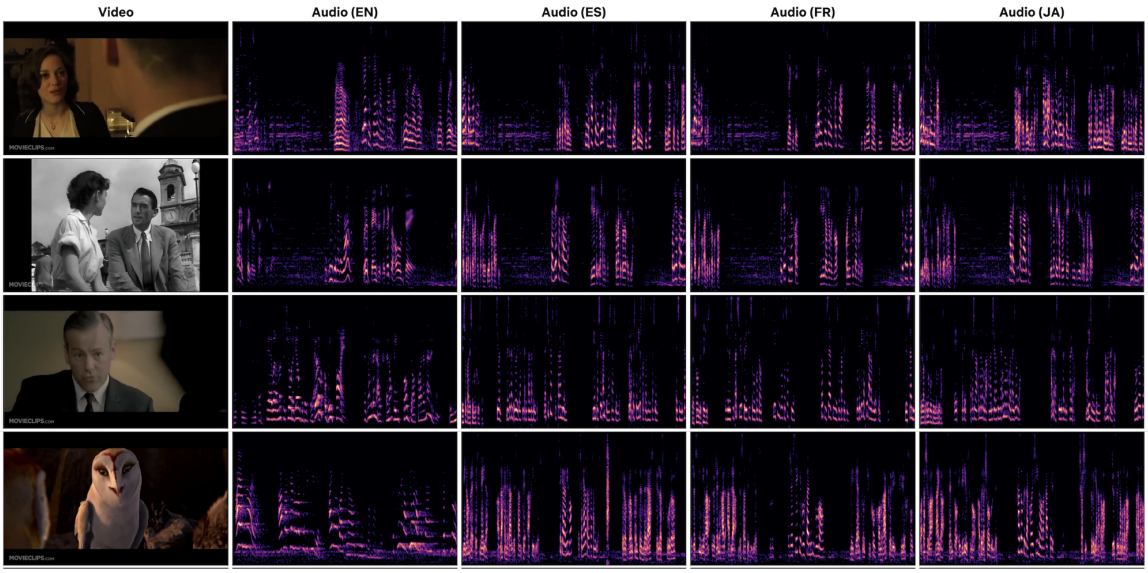

Figure A-13: Examples of clips from **LVU-M**.

A.3 Supplement for Chapter 5

A.3.1 Supplementary Analyses

Generation Time

Iter/Popsize	25	50	100
50	5.49 ± 0.154	9.62 ± 0.452	18.43 ± 0.752
100	10.01 ± 0.194	18.05 ± 0.605	33.40 ± 0.331
300	27.61 ± 0.703	49.94 ± 0.424	97.23 ± 0.469

Table A.6: Time (in seconds) for different population sizes (columns) and iteration counts (rows).

In Table A.6 we illustrate the optimization times, in seconds, for different numbers of iterations (rows) and optimizer population sizes (columns) below, on a modest GPU, i.e. single V100. Note that the necessary number of iterations varies for different prompts, from 50 to 300+ to get optimal results.

CLAP Scores

Model	AudioSet-50	ESC-50
AudioGen	0.249 ± 0.160	0.277 ± 0.180
AudioLDM	0.166 ± 0.128	0.173 ± 0.142
CTAG	**0.573** ± 0.126	**0.585** ± 0.130
Real	–	0.416 ± 0.139

Table A.7: Comparison of CLAP scores between *CTAG* and other generative models on AudioSet-50 and ESC-50 datasets

Table A.7 shows the CLAP [559] evaluations for each model with AudioSet-50 and ESC-50 prompts, as well as for the actual ESC-50 dataset of real sounds. CLAP is the objective that we optimize in our synthesis-by-optimization approach, and these results show how *CTAG* trivially achieves a higher score compared to all other models and even the real data. This highlights the ability of our optimization strategy to effectively maximize the CLAP score, and also the importance of finding alternative and distinct evaluation metrics as we showed in Section 5.3.4.

Dataset	Metric	Model	Sounds	Template	Caption
AudioSet-50	Top-1	*AudioGen*	51.6	**57.0**	48.8
		AudioLDM	17.4	**21.0**	16.6
		CTAG	**26.2**	25.2	23.6
	Top-5	*AudioGen*	77.4	**84.8**	80.8
		AudioLDM	44.2	**49.8**	48.0
		CTAG	45.2	**52.2**	51.6
ESC-50	Top-1	*AudioGen*	54.0	**69.0**	62.0
		AudioLDM	23.0	20.2	**29.4**
		CTAG	**16.4**	11.4	13.8
	Top-5	*AudioGen*	71.8	**85.2**	81.8
		AudioLDM	49.4	47.0	**58.4**
		CTAG	30.4	26.4	**31.0**

Table A.8: Performance comparison, with different prompting strategies, of models on AudioSet-50 and ESC-50 datasets

Prompting Strategies for All Tested Models

For completeness, Table A.8 provides all the results for all different models with templates and captions as we showed for *CTAG* in Section 5.4.3. The performance of *AudioGen* shows a notable boost when using the +T (Template) strategy. However, the impact of these strategies on the other models and datasets is less consistent, with some cases showing modest improvements and others exhibiting a decrease in performance (e.g., *AudioLDM* ESC-50 +T, *AudioLDM* AudioSet-50 +C). Given the variability in results, it is difficult to make a definitive statement about the effectiveness of these strategies across all baselines. While they may prove beneficial in certain scenarios, their impact appears to be context-dependent.

User Study Statistical Models

We report post-hoc contrasts for the user study results in Tables A.9 to A.11.

User Study Per-Prompt Accuracy

Figure A-14 shows the accuracy of our user study participants at classifying sounds generated with *CTAG*, *AudioGen*, and *AudioLDM*. Reviewing these differences shows that some sounds are overall more difficult to identify, for instance; “Truck air brake”. This may be due to the ambiguity in what this can sound like, as it is not as common

contrast	odds.ratio	SE	asymp.LCL	asymp.UCL	z.ratio	p.value
AudioLDM / AudioGen	0.31	0.07	0.19	0.53	-5.28	<1e-04
CTAG / AudioGen	0.85	0.18	0.51	1.42	-0.75	1
CTAG / AudioLDM	2.72	0.59	1.61	4.58	4.59	<1e-04

Table A.9: Post-hoc contrasts from a mixed-effects logistic regression for accuracy.

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
AudioLDM - AudioGen	-0.53	0.12	579	-0.82	-0.24	-4.34	<1e-04
CTAG - AudioGen	-0.48	0.12	579	-0.78	-0.19	-3.97	0.00024
CTAG - AudioLDM	0.04	0.12	579	-0.25	0.34	0.37	1

Table A.10: Post-hoc contrasts from a mixed-effects linear regression for confidence ratings.

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
AudioLDM - AudioGen	0.57	0.12	579	0.29	0.86	4.81	<1e-04
CTAG - AudioGen	1.22	0.12	579	0.93	1.51	10.20	<1e-04
CTAG - AudioLDM	0.65	0.12	579	0.36	0.93	5.39	<1e-04

Table A.11: Post-hoc contrasts from a mixed-effects linear regression for artistic interpretativeness.

a sound as “Bicycle bell”.

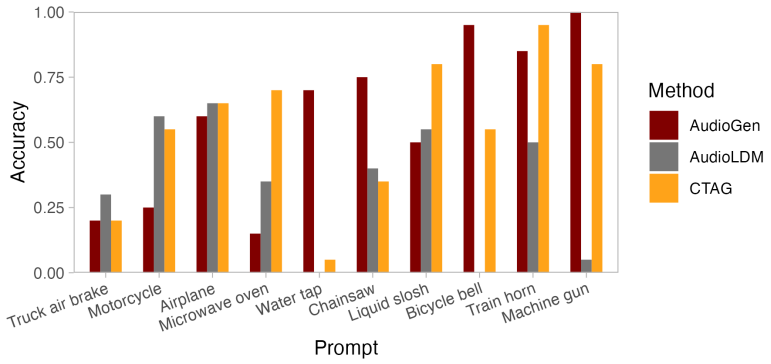

Figure A-14: User study classification accuracy per prompt, for *CTAG*, *AudioGen*, and *AudioLDM*.

Dimensionality Reduction

Having access to the parameters of the synthesizer also allows us to project them into a two-dimensional space to explore the relationship between sounds. Leveraging the Uniform Manifold Approximation and Projection (UMAP) [338] algorithm for dimensionality reduction of the synthesizer parameters, Figure A-15 shows how the representation delineates clusters for each distinct sound class while retaining semantic meaning—sounds with similar acoustic properties cluster together.

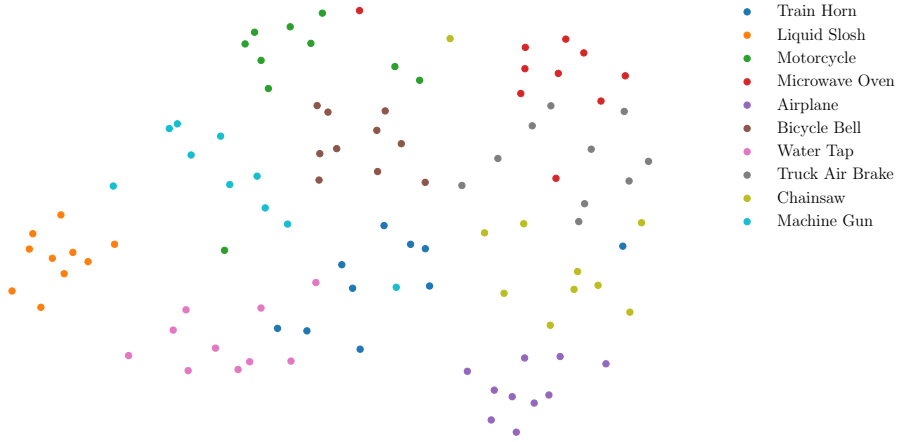

Figure A-15: Dimensionality reduction of the *Voice* synthesizer parameters using UMAP applied to 10 sounds from each of the 10 classes from the user study. It distinctly reveals clusters corresponding to individual sounds, and it shows how conceptually similar sounds such as "water tap" and "liquid slosh" are closer in space.

A.3.2 Caption Prompt

We used the following instructions to generate caption-like prompts from class labels:

> *"Write a simple one-sentence audio caption that describes objectively each sound itself in a real scenario without making up any extra details about other possible sounds or places. You should define the most common action for such an entity when multiple options are available. Avoid using templates such as 'A sound of' or 'The sound of'. Sounds: [List]"*

A.3.3 Listener Survey

In this section, we provide information about the survey design we used to collect human ratings.

Survey Flow

- Standard: Introduction (3 Questions)
- Block: Audio (4 Questions)
- Standard: Additional (2 Questions)

Start of Block: Introduction

Q1: We are conducting a survey to assess the quality of a novel method for text-to-audio generation. You will be presented with a series of short sounds, and asked to select the closest category from a given list, the confidence in your prediction, and how artistically designed the sound is compared to a more realistic interpretation.

Q2: I consent to participate. I understand that my participation is voluntary and I may withdraw my consent at any time.

- Yes (1)
- No (2)

Q3: I am at least 18 years old.

- Yes (1)
- No (2)

Q4: Do you have any hearing loss or hearing difficulties?

- Yes (1)
- No (2)

Q5: Are you fluent in English?

- Yes (1)
- No (2)

Q5: What is your Prolific ID? Please note that this response should auto-fill with the correct ID

Start of Block: Audio

We use Qualtrics' Loop & Merge functionality to loop through the sounds.

A: Select the closest category for the following sound: **[Audio Clip]**

- Truck air brake (1)
- Water tap (2)
- Train horn (3)
- Motorcycle (4)
- Microwave oven (5)
- Liquid slosh (6)
- Chainsaw (7)
- Airplane (8)
- Bicycle bell (9)
- Machine gun (10)

B: How confident are you in your selected answer?

- Completely confident (1)
- Fairly confident (2)
- Somewhat confident (3)
- Slightly confident (4)
- Not confident at all (5)

C: Would you associate this sound more with a realistic portrayal or an artistic interpretation of the category that you selected?

- 1 (1) Realistic Portrayal
- 2 (2) •

- 3 (3) •
- 4 (4) •
- 5 (5) Artistic Interpretation

Start of Block: Additional

We have two questions to check that participants were paying attention.

A1 Please select "Chainsaw" from the options below:

- Truck air brake (1)
- Water tap (2)
- Train horn (3)
- Motorcycle (4)
- Microwave oven (5)
- Liquid slosh (6)
- Chainsaw (7)
- Airplane (8)
- Bicycle bell (9)
- Machine gun (10)

A2: All of the sounds you heard during this survey were the same.

- Yes (1)
- No (2)

Completion Message: Thank you for taking part in this study. Please click the button below to be redirected back to Prolific and register your submission.

A.4 Supplement for Chapter 6

A.4.1 Additional Results

Comparison of Different Architectures Across Tasks

In Figure A-16 we show the relative performance of models trained with data from different synthesizer architectures with $\delta = 0.25$. These results illustrate that, though *Voice*-generated sounds appear strongest overall, there is some task specialization of these different synthesis approaches. For example, on LibriCount and NSynth, *Voice* is the lowest performer here.

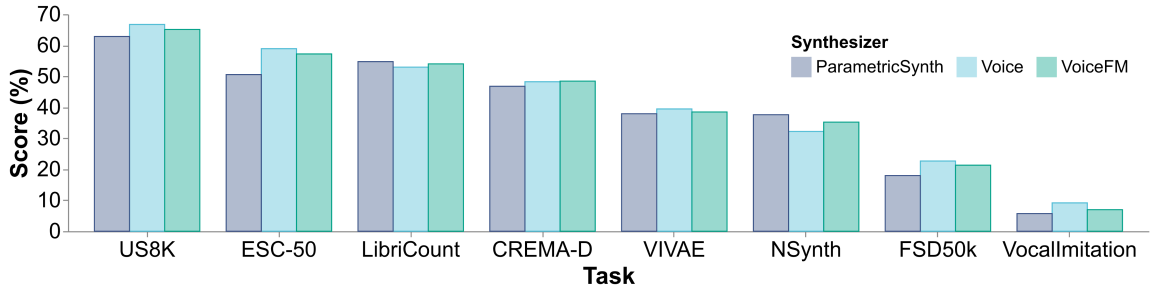

Figure A-16: Scores with a fixed $\delta = 0.25$ and different synthesizer architectures for a suite of tasks including (from left to right) UrbanSound8k [441], ESC-50 [393], LibriCount [486], CREMA-D [75], VIVAE [215], NSynth Pitch 5h [134], FSD50k [157], and Vocal Imitation [255]

Effects of Increasing Perturbation Factor δ on Training

We seek to understand how increasing δ impacts the training dynamics. In particular, the alignment and uniformity objectives are in tension [538]. A small δ leads to easy positive pairs (high similarity), resulting in low alignment cost but potentially poor generalization. Conversely, too large δ produces hard positive pairs (low similarity), increasing the alignment cost but potentially hindering optimization. The optimal δ should balance this trade-off, however the complexity of the synthesizer function and the embedding function make deriving a closed-form solution for this infeasible. As such, we must explore the effect empirically.

Figure A-17 shows the impact of different δ on the final validation value of the alignment and uniformity costs respectively. Alignment cost increases monotonically with δ, which shows the increased difficulty of aligning increasingly distant pairs.

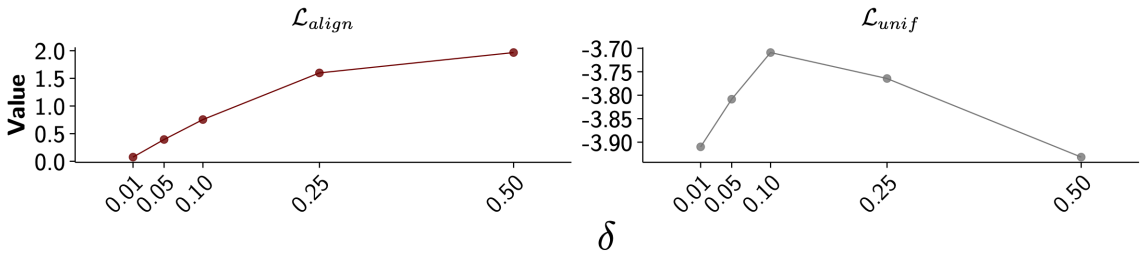

Figure A-17: Final validation scores showing the effect of δ on $\mathcal{L}_{align}$ and $\mathcal{L}_{unif}$. $\mathcal{L}_{align}$ increases monotonically with δ, since the difficulty of aligning more distinct samples goes up. $\mathcal{L}_{unif}$, on the other hand, shows an inverse-U-shaped relationship with δ.

Uniformity has an inverted-U-shaped relationship with δ, suggesting that as the model struggles to align positives with moderate noise driven variation, it incurs a cost in uniformity in order to do so (e.g. creating clusters). With large δ, the amount of noise present is significant, alignment is difficult, and the representations can be more spread out. The theoretically optimal value of L_{unif} is $-2t = 4$, which all values of δ remain close to. In Figure 8 of [538], the best performance with a more complex task and encoder (review classification) is observed when alignment is on the higher side (but not the maximum), and uniform is low (close to the optimal value). In our experiments, $\delta = 0.25$ gets closest to this, and we observe it to be the strongest as well.

Results for all Variants

We give results for all synthetic model variants below, in Table A.12.

A.4.2 Additional Details on Training

Augmentation Batching

Due to practical considerations in batching and memory management, augmentations are applied differently for real and synthetic data. In real data, augmentations are applied per-example within distributed data-loading workers. Synthetic data is batch-generated within the main process to avoid concurrency issues between JAX's multithreading and PyTorch's data loading. Individually augmenting examples in this synthetic data environment is prohibitively slow. As a solution, we mini-batched augmentations with a default size ≤ 100. This allows us to memory-efficiently

Data/Model	ESC	US8K	VIV	NSyn	C-D	FSD	VI	LCount
External Baselines								
HEAR/ARCH Top	96.65	79.09	44.28	87.80	75.21	65.48	22.69	78.53
HEAR/ARCH SSL	80.50	79.09	44.28	52.40	75.21	50.88	18.48	78.53
MS-CLAP Linear	89.95	82.29	–	–	23.15	50.24	–	54.51
GURA (HEAR)	74.35	–	–	38.20	75.21	41.32	18.48	68.34
VGGSound Sup.	87.45	77.57	39.38	43.80	54.36	43.76	14.06	56.10
Internal Baselines								
Random Init.	22.45	55.03	33.81	36.20	38.91	9.03	2.43	44.91
Voice (Ours, No-δ, Aug.)	48.65	59.46	36.31	32.80	46.32	16.88	7.12	47.64
VGGSound SSL (Aug.)	48.85	61.91	32.67	39.60	47.86	19.63	6.03	53.46
VGGSound SSL (Jitter)	52.95	63.82	38.12	14.20	**50.03**	24.02	3.43	**69.77**
VGGSound-Mix 5s	43.95	59.69	33.31	40.80	46.10	14.71	5.95	52.57
VGGSound-Mix 10s	42.95	57.40	32.03	40.20	46.57	15.77	6.43	51.07
Audio Doppelgängers (Ours)								
Best Synthetic	**58.90**	**66.71**	**39.45**	**44.40**	**48.43**	**24.12**	**9.15**	**58.60**
Voice ($\delta = 0.01$)	47.55	59.56	38.62	11.40	47.53	17.15	6.67	55.56
Voice ($\delta = 0.05$)	47.90	64.02	37.93	13.80	46.45	17.77	7.72	51.52
Voice ($\delta = 0.10$)	48.40	63.92	38.74	11.40	45.13	18.40	7.67	49.32
Voice ($\delta = 0.25$)	**58.90**	**66.71**	**39.45**	32.20	**48.24**	**24.12**	**9.15**	52.95
Voice ($\delta = 0.50$)	41.85	54.03	28.54	40.60	45.78	17.14	4.69	43.85
VoiceFM ($\delta = 0.01$)	42.40	59.89	36.58	9.20	44.31	15.34	5.15	57.13
VoiceFM ($\delta = 0.05$)	42.90	62.96	36.54	14.20	44.93	15.64	5.79	50.61
VoiceFM ($\delta = 0.10$)	44.80	62.03	35.73	14.80	43.99	15.67	5.60	50.56
VoiceFM ($\delta = 0.25$)	57.20	65.11	38.48	35.20	48.43	22.15	6.96	54.00
VoiceFM ($\delta = 0.50$)	43.50	60.98	39.04	12.20	44.17	15.25	6.06	51.07
Parametric ($\delta = 0.01$)	39.50	58.95	36.87	12.20	42.16	13.92	4.53	**58.60**
Parametric ($\delta = 0.05$)	40.15	57.22	35.11	14.60	42.65	12.87	4.78	55.37
Parametric ($\delta = 0.10$)	42.50	59.65	34.12	14.20	43.01	13.41	4.97	53.43
Parametric ($\delta = 0.25$)	50.55	62.83	37.91	37.60	46.77	18.68	5.70	54.72
Parametric ($\delta = 0.50$)	41.15	56.86	35.41	10.40	41.73	12.76	4.48	54.27
Voice ($\delta = 0.01$, Aug.)	52.55	62.92	34.82	23.60	46.96	18.18	8.17	51.01
Voice ($\delta = 0.05$, Aug.)	53.00	65.17	34.49	19.40	45.39	19.79	8.32	49.84
Voice ($\delta = 0.10$, Aug.)	54.20	65.89	33.78	23.40	45.71	20.38	8.50	50.42
Voice ($\delta = 0.25$, Aug.)	58.75	65.01	34.81	**44.40**	46.17	21.76	8.54	50.70
Voice ($\delta = 0.50$, Aug.)	32.25	48.40	25.41	36.20	41.38	11.82	3.26	44.74

Table A.12: Complete results for all model variants.

leverage GPU processing and introduces variation within each training batch. While per-example augmentations might further enhance performance of synthetic data

with augmentations, we believe our current approach is a conservative yet effective option and expect minimal impact.

A.5 Supplement for Chapter 8

A.5.1 Questions

ID	QUESTIONS
6.1	I quickly figured out how to use Editor-Red
6.2	It was easy to come up with ideas while writing
6.3	It was easy to decide how I will continue this story
6.4	The more time I spend writing with Editor-Red, the better it gets.
6.5	The pictures used in Editor-Red distracted me from my task
6.6	The pictures used in Editor-red were helpful
6.7	The sounds used in Editor-Red distracted me from my task
6.8	The sounds used in Editor-Red were helpful
6.9	The story that I wrote in Editor-Red is coherent
6.10	The story that I wrote in Editor-Red is creative
6.11	Using Editor-Red felt intuitive
6.12	Using Editor-Red was easy
14.1	I did most of the creative writing, using Editor-Red just for suggestions.
14.2	I enjoyed co-writing with Editor-Red
14.3	I enjoyed collaborating with Editor-Red
14.4	I equally used textual suggestions and pictures and sounds
14.5	I mostly used the textual suggestions and not pictures or sounds
14.6	The final product of writing is a result of joint efforts of Editor-Red and myself
14.7	The suggestions made by Editor-Red were coherent
14.8	The suggestions made by Editor-Red were creative
14.9	The suggestions made by Editor-Red were grammatically correct
14.10	The suggestions made by Editor-Red were relevant
14.11	The suggestions made by Editor-Red were surprising
16.1	It was easy to come up with ideas while writing
16.2	It was easy to decide how I will continue this story
16.3	The story that I wrote in Editor-Green is coherent

16.4	The story that I wrote in Editor-Green is creative
16.5	Using Editor-Green felt intuitive
16.6	Using Editor-Green was easy

21	In which editor was the text that you wrote more creative?
22	In which editor was the text that you wrote more coherent?
23	Where did you feel more relaxed when writing a story?
24	Where was it easier to write a text?
25	Which editor did you prefer for writing a creative text?
26	Where did you feel more focused when writing a text?
27	Where did it feel more demanding when writing a text?
28	Where did it feel more rushed when writing a text?
29	Where did you feel you had to work harder when writing a text?
30	Which editor made you feel more discouraged or annoyed when writing a text?

A.6 Supplement for Chapter 9

A.6.1 Surveys

Draft+Revise Survey

Please rate the following factors (Very Low to Very High) based on your experience with the Figure Description Writing Assistant.

- Mental Demand: How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
- Temporal Demand: How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
- Own Performance: How successful were you in performing the task? How satisfied were you with your performance?
- Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?
- Frustration Level: How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- The tool helped me produce alt text more efficiently.
- The tool helped me think to describe figure elements I would not have thought to describe otherwise.
- The tool helped me produce better alt text.

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- The draft alt texts were helpful
- The generated draft for the summarized figure description was helpful

Please explain your ratings for each of the above statements.

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- I would use the tool if it were available.
- I would recommend the tool to my friends and colleagues.
- I found the tool to be helpful.
- I found the tool to be able to improve my productivity.
- I found the tool to be annoying or distracting.

Interactive Assistance Survey

Please rate the following factors (Very Low to Very High) based on your experience with the Figure Description Writing Assistant.

- Mental Demand: How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
- Temporal Demand: How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?

- Own Performance: How successful were you in performing the task? How satisfied were you with your performance?
- Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?
- Frustration Level: How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- The tool helped me produce alt text more efficiently.
- The tool helped me think to describe figure elements I would not have thought to describe otherwise.
- The tool helped me produce better alt text.

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- The draft alt texts were helpful
- The Potential User Question type suggestions were helpful
- The Generate at Cursor type suggestions were helpful
- The generated draft for the summarized figure description was helpful

Please explain your ratings for each of the above statements.

Please indicate your level of agreement with the following statements based on your experience with the Figure Description Writing Assistant.

- I would use the tool if it were available.
- I would recommend the tool to my friends and colleagues.
- I found the tool to be helpful.
- I found the tool to be able to improve my productivity.
- I found the tool to be annoying or distracting.

Comparison Survey

Please reflect back on both interfaces.

What aspects of each interface did you like, and why?

Please explain any situations where the tool was especially helpful:
For example, if suggestions drew your attention to specific visual elements of the figure or ways to describe them, or provided text that did so which you were able to incorporate directly.

Do you have any other feedback about problems, bugs, or areas for improvement with regard to the interfaces?

Please explain any situations where the tool was unhelpful or detrimental:
Can you provide an example where a suggestion was unhelpful or misleading? If so, why?
Did any suggestions make your alt text worse in a significant way? Please explain.

What changes to each tool would make it more helpful?

Anything else that you would like to share with us?

Which version of the system did you prefer? (Without vs. With Suggestions)

A.6.2 Prompt Design

The overall prompt structure is given as follows:

- **Instruction Prompt**
- **Metadata Prompt**
- **Description Content**

Instruction Prompt

We defined several different versions of the instruction prompt, toward different goals. The first two below form part of the *Generate at Cursor* feature, while the third is used for pre-generating drafts.

Initial High-Level Summary `Your goal is to assist in writing an alt text description of a figure that is as informative and accessible as possible, based on metadata provided to you.`
`Some of this data is automatically extracted from the figure, and may contain errors. Infer as much detail as possible from the information given.`
`Respond with only a brief and high-level overview (1-2 sentences), with no additional content. In your response, do not explicitly refer to the metadata (such as "caption" or "OCR text"). These are provided to help you write descriptive responses only.`

Text Continuation and Infilling `Your goal is to assist in writing an alt text description of a figure that is as informative and accessible as possible, based on metadata provided to you.`
`Some of this data is automatically extracted from the figure, and may contain errors. Infer as much detail as possible from the information given. Only include clear and helpful statements for understanding the figure. Do not make explicit reference to the metadata (such as "caption" or "OCR text"). These are provided to help you write descriptive responses only.`
`Respond with only a continuation of the given description itself (1-4 sentences), with no additional content. Add as much detail as possible. You may also be given a DESCRIPTION CONTEXT, which contains text after your response. In this case, provide text that bridges the gap between the description, and additional text the user has already written.`
`In your response, do not explicitly refer to the metadata (such as "caption" or "OCR text"). These are provided to help you write descriptive responses only.`

Full Draft `Your goal is to assist in writing an alt text description of a figure that is as informative and accessible as possible, based on metadata provided to you.`
`Some of this data is automatically extracted from the figure, and may contain errors. Infer as much detail as possible from the information given.`
`Respond with a full description of the figure, with no additional content. In your response, do not explicitly refer to the metadata (such as "caption" or "OCR text"). These are provided to help you write descriptive responses only.`

Potential User Questions `Your goal is to assist in writing an alt text description of a figure that is as informative and accessible as possible. Infer as much detail as possible from the information given.`

```
What visual aspects of the figure are unclear from the given alt text description?  Ask
a series of questions to elicit all the necessary information about the figure to describe
these elements.  Based on the type of figure, focus on essential visual aspects that
someone who cannot see the figure would need to know.  Based on the guidelines and metadata
you have access to, suggest an answer for each question.  In your response, do not explicitly
refer to the metadata (such as "caption" or "OCR text").  These are provided to help
you write descriptive responses only.  Do not repeat any existing questions.
```

Metadata Prompt

We define the **Metadata Prompt** as:

```
---CAPTION
        <Caption Text>

---FIGURE MENTIONS FROM PAPER
        <Mentioning Paragraphs>

---OCR TEXT RECOGNIZED FROM FIGURE (MAY CONTAIN ERRORS)
        <Layout Preserving OCR Text>

---DATA TABLE EXTRACTED FROM FIGURE (MAY CONTAIN ERRORS)
        <Automatically Extracted Data Table>

---Please refer to the following guidelines when writing
   your description:
        <Selected Guidelines>
---
```

A.6.3 Event Traces

Fig. A-18 shows event traces for all logged participants in our study (i.e. P6 through P14). We provide them here to give a broader sense of the diversity of strategies we observed.

A.6.4 Additional Interface Features

Authors can selectively ablate certain metadata they deem irrelevant or erroneous via interface settings (Fig. A-19A). Also present in this menu is a set of guidelines which our pipeline selects based on the figure type (expanded in Fig. A-19B), incorporating

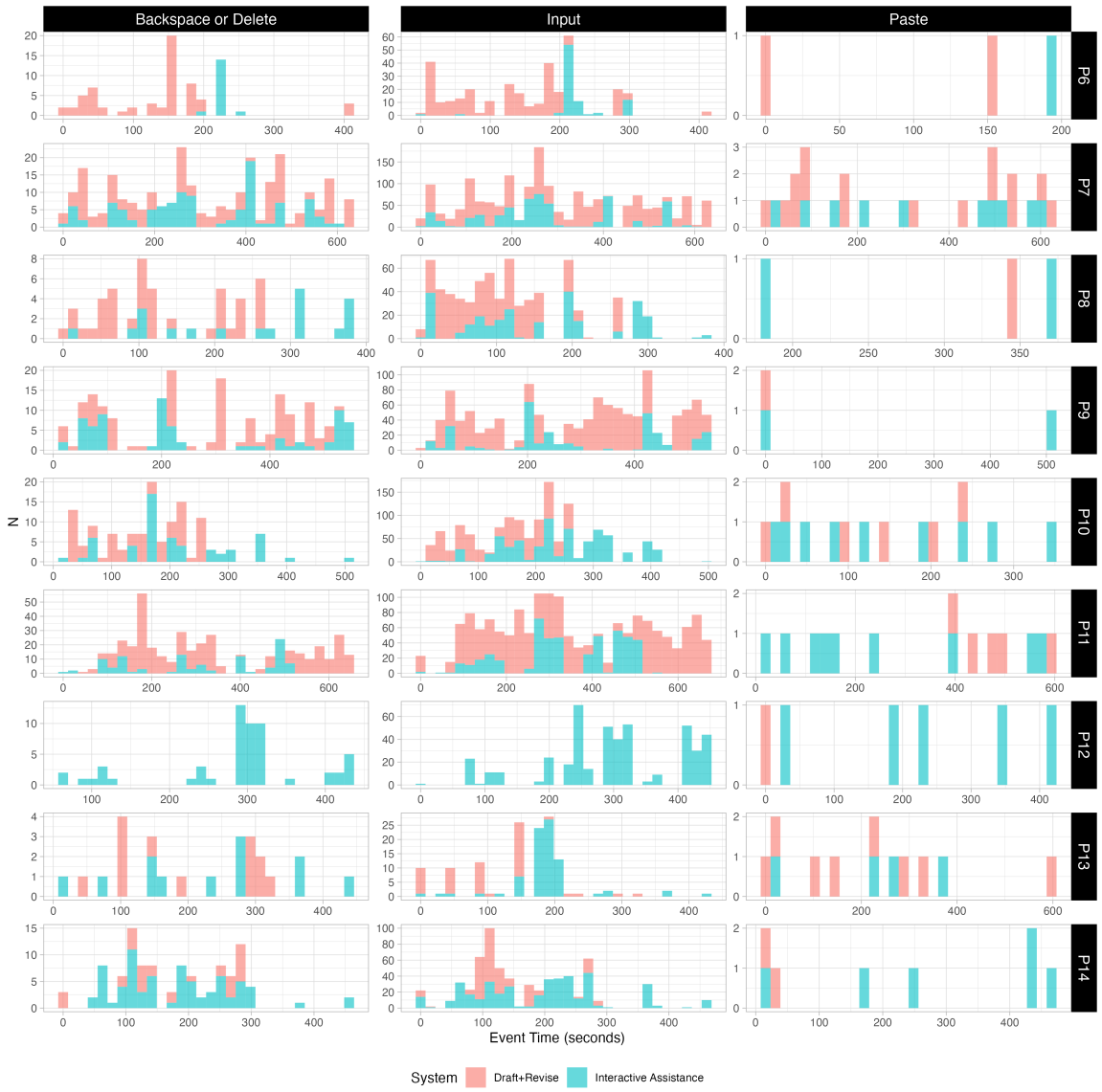

Figure A-18: Event traces for all logged participants (N=9) in our study. Different patterns show a wide range of strategies for using our systems' features to produce detailed alt text.

general figure description guidelines along with domain-specific items (e.g. for general plots), and figure type-specific ones as well (e.g. describing the change of concentration of datapoints for a scatter plot). The summarization workflow is shown in Fig. A-19C).

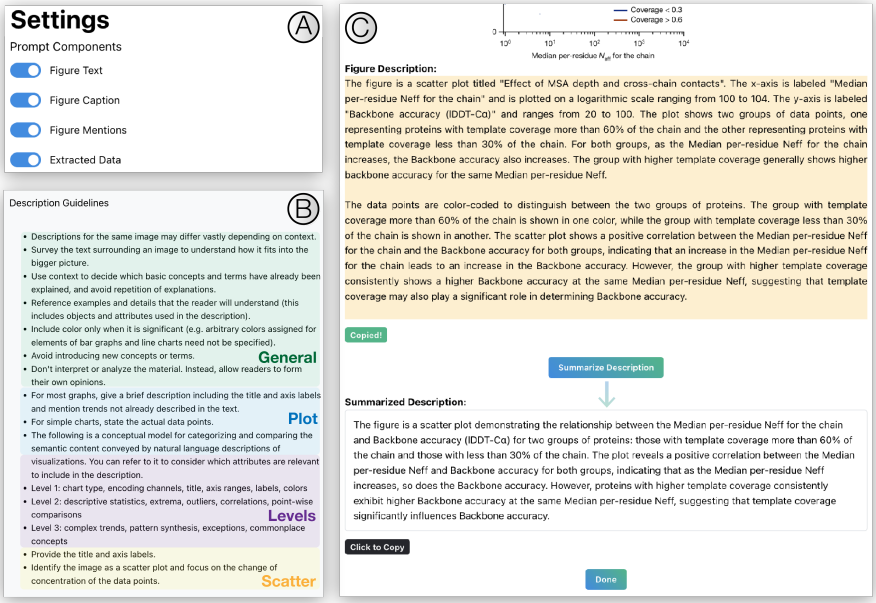

Figure A-19: Additional features that our system versions implement. **(A)** Prompt ablation settings (in **Interactive Assistance**), wherein the user can de-select metadata components for use in suggestion and question generation, to account for highly erroneous extractions or irrelevant information. **(B)** Figure description guidelines (both versions). These begin with general guidelines for descriptions, then plot-specific guidelines, then the semantic level framework introduced by Lundgard and Satyanarayanan [317] for data visualizations, then scatterplot-specific items, to construct a full set of guidelines for both prompting and user review. A link to the DIAGRAM Center's original guidelines is also provided. **(C)** After writing the full description, we implement a summarization workflow to produce more concise descriptions (both versions; one paragraph long by default). This also serves as a description review stage.

References

[1] FMA: A Dataset For Music Analysis. 2017. Meeting Name: 18th International Society for Music Information Retrieval Conference.

[2] Re-imagining the opera of the future, September 2023. URL https://news.mit.edu/2023/re-imagining-opera-of-the-future-valis-0927.

[3] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[4] Haider Al-Tahan and Yalda Mohsenzadeh. Clar: Contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics*, pages 2530–2538. PMLR, 2021.

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[6] Abdullah Almaatouq, Thomas L Griffiths, Jordan W Suchow, Mark E Whiting, James Evans, and Duncan J Watts. Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences*, 47:e33, 2024.

[7] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-Supervised Learning by Cross-Modal Audio-Video Clustering. In *Advances in Neural Information Processing Systems*, volume 33, pages 9758–9770, 2020.

[8] Teresa M. Amabile. *Creativity In Context: Update To The Social Psychology Of Creativity*. Routledge, New York, June 2019. ISBN 978-0-429-50123-4. doi: 10.4324/9780429501234.

[9] Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara Martin, and Mark Riedl. Guided neural language generation for automated storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 46–55, 2019.

[10] Ishwarya Ananthabhotla, David B Ramsay, and Joseph A Paradiso. HCU400: An annotated dataset for exploring aural phenomenology through causal uncertainty. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE, 2019.

[11] Rolf Anderegg, Norbert Felber, Wolfgang Fichtner, and Ulrich Franke. Implementation of High-Order Convolution Algorithms with Low Latency on Silicon Chips. In *Audio Engineering Society Convention 117*. Audio Engineering Society, 2004.

[12] Manuel Anglada-Tort, Peter M. C. Harrison, Harin Lee, and Nori Jacoby. Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution. *Current Biology*, March 2023. ISSN 0960-9822. doi: 10.1016/j.cub.2023.02.070. URL https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3501820. Publisher: Cell Press.

[13] Andrey Anikin. Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior research methods*, 51:778–792, 2019.

[14] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[15] Ian Arawjo. Kicking the Leg out from the Table: On Contrived Controls in HCI Systems Research. *Medium*, October 4 2023. https://ianarawjo.medium.com/kicking-the-leg-out-from-the-table-on-contrived-controls-in-hci-systems-research-9f77b28fef39.

[16] Kenneth C Arnold, Krzysztof Z Gajos, and Adam T Kalai. On suggesting phrases vs. predicting words for mobile text composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 603–608, 2016.

[17] Kenneth C Arnold, Kai-Wei Chang, and Adam T Kalai. Counterfactual language model adaptation for suggesting phrases. *arXiv preprint arXiv:1710.01799*, 2017.

[18] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 128–138, 2020.

[19] Bishnu S Atal, Jih Jie Chang, Max V Mathews, and John W Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, 63(5):1535–1555, 1978.

[20] Jacques Attali. Noise, 1977. URL https://www.upress.umn.edu/book-division/books/noise.

[21] Dana Aubakirova, Kim Gerdes, and Lufei Liu. PatFig: Generating Short and Long Captions for Patent Figures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2843–2849, 2023.

[22] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. pages 892–900, 2016.

[23] Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem. URL http://github.com/deepmind, 2020.

[24] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

[25] Alan D Baddeley and Graham Hitch. Working memory. In *Psychology of learning and motivation*, volume 8, pages 47–89. Elsevier, 1974.

[26] Roland Badeau. Common mathematical framework for stochastic reverberation models. *The Journal of the Acoustical Society of America*, 145(4):2733–2745, 2019.

[27] Eytan Bakshy, Lili Dworkin, Brian Karrer, Konstantin Kashin, Benjamin Letham, Ashwin Murthy, and Shaun Singh. AE: A domain-agnostic platform for adaptive experimentation. In *Conference on neural information processing systems*, pages 1–8, 2018.

[28] JA Ballas and MJ Sliwinski. Causal uncertainty in the identification of environmental sounds. *Georgetown University, Washington, DC*, 1986.

[29] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine*

intelligence, 41(2):423–443, 2018.

[30] David Bamman, Brendan O'Connor, and Noah A Smith. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, 2013.

[31] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.

[32] Manel Baradad, Richard Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, and Phillip Isola. Procedural image programs for representation learning. *Advances in Neural Information Processing Systems*, 35:6450–6462, 2022.

[33] Manel Baradad Jurjo, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. *Advances in Neural Information Processing Systems*, 34:2556–2569, 2021.

[34] Jeffrey Bardzell and Shaowen Bardzell. Humanistic Hci. *Interactions*, 23(2):20–29, 2016.

[35] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *arXiv preprint arXiv:2308.11596*, 2023.

[36] Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *arXiv preprint arXiv:2305.10951*, 2023.

[37] Mark Batey. The measurement of creativity: From definitional consensus to the introduction of a new heuristic framework. *Creativity Research Journal*, 24(1):55–65, 2012.

[38] Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 2023. ISSN 0016-0032. doi: https://doi.org/10.1016/j.jfranklin.2023.11.038. URL

https://www.sciencedirect.com/science/article/pii/S0016003223007536.

[39] Dan Bennett, Alan Dix, Parisa Eslambolchilar, Feng Feng, Tom Froese, Vassilis Kostakos, Sebastien Lerique, and Niels Van Berkel. Emergent interaction: Complexity, dynamics, and enaction in HCI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[40] Eden Bensaid. Multimodal generative models for storytelling. Master's thesis, Massachusetts Institute of Technology, 2021.

[41] Eden Bensaid, Mauro Martino, Benjamin Hoover, Jacob Andreas, and Hendrik Strobelt. FairyTailor: A Multimodal Generative Framework for Storytelling. *arXiv preprint arXiv:2108.04324*, 2021.

[42] Paul F. Berliner. *Thinking in Jazz: The Infinite Art of Improvisation*. University of Chicago Press, October 2009. ISBN 978-0-226-04452-1. Google-Books-ID: tqPnM_e4CPYC.

[43] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. pages 591–596, 2011. doi: 10.7916/D8NZ8J07. URL https://doi.org/10.7916/D8NZ8J07.

[44] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020.

[45] Xiaojun Bi, Tom Ouyang, and Shumin Zhai. Both complete and correct? multi-objective optimization of touchscreen keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2297–2306, 2014.

[46] Timothy W. Bickmore and Rosalind W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12 (2):293–327, June 2005. ISSN 1073-0516, 1557-7325. doi: 10.1145/1067860.1067867. URL https://dl.acm.org/doi/10.1145/1067860.1067867.

[47] Lukas Bieringer, Kathrin Grosse, Michael Backes, and Katharina Krombholz. Mental Models of Adversarial Machine Learning. *arXiv preprint arXiv:2105.03726*, 2021.

[48] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C

Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[49] Peter Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8(4):e60603, 2013.

[50] Margaret A. Boden. *The creative mind: Myths & mechanisms*. The creative mind: Myths & mechanisms. Basic Books, New York, NY, US, 1991. ISBN 978-0-465-01452-1. Pages: xii, 303.

[51] Margaret A Boden et al. *The creative mind: Myths and mechanisms*. Psychology Press, 2004.

[52] Konrad Boehmer. Chance as Ideology. In Julia Robinson, editor, *John Cage (October Files #12)*, pages 17–34. 2011. URL https://mitpressbookstore.mit.edu/book/9780262516303. ISBN: 9780262516303.

[53] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Ricardo Zapata, and Xavier Serra. Essentia: An Audio Analysis Library for Music Information Retrieval. In *International Society for Music Information Retrieval Conference*, 2013. URL https://api.semanticscholar.org/CorpusID:11200511.

[54] Tal Boger, Ishwarya Ananthabhotla, and Joseph Paradiso. Manipulating causal uncertainty in sound objects. In *Proceedings of the 16th International Audio Mostly Conference*, pages 9–15, 2021.

[55] Marilyn G. Boltz. Musical Soundtracks as a Schematic Influence on the Cognitive Processing of Filmed Events. *Music Perception*, 18(4):427–454, June 2001. ISSN 0730-7829. doi: 10.1525/mp.2001.18.4.427. URL https://doi.org/10.1525/mp.2001.18.4.427.

[56] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

[57] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[58] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[59] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. JAX: composable transformations of Python+ NumPy programs, 2018.

[60] Gwern Branwen. GPT-3 creative fiction. 2020.

[61] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[62] Jacob Browning and Yann LeCun. AI And The Limits Of Language. August 2022. URL https://www.noemamag.com/ai-and-the-limits-of-language.

[63] Nicholas J Bryan. Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020.

[64] Erik Brynjolfsson. The turing trap: The promise & peril of human-like artificial intelligence. *Daedalus*, 151(2):272–287, 2022.

[65] Sture Brändström. Music Teachers' Everyday Conceptions of Musicality. *Bulletin of the Council for Research in Music Education*, (141):21–25, 1999. ISSN 0010-9894. URL https://www.jstor.org/stable/40318978. Publisher: University of Illinois Press.

[66] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, January 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html. ISSN: 2640-3498.

[67] Richard R. Burton and John Seely Brown. An investigation of computer coaching

for informal learning activities. *International Journal of Man-Machine Studies*, 11(1): 5–24, January 1979. ISSN 0020-7373. doi: 10.1016/S0020-7373(79)80003-6. URL https://www.sciencedirect.com/science/article/pii/S0020737379800036.

[68] Daniel Buschek, Martin Zürn, and Malin Eiband. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[69] Vannevar Bush. As we may think. *The atlantic monthly*, 176(1):101–108, 1945.

[70] William Byrne, Peter Beyerlein, Juan M Huerta, Sanjeev Khudanpur, Bhaskara Marthi, John Morgan, Nino Peterek, Joe Picone, Dimitra Vergyri, and T Wang. Towards language independent acoustic modeling. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 2, pages II1029–II1032. IEEE, 2000.

[71] Antoine Caillon and Philippe Esling. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis, December 2021. URL http://arxiv.org/abs/2111.05011. arXiv:2111.05011 [cs, eess].

[72] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. How Novelists Use Generative Language Models: An Exploratory User Study. In *HAI-GEN+ user2agent@ IUI*, 2020.

[73] Mateo Cámara, Zhiyuan Xu, Yisu Zong, José Luis Blanco, and Joshua D Reiss. Optimization Techniques for a Physical Model of Human Vocalisation. *arXiv preprint arXiv:2309.14761*, 2023.

[74] Mateo Cámara, Fernando Marcos, and José Luis Blanco. Decoding Vocal Articulations from Acoustic Latent Representations. *arXiv preprint arXiv:2406.14379*, 2024.

[75] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[76] William E. Caplin. *Classical Form: A Theory of Formal Functions for the Instrumental Music of Haydn, Mozart, and Beethoven*. Oxford University Press, USA, 1998. ISBN 978-0-19-514399-7. Google-Books-ID: PBYjpKypCskC.

[77] Orson Scott Card. *How to write science fiction and fantasy*. Writer's digest books Cincinnati, OH, 1990.

[78] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1213, 2021.

[79] Frederic Chaume. The turn of audiovisual translation: New audiences and new technologies. *Translation spaces*, 2(1):105–123, 2013.

[80] Frederic Chaume. *Audiovisual translation: dubbing*. Routledge, 2020.

[81] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual Acoustic Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022.

[82] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[83] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022.

[84] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep Cross-Modal Audio-Visual Generation. *CoRR*, abs/1704.08292, 2017. URL http://arxiv.org/abs/1704.08292.

[85] Shixing Chen, Chun-Hao Liu, Xiang Hao, Xiaohan Nie, Maxim Arap, and Raffay Hamid. Movies2Scenes: Using movie metadata to learn scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6535–6544, 2023.

[86] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[87] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation

from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1841–1850, 2019.

[88] Zhehuai Chen, Andrew Rosenberg, Yu Zhang, Gary Wang, Bhuvana Ramabhadran, and Pedro J Moreno. Improving Speech Recognition Using GAN-Based Speech Synthesis and Contrastive Unspoken Text Selection. In *Interspeech*, pages 556–560, 2020.

[89] Manuel Cherep* and Nikhil Singh*. Synthax: A fast modular synthesizer in jax. In *Audio Engineering Society Convention 155*. Audio Engineering Society, 2023.

[90] Manuel Cherep* and Nikhil Singh*. Contrastive learning from synthetic audio doppelgangers. *arXiv preprint arXiv:2406.05923*, 2024.

[91] Manuel Cherep*, Nikhil Singh*, and Jessica Shand. Creative text-to-audio generation via synthesizer programming. In *ICML*, 2024.

[92] Michelene TH Chi. Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in instructional psychology*, 5: 161–238, 2000.

[93] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350. IEEE, 2021.

[94] Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. A dataset of alt texts from HCI publications: Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–12, 2022.

[95] Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K Anumanchipalli. Evidence of vocal tract articulation in self-supervised learning of speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[96] Hwan-Hee Choi, Jeroen JG Van Merriënboer, and Fred Paas. Effects of the physical environment on cognitive load and learning: towards a new model of cognitive load. *Educational Psychology Review*, 26(2):225–244, 2014.

[97] Yejin Choi. The Curious Case of Commonsense Intelligence. *Daedalus*, 151(2):139–155, May 2022. ISSN 0011-5266. doi: 10.1162/daed_a_01906. URL https://doi.org/10.1162/daed_a_01906.

[98] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[99] Rudi Cilibrasi and Paul MB Vitányi. Clustering by compression. *IEEE Transactions on Information theory*, 51(4):1523–1545, 2005.

[100] Christopher Clark and Santosh Divvala. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 143–152, 2016.

[101] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, pages 329–340, 2018.

[102] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. Wordcraft: a Human-AI Collaborative Editor for Story Writing. *arXiv preprint arXiv:2107.07430*, 2021.

[103] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*, 2020.

[104] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[105] Crispin Cooper, Damian Murphy, David Howard, and Alexander Tyrrell. Singing synthesis with an evolved physical model. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1454–1461, 2006.

[106] David Cope. Computer Modeling of Musical Intelligence in EMI. *Computer Music Journal*, 16(2):69–83, 1992. ISSN 0148-9267. doi: 10.2307/3680717. URL https://www.jstor.org/stable/3680717. Publisher: The MIT Press.

[107] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi

Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[108] Fintan J Costello and Mark T Keane. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science*, 24(2):299–349, 2000.

[109] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022.

[110] Mihaly Csikszentmihalyi. *Creativity: Flow and the psychology of discovery and invention*. Creativity: Flow and the psychology of discovery and invention. HarperCollins Publishers, New York, NY, US, 1997. ISBN 978-0-06-017133-9. Pages: viii, 456.

[111] Mihaly Csikszentmihalyi. *Flow: The psychology of happiness*. Random House, 2013. URL https://scholar.google.com/scholar?cluster=15319637431267571346&hl=en&oi=scholarr.

[112] Barbara Dancygier. What can blending do for you? *Language and Literature*, 15(1): 5–15, 2006.

[113] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–13, 2022.

[114] Jianwu Dang and Kiyoshi Honda. Estimation of vocal tract shapes from speech sounds with a physiological articulatory model. *Journal of Phonetics*, 30(3):511–532, 2002.

[115] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, August 2015. ISSN 0001-0782. doi: 10.1145/2701413. URL https://dl.acm.org/doi/10.1145/2701413.

[116] Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. Technical report, Technical report, Stanford University, 2008.

[117] Dani Deahl. This live stream plays endless death metal produced by an AI, April 2019.

URL https://www.theverge.com/2019/4/27/18518170/algorithm-ai-death-metal-dadabots-live-stream-youtube-cj-carr-zack-zukowski.

[118] Leslie A DeChurch and Jessica R Mesmer-Magnus. The cognitive underpinnings of effective teamwork: a meta-analysis. *Journal of applied psychology*, 95(1):32, 2010.

[119] DeepLearningAI. A Chat with Andrew on MLOps: From Model-centric to Data-centric AI. URL https://www.youtube.com/watch?v=06-AZXmwHjo.

[120] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *arXiv preprint arXiv:1909.01174*, 2019.

[121] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. " Algorithms ruin everything" # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3163–3174, 2017.

[122] Michael A DeVito, Jeffrey T Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. The algorithm and the user: How can hci use lay understandings of algorithmic systems? In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018.

[123] Terrance DeVries, Adriana Romero, Luis Pineda, Graham W. Taylor, and Michal Drozdzal. On the Evaluation of Conditional {GAN}s, 2020. URL https://openreview.net/forum?id=rylxpA4YwH.

[124] Ali Diba, Vivek Sharma, Mohammad Arzani, Luc Van Gool, et al. Spatio-Temporal Convolution-Attention Video Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 859–869, 2023.

[125] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9726, 2022.

[126] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[127] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

[128] Stephen R Donaldson. *The gap into conflict: The real story*. CNIB, 2008.

[129] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[130] Gayle T Dow. Defining creativity. In *Creativity and Innovation*, pages 5–21. Routledge, 2022.

[131] August DuMont Schütte, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick Schwab. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ digital medicine*, 4(1):141, 2021.

[132] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[133] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[134] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017.

[135] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. GANSynth: Adversarial Neural Audio Synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1xQVn09FX.

[136] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint*

arXiv:1902.08710, 2019.

[137] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.

[138] Douglas C Engelbart. *Augmenting human intellect: A conceptual framework*. 1962.

[139] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*, pages 2371–2382, 2016.

[140] Philippe Esling, Axel Chemla-Romeu-Santos, and Adrien Bitton. Bridging Audio Analysis, Perception and Synthesis with Perceptually-regularized Variational Timbre Spaces. In *International Society for Music Information Retrieval Conference*, 2018. URL https://api.semanticscholar.org/CorpusID:53873046.

[141] Philippe Esling, Naotake Masuda, and Axel Chemla-Romeu-Santos. FlowSynth: Simplifying Complex Audio Generation Through Explorable Latent Spaces with Normalizing Flows. In *International Joint Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:220484250.

[142] Philippe Esling, Naotake Masuda, and Axel Chemla-Romeu-Santos. FlowSynth: simplifying complex audio generation through explorable latent spaces with normalizing flows. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 5273–5275, 2021.

[143] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

[144] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

[145] Gunnar Fant. The LF-model revisited. Transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2(3):40, 1995.

[146] Gunnar Fant, Johan Liljencrants, Qi-guang Lin, et al. A four-parameter model of glottal flow. *STL-QPSR*, 4(1985):1–13, 1985.

[147] Morwaread Mary Farbood. *Hyperscore : a new approach to interactive, computer-generated music*. Thesis, Massachusetts Institute of Technology, 2001. URL https://dspace.mit.edu/handle/1721.1/61122. Accepted: 2011-02-23T14:16:23Z.

[148] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 15–29. Springer, 2010.

[149] Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657, 2016.

[150] Haytham M Fayek and Anurag Kumar. Large scale audiovisual learning of sounds with weakly labeled data. *arXiv preprint arXiv:2006.01595*, 2020.

[151] Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. Synthasr: Unlocking synthetic data for speech recognition. *arXiv preprint arXiv:2106.07803*, 2021.

[152] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[153] Jonathan S Feinstein. *Creativity in Large-Scale Contexts: Guiding Creative Engagement and Exploration*. Stanford University Press, 2023.

[154] Ronald A Finke, Thomas B Ward, and Steven M Smith. Creative Cognition: Theory. *Research and Applications: MIT press Cambridge, MA*, 1992.

[155] Linda Flower and John R Hayes. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387, 1981.

[156] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound Datasets: A Platform for the Creation of Open Audio Datasets. pages 486–493, October 2017.

[157] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k:

an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

[158] Eduardo Fonseca, Diego Ortego, Kevin McGuinness, Noel E O'Connor, and Xavier Serra. Unsupervised contrastive learning of sound event representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 371–375. IEEE, 2021.

[159] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 411–412, 2013.

[160] Andrew Fowler, Kurt Partridge, Ciprian Chelba, Xiaojun Bi, Tom Ouyang, and Shumin Zhai. Effects of language modeling and its personalization on touchscreen typing performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 649–658, 2015.

[161] Marcelle Freiman. A 'cognitive turn'in creative writing–Cognition, body and imagination. *New Writing*, 12(2):127–142, 2015.

[162] Megan French and Jeff Hancock. What's the folk theory? Reasoning about cyber-social systems. *Reasoning About Cyber-Social Systems (February 2, 2017)*, 2017.

[163] Richard P Gabriel, Jilin Chen, and Jeffrey Nichols. InkWell: A Creative Writer's Creative Assistant. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pages 93–102, 2015.

[164] Leonardo Gabrielli, Stefano Tomassetti, Stefano Squartini, Carlo Zinato, et al. Introducing deep machine learning for parameter estimation in physical modelling. In *Proceedings of the 20th international conference on digital audio effects*, 2017.

[165] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. Design patterns: Abstraction and reuse of object-oriented design. In *ECOOP'93—Object-Oriented Programming: 7th European Conference Kaiserslautern, Germany, July 26–30, 1993 Proceedings 7*, pages 406–431. Springer, 1993.

[166] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 758–775. Springer, 2020.

[167] Heting Gao, Kaizhi Qian, Junrui Ni, Chuang Gan, Mark A Hasegawa-Johnson, Shiyu Chang, and Yang Zhang. Speech self-supervised learning using diffusion model synthetic data. In *Forty-first International Conference on Machine Learning*, 2024.

[168] Ruohan Gao and Kristen Grauman. 2.5D Visual Sound. In *CVPR*, 2019.

[169] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP (1)*, 2021.

[170] Yingming Gao, Simon Stone, and Peter Birkholz. Articulatory Copy Synthesis Based on a Genetic Algorithm. In *INTERSPEECH*, pages 3770–3774, 2019.

[171] Ricardo Antonio García. *Automatic generation of sound synthesis techniques*. PhD thesis, Massachusetts Institute of Technology, 2001.

[172] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. *arXiv preprint arXiv:2111.10882*, 2021.

[173] John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

[174] William Gaver. What should we expect from research through design? In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 937–946, 2012.

[175] William W. Gaver, Randall B. Smith, and Tim O'Shea. Effective Sounds in Complex Systems: The ARKOLA Simulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '91, page 85–90, New York, NY, USA, 1991. Association for Computing Machinery. ISBN 0897913833. doi: 10.1145/108844.108857. URL https://doi.org/10.1145/108844.108857.

[176] Clifford Geertz. *The interpretation of cultures*, volume 5019. Basic books, 1973.

[177] Susan A Gelman and Cristine H Legare. Concepts and folk theories. *Annual review of anthropology*, 40:379–398, 2011.

[178] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. URL

https://api.semanticscholar.org/CorpusID:21519176.

[179] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[180] John S Gero and Bimal Kumar. Expanding design spaces through new design variables. *Design Studies*, 14(2):210–221, 1993.

[181] Katy Ilonka Gero and Lydia B Chilton. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[182] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376316. URL https://doi.org/10.1145/3313831.3376316.

[183] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pages 1002–1019, 2022.

[184] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals. Multilingual training of deep neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 7319–7323. IEEE, 2013.

[185] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*, 2018.

[186] Elena L. Glassman. Designing Interfaces for Human-Computer Communication: An On-Going Collection of Considerations, September 2023. URL http://arxiv.org/abs/2309.02257. arXiv:2309.02257 [cs].

[187] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M

Kitani, and Jeffrey P Bigham. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference*, pages 549–559, 2019.

[188] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. Twitter A11y: A browser extension to make Twitter images accessible. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–12, 2020.

[189] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into Self-Supervised Monocular Depth Prediction. In *The International Conference on Computer Vision (ICCV)*, October 2019.

[190] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[191] Yuan Gong, Yu-An Chung, and James R. Glass. AST: Audio Spectrogram Transformer. *ArXiv*, abs/2104.01778, 2021. URL https://api.semanticscholar.org/CorpusID:233024831.

[192] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[193] Skylar Gordon, Robert Mahari, Manaswi Mishra, and Ziv Epstein. Co-creation and ownership for AI radio. In *International Conference on Computational Creativity*, 2022.

[194] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*, 2022.

[195] Samuel W Greenhouse and Seymour Geisser. On methods in the analysis of profile data. *Psychometrika*, 24(2):95–112, 1959.

[196] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1331–1335. IEEE, 2024.

[197] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A Review on

Generative Adversarial Networks: Algorithms, Theory, and Applications, 2020.

[198] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021.

[199] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023.

[200] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 417–434. Springer, 2020.

[201] Juan Sebastián Gómez-Cañón, Estefanía Cano, Tuomas Eerola, Perfecto Herrera, Xiao Hu, Yi-Hsuan Yang, and Emilia Gómez. Music Emotion Recognition: Toward new, robust standards in personalized and context-sensitive applications. *IEEE Signal Processing Magazine*, 38(6):106–114, November 2021. ISSN 1558-0792. doi: 10.1109/MSP.2021.3106232. URL https://ieeexplore.ieee.org/abstract/document/9591555. Conference Name: IEEE Signal Processing Magazine.

[202] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[203] Masato Hagiwara, Maddie Cusimano, and Jen-Yu Liu. Modeling Animal Vocalizations through Synthesizers. *arXiv preprint arXiv:2210.10857*, 2022.

[204] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.

[205] Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. Gibbs sampling with people. *Advances in neural information processing systems*, 33:10659–10671, 2020.

[206] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[207] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[208] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[209] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX. URL http://github.com/google/flax, 2023.

[210] Hermann von Helmholtz. Concerning the perceptions in general. 1867.

[211] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[212] Jr Hiller and L. M. Isaacson. Musical Composition with a High Speed Digital Computer. Audio Engineering Society, October 1957. URL https://www.aes.org/e-lib/browse.cfm?elib=189.

[213] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[214] Charles Joseph Holbrow. *Fluid Music: A New Model for Radically Collaborative Music Production*. PhD thesis, Massachusetts Institute of Technology, 2021.

[215] Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. The variably intense vocalizations of affect and emotion (VIVAE) corpus prompts new perspective on nonspeech perception. *Emotion*, 22(1):213, 2022.

[216] Henkjan Honing, Carel ten Cate, Isabelle Peretz, and Sandra E. Trehub. Without it no music: cognition, biology and evolution of musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1664):20140088, March 2015. doi: 10.1098/rstb.2014.0088. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2014.0088. Publisher: Royal Society.

[217] Stephanie Houde and Charles Hill. What do prototypes prototype? In *Handbook of human-computer interaction*, pages 367–381. Elsevier, 1997.

[218] Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. SciCap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624*, 2021.

[219] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019.

[220] Ting-Yao Hu, Mohammadreza Armandpour, Ashish Shrivastava, Jen-Hao Rick Chang, Hema Koppula, and Oncel Tuzel. Synt++: Utilizing imperfect synthetic data to improve speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7682–7686. IEEE, 2022.

[221] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music Transformer: Generating Music with Long-Term Structure. September 2018. URL https://openreview.net/forum?id=rJe4ShAcF7.

[222] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-An-Audio 2: Temporal-Enhanced Text-to-Audio Generation. *arXiv preprint arXiv:2305.18474*, 2023.

[223] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. Do LVLMs Understand Charts? Analyzing and Correcting Factual Errors in Chart Captioning. *arXiv preprint arXiv:2312.10160*, 2023.

[224] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022.

[225] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision*, pages 709–727. Springer, 2020.

[226] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation

with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023.

[227] Arefin Huq, Juan Pablo Bello, and Robert Rowe. Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research*, 39(3):227–244, 2010.

[228] Alyssa Hwang, Andrew Head, and Chris Callison-Burch. Grounded Intuition of GPT-Vision's Abilities with Scientific Images. *arXiv preprint arXiv:2311.02069*, 2023.

[229] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339, 2013.

[230] Khalil Iskarous, Louis Goldstein, Douglas H Whalen, Mark Tiede, and Philip Rubin. CASY: The Haskins configurable articulatory synthesizer. In *International Congress of Phonetic Sciences, Barcelona, Spain*, pages 185–188, 2003.

[231] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.

[232] Mahmoud A Ismail. Vocal Tract Area Function Estimation Using Particle Swarm. *J. Comput.*, 3(6):32–38, 2008.

[233] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021.

[234] Jongheon Jeong and Jinwoo Shin. Training gans with stronger augmentations via contrastive discriminator. *arXiv preprint arXiv:2103.09742*, 2021.

[235] Elena Jessop, Peter A Torpey, and Benjamin Bloomberg. Music and Technology in Death and the Powers. 2011.

[236] KV Jobin, Ajoy Mondal, and CV Jawahar. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79. IEEE, 2019.

[237] Nicolas Jonason and Bob LT Sturm. TimbreCLIP: Connecting Timbre to Text and

Images. *arXiv preprint arXiv:2211.11225*, 2022.

[238] Quincy Jones. *12 Notes: On Life and Creativity*. Harry N. Abrams, April 2022. ISBN 978-1-4197-5256-8.

[239] Patrick W Jordan, Bruce Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. *Usability evaluation in industry*. CRC Press, 1996.

[240] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589, 2021.

[241] Patrik N. Juslin and John A. Sloboda, editors. *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, January 2010. ISBN 978-0-19-169643-5. doi: 10.1093/acprof:oso/9780199230143.001.0001. URL https://academic.oup.com/book/38621.

[242] Mahdi M Kalayeh, Shervin Ardeshir, Lingyi Liu, Nagendra Kamath, and Ashok Chandrashekar. On negative sampling for audio-visual contrastive learning from movies. *arXiv preprint arXiv:2205.00073*, 2022.

[243] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2018.

[244] Wonjune Kang, Mark Hasegawa-Johnson, and Deb Roy. End-to-End Zero-Shot Voice Conversion with Location-Variable Convolutions. In *Proc. INTERSPEECH 2023*, pages 2303–2307, 2023. doi: 10.21437/Interspeech.2023-2298.

[245] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[246] Scott Barry Kaufman and James C Kaufman. *The psychology of creative writing*. Cambridge University Press, 2009.

[247] Margarita Kaushanskaya, Henrike K Blumenfeld, and Viorica Marian. The language

experience and proficiency questionnaire (leap-q): Ten years later. *Bilingualism: Language and Cognition*, 23(5):945–950, 2020.

[248] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.

[249] Peter Keller. Ensemble performance: Interpersonal alignment of musical expression. In *Expressiveness in music performance: Empirical approaches across styles and cultures*, pages 260–282. Oxford University Press, New York, NY, US, 2014. ISBN 978-0-19-965964-7. doi: 10.1093/acprof:oso/9780199659647.003.0015.

[250] Peter E. Keller. Joint action in music performance. In *Enacting intersubjectivity: A cognitive and social perspective on the study of interactions*, Emerging communication: Studies on new technologies and practices in communication, pages 205–221. IOS Press, Amsterdam, Netherlands, 2008. ISBN 978-1-58603-850-2.

[251] John L. Kelly and Carol C. Lochbaum. Speech synthesis. *Proceedings of the Fourth International Congress on Acoustics*, 1962.

[252] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. IEEE, 1995.

[253] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[254] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

[255] Bongjun Kim, Madhav Ghei, Bryan Pardo, and Zhiyao Duan. Vocal Imitation Set: a dataset of vocally imitated sound events using the AudioSet ontology. In *DCASE*, pages 148–152, 2018.

[256] H. Kim, L. Remaggi, P. J. B. Jackson, and A. Hilton. Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360° Images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 120–126, 2019. doi:

10.1109/VR.2019.8798247.

[257] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165. IEEE, 2018.

[258] Gary King, Robert O Keohane, and Sidney Verba. *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press, 2021.

[259] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[260] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[261] Paul Kockelman. Agent, person, subject, self. 2006.

[262] Paul Kockelman. *Agent, person, subject, self: A theory of ontology, interaction, and infrastructure*. Oxford University Press, 2013.

[263] Homare Kon and Hideki Koike. Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images. In *Audio Engineering Society*, 05 2018.

[264] Homare Kon and Hideki Koike. Estimation of Late Reverberation Characteristics from a Single Two-Dimensional Environmental Image Using Convolutional Neural Networks. *Journal of the Audio Engineering Society*, 67:540–548, 08 2019. doi: 10.17743/jaes.2018.0069.

[265] Homare Kon and Hideki Koike. An auditory scaling method for reverb synthesis from a single two-dimensional image. *Acoustical Science and Technology*, 41(4):675–685, 2020. doi: 10.1250/ast.41.675.

[266] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.

[267] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient

training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

[268] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*, 2020.

[269] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

[270] Paul Konstantin Krug, Simon Stone, and Peter Birkholz. Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies. *Proc. SSW*, 11:102–107, 2021.

[271] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[272] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35 (12):2891–2903, 2013.

[273] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020.

[274] Ziva Kunda, Dale T Miller, and Theresa Claire. Combining social concepts: The role of causal reasoning. *Cognitive Science*, 14(4):551–577, 1990.

[275] Moreno La Quatra, Alkis Koudounas, Lorenzo Vaiani, Elena Baralis, Paolo Garza, Luca Cagliero, and Sabato Marco Siniscalchi. Benchmarking Representations for Speech, Music, and Acoustic Events. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024.

[276] Brandon LaBelle. *Sonic Agency: Sound and Emergent Forms of Resistance*. 2023. URL https://mitpress.mit.edu/9781912685950/sonic-agency/.

[277] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40,

2017.

[278] Carolyn Lamb, Daniel G Brown, and Charles LA Clarke. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR)*, 51(2):1–34, 2018.

[279] Brian M Landry and Mark Guzdial. iTell: supporting retrospective storytelling with digital photos. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 160–168, 2006.

[280] Robert Lange, Tom Schaul, Yutian Chen, Chris Lu, Tom Zahavy, Valentin Dalibard, and Sebastian Flennerhag. Discovering attention-based genetic algorithms via meta-black-box optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 929–937, 2023.

[281] Robert Tjarko Lange. evosax: JAX-based Evolution Strategies. *arXiv preprint arXiv:2212.04180*, 2022.

[282] Robert Tjarko Lange. evosax: Jax-based evolution strategies. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pages 659–662, 2023.

[283] Jaron Lanier. *You are not a gadget: A manifesto*. Vintage, 2011.

[284] Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444. IEEE, 2020.

[285] Cyril Laurier, Owen Meyers, Joan Serra, Martin Blech, Perfecto Herrera, and Xavier Serra. Indexing music by mood: design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*, 48:161–184, 2010.

[286] Kimaya Lecamwasam*, Samantha Gutierrez Arango*, Nikhil Singh, Neska Elhaouij, Max Addae, and Rosalind Picard. Investigating the Physiological and Psychological Effect of an Interactive Musical Interface for Stress and Anxiety Reduction. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–9, 2023.

[287] Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*, 2021.

[288] Keun Sup Lee, Nicholas J Bryan, and Jonathan S Abel. Approximating measured reverberation using a hybrid fixed/switched convolution structure. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx'10)*, 2010.

[289] Mina Lee, Percy Liang, and Qian Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19, 2022.

[290] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–35, 2024.

[291] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[292] Roger Levy and Galen Andrew. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer, 2006.

[293] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021.

[294] Dingzeyu Li, Timothy R. Langlois, and Changxi Zheng. Scene-Aware Audio for 360° Videos. *ACM Trans. Graph.*, 37(4), July 2018. ISSN 0730-0301. doi: 10.1145/3197517.3201391. URL https://doi.org/10.1145/3197517.3201391.

[295] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[296] Bryan Lim, Maxime Allard, Luca Grillotti, and Antoine Cully. Accelerated Quality-

Diversity for Robotics through Massive Parallelism. *arXiv preprint arXiv:2202.01258*, 2022.

[297] Youn-Kyung Lim, Erik Stolterman, and Josh Tenenberg. The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 15(2):1–27, 2008.

[298] Zhiyu Lin and Mark O Riedl. Plug-and-Blend: A Framework for Plug-and-Play Controllable Story Generation with Sketches. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 17, pages 58–65, 2021.

[299] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE, 2020.

[300] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*, 2022.

[301] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[302] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[303] Li Liu, Gang Feng, and Denis Beautemps. Inner lips feature extraction based on CLNF with hybrid dynamic template for Cued Speech. *EURASIP Journal on Image and Video Processing*, 2017:1–15, 2017.

[304] Li Liu, Gang Feng, Denis Beautemps, and Xiao-Ping Zhang. Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition. *IEEE Transactions on Multimedia*, 23:292–305, 2020.

[305] Ming Liu, Rafael A Calvo, and Vasile Rus. G-Asks: An intelligent automatic question generation system for academic writing support. *Dialogue & Discourse*, 3(2):101–124,

2012.

[306] Peng Liu, Quanjie Yu, Zhiyong Wu, Shiyin Kang, Helen Meng, and Lianhong Cai. A deep recurrent approach for acoustic-to-articulatory inversion. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4450–4454. IEEE, 2015.

[307] Xubo Liu, Egor Lakomkin, Konstantinos Vougioukas, Pingchuan Ma, Honglie Chen, Ruiming Xie, Morrie Doulaty, Niko Moritz, Jachym Kolar, Stavros Petridis, et al. Synthvsr: Scaling up visual speech recognition with synthetic supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18806–18815, 2023.

[308] Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, Qiushi Huang, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D Plumbley, et al. WavJourney: Compositional Audio Creation with Large Language Models. *arXiv preprint arXiv:2307.14335*, 2023.

[309] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[310] Sebastian Löbbers, Louise Thorpe, and György Fazekas. SketchSynth: Cross-Modal Control of Sound Synthesis. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, pages 164–179. Springer, 2023.

[311] Shayne Longpre, Robert Mahari, Ariel N Lee, Campbell S Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole J Hunter, et al. Consent in crisis: The rapid decline of the ai data commons. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[312] Patrice Lopez. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, pages 473–474. Springer, 2009.

[313] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[314] Augusta Ada Lovelace. Sketch of the analytical engine invented by Charles Babbage, by LF Menabrea, officer of the military engineers, with notes upon the memoir by the translator. *Taylor's Scientific Memoirs*, 3:666–731, 1842.

[315] Jens Ludwig and Sendhil Mullainathan. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, 2024.

[316] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*, 2019.

[317] Alan Lundgard and Arvind Satyanarayan. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, 28(1):1073–1083, 2021.

[318] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*, 2020.

[319] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[320] N. Macdonald, L. Frase, P. Gingrich, and S. Keenan. The Writer's Workbench: Computer Aids for Text Analysis. *IEEE Transactions on Communications*, 30(1):105–110, 1982. doi: 10.1109/TCOM.1982.1095380.

[321] Tod Machover. Repertoire Remix Live Demonstration, August 2013. URL https://operaofthefuture.com/2013/08/09/repertoire-remix-video/.

[322] Tod Machover and Charles Holbrow. Toward New Musics: What The Future Holds For Sound Creativity. *NPR*, July 2019. URL https://www.npr.org/2019/07/26/745315045/towards-new-musics-what-the-future-holds-for-sound-creativity.

[323] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. Designing tools for high-quality alt text authoring. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14, 2021.

[324] Ken MacLean. Voxforge. *Ken MacLean.[Online]. Available: http://www. voxforge. org/home.[Acedido em 2012]*, 2018.

[325] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-Shot Audio-Visual Learning of Environment Acoustics. *arXiv preprint arXiv:2206.04006*, 2022.

[326] Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. CDPAM: Contrastive learning for perceptual audio similarity. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 196–200. IEEE, 2021.

[327] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[328] Viorica Marian, Henrike K Blumenfeld, and Margarita Kaushanskaya. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of speech, language, and hearing research*, 2007.

[329] Keith D Martin and Youngmoo E Kim. 2pMU9. Musical instrument identification: A pattern-recognition approach. In *Presented at the 136th meeting of the Acoustical Society of America*, 1998.

[330] Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[331] Lee Martin and Nick Wilson. Defining creativity with discovery. *Creativity Research Journal*, 29(4):417–425, 2017.

[332] Ninon Lizé Masclef and T. Anderson Keller. Deep Generative Models of Music Expectation. In *NeurIPS 2023 Workshop on Machine Learning for Audio*, October 2023. URL http://arxiv.org/abs/2310.03500. arXiv:2310.03500 [cs, eess].

[333] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning. *arXiv preprint arXiv:2305.14761*, 2023.

[334] Naotake Masuda and Daisuke Saito. Synthesizer Sound Matching with Differentiable DSP. In *ISMIR*, pages 428–434, 2021.

[335] Naotake Masuda and Daisuke Saito. Quality-diversity for Synthesizer Sound Matching. *Journal of Information Processing*, 31:220–228, 2023.

[336] Max V Mathews, Joan E Miller, F Richard Moore, John R Pierce, and Jean-Claude Risset. *The technology of computer music*, volume 5. MIT press Cambridge, MA, 1969.

[337] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[338] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[339] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. SampleRNN: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016.

[340] Ravi Mehta, Rui Zhu, and Amar Cheema. Is noise always bad? Exploring the effects of ambient noise on creative cognition. *Journal of Consumer Research*, 39(4):784–799, 2012.

[341] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477, 2022.

[342] Z. Meng, F. Zhao, and M. He. The Just Noticeable Difference of Noise Length and Reverberation Perception. In *2006 International Symposium on Communications and Information Technologies*, pages 418–421, 2006. doi: 10.1109/ISCIT.2006.339980.

[343] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-Fidelity Generative Image Compression. *Advances in Neural Information Processing Systems*, 33, 2020.

[344] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification. *arXiv preprint arXiv:1807.09840*, 2018.

[345] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. Co-Writing

Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34, 2023.

[346] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[347] Manaswi Mishra. *Living, Singing AI: An evolving, intelligent, scalable, bespoke composition system*. PhD thesis, Massachusetts Institute of Technology, 2021.

[348] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[349] Luke Mo*, Manuel Cherep*, Nikhil Singh*, Quinn Langford, and Pattie Maes. Articulatory synthesis of speech and diverse vocal sounds via optimization. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

[350] Valerie S Morash, Yue-Ting Siu, Joshua A Miele, Lucia Hasty, and Steven Landau. Guiding novice web workers in making image descriptions using templates. *ACM Transactions on Accessible Computing (TACCESS)*, 7(4):1–21, 2015.

[351] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31, 2018.

[352] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021.

[353] Philippa Mothersill and V Michael Bove. Design Daydreams: Juxtaposing Digital and Physical Inspiration. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*, pages 265–269, 2019.

[354] Damian T Murphy and Simon Shelley. Openair: An interactive auralization web resource and database. In *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[355] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-

scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020.

[356] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2Action: Cross-modal Supervision for Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10317–10326, 2020.

[357] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34, 2021.

[358] Frieder Nake. Human-computer interaction: signs and signals interfacing. *Languages of design*, 2(193-205), 1994.

[359] Allen Newell. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. *Computer Science Department*, 1973.

[360] Mickey Nguyen, Matthew Crane, John Romley, and Yannis M Paulus. Accessibility of Figures in Leading Biomedical and Ophthalmology Journals: Analysis of Alternative Text Use. *Investigative Ophthalmology & Visual Science*, 64(8):2806–2806, 2023.

[361] Eric Nichols, Leo Gao, and Randy Gomez. Collaborative storytelling with large-scale neural language models. In *Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–10, 2020.

[362] Hugo Nicolau, André Rodrigues, André Santos, Tiago Guerreiro, Kyle Montague, and João Guerreiro. The Design Space of Nonvisual Word Completion. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 249–261, 2019.

[363] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[364] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for Audio: Exploring Pre-trained General-purpose Audio Representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:137–151, 2022.

[365] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Masked Modeling Duo: Learning Representations by Encouraging Both Networks to Model the Input. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[366] Ellen W Nold. Revising. *Writing: the nature, development, and teaching of written communication*, 2:67–79, 1981.

[367] Donald A Norman. Some observations on mental models. *Mental models*, 7(112): 7–14, 1983.

[368] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[369] Martin Nystrand. A social-interactive model of writing. *Written communication*, 6(1): 66–85, 1989.

[370] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[371] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[372] OpenAI. GPT-4 Technical Report. *ArXiv*, abs/2303.08774, 2023.

[373] OpenAI. GPT-4 Technical Report, 2023.

[374] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2021.

[375] Sharon Oviatt. Mutual disambiguation of recognition errors in a multimodel architecture. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 576–583, 1999.

[376] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[377] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016.

[378] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016.

[379] François Pachet, Pierre Roy, and Benoit Carré. Assisted Music Creation with Flow Machines: Towards New Categories of New. In Eduardo Reck Miranda, editor, *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*, pages 485–520. Springer International Publishing, Cham, 2021. ISBN 978-3-030-72116-9. doi: 10.1007/978-3-030-72116-9_18. URL https://doi.org/10.1007/978-3-030-72116-9_18.

[380] Allan Paivio. *Mental representations: A dual coding approach*. Oxford University Press, 1990.

[381] Seymour A. Papert. *Mindstorms: Children, Computers, And Powerful Ideas*. Basic Books, October 2020. ISBN 978-1-5416-7510-0. Google-Books-ID: nDjRDwAAQBAJ.

[382] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[383] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[384] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9587, 2021.

[385] Geoffroy Peeters. A large set of audio features for sound description (similarity and

classification) in the CUIDADO project. *CUIDADO Ist Project Report*, 54(0):1–25, 2004.

[386] Renato S Pellegrini. Quality assessment of auditory virtual environments. In *Proceedings of the 2001 International Conference on Auditory Display*, 2001.

[387] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[388] Jean Piaget. *The equilibration of cognitive structures: The central problem of intellectual development*. University of Chicago press, 1985.

[389] Rosalind W. Picard. Computer learning of subjectivity. *ACM Computing Surveys*, 27 (4):621–623, December 1995. ISSN 0360-0300, 1557-7341. doi: 10.1145/234782.234805. URL https://dl.acm.org/doi/10.1145/234782.234805.

[390] Rosalind W. Picard. *Affective Computing*. The MIT Press, July 2000. ISBN 978-0-262-28158-4. doi: 10.7551/mitpress/1140.001.0001. URL https://direct.mit.edu/books/book/4296/Affective-Computing.

[391] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. *Proceedings of the 23rd ACM international conference on Multimedia*, 2015. URL https://api.semanticscholar.org/CorpusID:17567398.

[392] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2015.

[393] Karol J Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[394] Ben Pietrzak, Ben Swanson, Kory Mathewson, Monica Dinculescu, and Sherol Chen. Story Centaur: Large Language Model Few Shot Learning as a Creative Writing Tool. 2021.

[395] Trevor Pinch and Frank Trocco. *Analog days: The invention and impact of the Moog synthesizer*. Harvard University Press, 2004.

[396] Reinier Plomp and Willem Johannes Maria Levelt. Tonal consonance and critical

bandwidth. *The journal of the Acoustical Society of America*, 38(4):548–560, 1965.

[397] Jonathan A Plucker, Ronald A Beghetto, and Gayle T Dow. Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational psychologist*, 39(2):83–96, 2004.

[398] Paul Prior. A sociocultural theory of writing. *Handbook of writing research*, pages 54–66, 2006.

[399] Jeb S Puryear and Kristen N Lamb. Defining creativity: How far have we come since Plucker, Beghetto, and Dow? *Creativity Research Journal*, 32(3):206–214, 2020.

[400] Philip Quinn and Shumin Zhai. A cost-benefit study of text entry suggestion interaction. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 83–88, 2016.

[401] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[402] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[403] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[404] Janet Rafner, Roger E Beaty, James C Kaufman, Todd Lubart, and Jacob Sherson. Creativity in the age of generative AI. *Nature Human Behaviour*, 7(11):1836–1838, 2023.

[405] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9301–9310, 2021.

[406] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, and Scott Gray. DALL·E: Creating Images from Text. *OpenAI Blog*, 2021.

[407] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[408] Marco A Martínez Ramírez, Oliver Wang, Paris Smaragdis, and Nicholas J Bryan. Differentiable signal processing with black-box audio effects. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 66–70. IEEE, 2021.

[409] David Ramsay, Ishwarya Ananthabhotla, and Joseph Paradiso. The Intrinsic Memorability of Everyday Sounds. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[410] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. IR-GAN: Room Impulse Response Generator for Speech Augmentation, 2021.

[411] Mirco Ravanelli and Yoshua Bengio. Learning Speaker Representations with Mutual Information. In *Interspeech*, pages 1153–1157, 2019.

[412] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.

[413] Andrew Reilly and David McGrath. Convolution processing for realistic reverberation. In *Audio Engineering Society Convention 98*. Audio Engineering Society, 1995.

[414] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[415] Luca Remaggi, Hansung Kim, Philip JB Jackson, and Adrian Hilton. Reproducing real world acoustics in virtual reality using spherical cameras. In *Proceedings of the 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[416] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 762–771, 2018.

[417] Harri Renney, Benedict Gaster, and Thomas J Mitchell. Survival of the synthesis—GPU

accelerating evolutionary sound matching. *Concurrency and Computation: Practice and Experience*, 34(10):e6824, 2022.

[418] Michael Rettinger. Reverberation chambers for broadcasting and recording studios. *Journal of the Audio Engineering Society*, 5(1):18–22, 1957.

[419] Julien Ricard. Towards computational morphological description of sound. *DEA pre-thesis research work, Universitat Pompeu Fabra, Barcelona*, 2004.

[420] Korin Richmond. Estimating articulatory parameters from the acoustic speech signal. *Annexe Thesis Digitisation Project 2017 Block 11*, 2002.

[421] Mark O. Riedl. Computational Narrative Intelligence: A Human-Centered Goal for Artificial Intelligence, February 2016. URL http://arxiv.org/abs/1602.06484. arXiv:1602.06484 [cs].

[422] Janne Riionheimo and Vesa Välimäki. Parameter estimation of a plucked string synthesis model using a genetic algorithm with perceptual fitness calculation. *EURASIP Journal on Advances in Signal Processing*, 2003:1–15, 2003.

[423] Arie Rip. Folk theories of nanotechnologists. *Science as culture*, 15(4):349–365, 2006.

[424] Bertrand Rivet, Wenwu Wang, Syed Mohsen Naqvi, and Jonathon A Chambers. Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, 31(3):125–134, 2014.

[425] Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. "I Can't Reply with That": Characterizing Problematic Email Reply Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2021.

[426] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[427] H Robjohns. Sony DRE S777 sampling digital reverb. *Sound on Sound*, 15, 1999.

[428] Davide Rocchesso, Guillaume Lemaitre, Patrick Susini, Sten Ternström, and Patrick Boussard. Sketching sound with voice and gesture. *interactions*, 22(1):38–41, 2015.

[429] Melissa Roemmele and Andrew S Gordon. Creative help: A story writing assistant. In *International Conference on Interactive Digital Storytelling*, pages 81–92. Springer, 2015.

[430] Melissa Roemmele and Andrew S Gordon. Automated assistance for creative writing with an rnn language model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, pages 1–2, 2018.

[431] Yvonne Rogers. New theoretical approaches for HCI. *Annual review of information science and technology*, 38(1):87–143, 2004.

[432] Martin Rohrmeier. On Creativity, Music's AI Completeness, and Four Challenges for Artificial Musical Creativity. 5(1):50–66, March 2022. ISSN 2514-3298. doi: 10.5334/tismir.104. URL https://transactions.ismir.net/articles/10.5334/tismir.104. Number: 1 Publisher: Ubiquity Press.

[433] Martin Rohrmeier and Patrick Rebuschat. Implicit Learning and Acquisition of Music. *Topics in Cognitive Science*, 4(4):525–553, 2012. ISSN 1756-8765. doi: 10.1111/j.1756-8765.2012.01223.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2012.01223.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1756-8765.2012.01223.x.

[434] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[435] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20, 2010.

[436] Charles Rosen. *Sonata Forms*. 1988. URL https://www.penguinbookshop.com/book/9780393302196.

[437] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech recognition with augmented synthesized speech. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 996–1002. IEEE, 2019.

[438] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7069–7073. IEEE, 2020.

[439] Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE, 2021.

[440] Pramit Saha and Sidney Fels. Learning joint articulatory-acoustic representations with normalizing flows. *arXiv preprint arXiv:2005.09463*, 2020.

[441] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

[442] Kaushal Sali and Alexander Lerch. Generating Impulse Responses using Recurrent Neural Networks, 2020. URL https://109ecc9c-0e76-482f-90c5-fe6cd93cf581.filesusr.com/ugd/4a27c6_fa8281568425494e8ca16133fe724c6e.pdf.

[443] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[444] Sepehr Sameni, Simon Jenni, and Paolo Favaro. Spatio-Temporal Crop Aggregation for Video Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5664–5674, 2023.

[445] Adam Sanborn and Thomas Griffiths. Markov chain Monte Carlo with people. *Advances in neural information processing systems*, 20, 2007.

[446] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.

[447] Patrick E. Savage. Cultural evolution of music. *Palgrave Communications*, 5(1):1–12, February 2019. ISSN 2055-1045. doi: 10.1057/s41599-019-0221-1. URL

https://www.nature.com/articles/s41599-019-0221-1. Number: 1 Publisher: Palgrave.

[448] Carl Schissler and Dinesh Manocha. Interactive Sound Propagation and Rendering for Large Multi-Source Scenes. *ACM Trans. Graph.*, 36(4), September 2016. ISSN 0730-0301. doi: 10.1145/3072959.2943779. URL https://doi.org/10.1145/3072959.2943779.

[449] Olaf Schleusing, Tomi Kinnunen, Brad Story, and Jean-Marc Vesin. Joint source-filter optimization for accurate vocal tract estimation using differential evolution. *IEEE transactions on audio, speech, and language processing*, 21(8):1560–1572, 2013.

[450] Arnold Schoenberg. *Structural Functions of Harmony*. W. W. Norton & Company, revised edition edition, June 1969. ISBN 978-1-934854-15-0.

[451] Jacob Schreiber, Jeffrey Bilmes, and William Stafford Noble. apricot: Submodular selection for data summarization in Python. *The Journal of Machine Learning Research*, 21(1):6474–6479, 2020.

[452] M. R. Schroeder and B. F. Logan. "Colorless" Artificial Reverberation. *IRE Transactions on Audio*, 9(6):209–214, 1961. ISSN 21682984. doi: 10.1109/TAU.1961.1166351.

[453] Hugo Scurto, Bavo Van Kerrebroeck, Baptiste Caramiaux, and Frédéric Bevilacqua. Designing deep reinforcement learning for human parameter exploration. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(1):1–35, 2021.

[454] Allan Seago, Simon Holland, and Paul Mulholland. A critical analysis of synthesizer user interfaces for timbre. 2004.

[455] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[456] Phoebe Sengers and Bill Gaver. Staying open to interpretation: engaging multiple meanings in design and evaluation. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 99–108, 2006.

[457] Shakhnarovich, Viola, and Darrell. Fast pose estimation with parameter-sensitive

hashing. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 750–757. IEEE, 2003.

[458] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.

[459] Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. Beyond Summarization: Designing AI Support for Real-World Expository Writing Tasks. *arXiv preprint arXiv:2304.02623*, 2023.

[460] Hayato Shibata, Mingxin Zhang, and Takahiro Shinozaki. Unsupervised acoustic-to-articulatory inversion neural network learning based on deterministic policy gradient. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 530–537. IEEE, 2021.

[461] Jordie Shier. The synthesizer programming problem: improving the usability of sound synthesizers, 2021.

[462] Jordie Shier, George Tzanetakis, and Kirk McNally. Spiegelib: An automatic synthesizer programming library. In *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

[463] Ben Shneiderman. The limits of speech recognition. *Communications of the ACM*, 43 (9):63–65, 2000.

[464] Ben Shneiderman. Creativity support tools: accelerating discovery and innovation. *Communications of the ACM*, 50(12):20–32, 2007.

[465] Ben Shneiderman. *Human-centered AI*. Oxford University Press, 2022.

[466] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[467] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering

the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[468] Nikhil Singh. The Sound Sketchpad: Expressively Combining Large and Diverse Audio Collections. In *International Conference on Intelligent User Interfaces (IUI)*, pages 297–301, 2021.

[469] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 286–295, 2021.

[470] Nikhil Singh*, Guillermo Bernal*, Daria Savchenko*, and Elena L Glassman. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction*, 30(5):1–57, 2022.

[471] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L Glassman. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction*, 30(5):1–57, 2023.

[472] Nikhil Singh, Manaswi Mishra, and Tod Machover. AI for Musical Discovery. *An MIT Exploration of Generative AI*, mar 27 2024. https://mit-genai.pubpub.org/pub/30vaia0v.

[473] Nikhil Singh, Lucy Lu Wang, and Jonathan Bragg. FigurA11y: AI Assistance for Writing Scientific Alt Text. In *International Conference on Intelligent User Interfaces (IUI)*, pages 886–906, 2024.

[474] Nikhil Singh, Chih-Wei Wu, Iroro Orife, and Mahdi Kalayeh. Looking Similar, Sounding Different: Leveraging Counterfactual Cross-Modal Pairs for Audiovisual Representation Learning. *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

[475] Sidney L Smith and Nancy C Goodwin. Alphabetic data entry via the Touch-Tone pad: A comment, 1971.

[476] Daria Soboleva, Ondrej Skopek, Márius Šajgalík, Victor Cărbune, Felix Weissenberger, Julia Proskurnia, Bogdan Prisacari, Daniel Valcarce, Justin Lu, Rohit Prabhavalkar, et al. Replacing human audio with synthetic audio for on-device unspoken punctuation prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7653–7657. IEEE, 2021.

[477] Yiren Song. CLIPTexture: Text-Driven Texture Synthesis. In *Proceedings of the 30th*

ACM International Conference on Multimedia, pages 5468–5476, 2022.

[478] Yiren Song, Xuning Shao, Kang Chen, Weidong Zhang, Zhongliang Jing, and Minzhe Li. CLIPVG: Text-Guided Image Manipulation Using Differentiable Vector Graphics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2312–2320, 2023.

[479] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[480] Victor N Sorokin, Alexander S Leonov, and Alexander V Trushkin. Estimation of stability and accuracy of inverse problem solution for the vocal tract. *Speech Communication*, 30(1):55–74, 2000.

[481] Janne Spijkervet and John Ashley Burgoyne. Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410*, 2021.

[482] Clay Spinuzzi. The methodology of participatory design. *Technical communication*, 52 (2):163–174, 2005.

[483] Christian Steinmetz. NeuralReverberator, 2018. URL https://www.christiansteinmetz.com/projects-blog/neuralreverberator.

[484] Christian J Steinmetz and Joshua Reiss. pyloudnorm: A simple yet flexible loudness meter in Python. In *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021.

[485] Brad H Story. History of speech synthesis. In *The Routledge Handbook of Phonetics*, pages 9–33. Routledge, 2019.

[486] Fabian-Robert Stöter, Soumitro Chakrabarty, Emanuël Habets, and Bernd Edler. LibriCount, a dataset for speaker count estimation. URL https://zenodo.org/record/1216072, 2018.

[487] David Su, Rosalind W. Picard, and Yan Liu. AMAI: Adaptive music for affect improvement. *International Computer Music Association*, 2018. URL https://dspace.mit.edu/handle/1721.1/138102.2. Accepted: 2021-11-10T20:28:56Z.

[488] Kun Su, Xiulong Liu, and Eli Shlizerman. Audeo: Audio generation for a silent

performance video. *Advances in Neural Information Processing Systems*, 33:3325–3337, 2020.

[489] Felipe Petroski Such, Vashisht Madhavan, Edoardo Conti, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning. *arXiv preprint arXiv:1712.06567*, 2017.

[490] David Südholt, Mateo Cámara, Zhiyuan Xu, and Joshua D Reiss. Vocal tract area estimation by gradient descent. *arXiv preprint arXiv:2307.04702*, 2023.

[491] Minhyang Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2021.

[492] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. It's time for artistic correspondence in music and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10564–10574, 2022.

[493] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.

[494] Shunryū Suzuki. *Zen mind, beginner's mind.* Weatherhill, New York, [1st ed.] edition, 1972. ISBN 978-0-8348-0052-6.

[495] Modan Tailleur, Junwon Lee, Mathieu Lagrange, Keunwoo Choi, Laurie M Heller, Keisuke Imoto, and Yuki Okamoto. Correlation of Fr\'echet Audio Distance With Human Perception of Environmental Audio Is Embedding Dependant. *arXiv preprint arXiv:2403.17508*, 2024.

[496] Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020.

[497] Siu-Lan Tan, Matthew P. Spackman, and Matthew A. Bezdek. Viewers' Interpretations of Film Characters' Emotions: Effects of Presenting Film Music Before or After a Character is Shown. *Music Perception*, 25(2):135–152, December 2007. ISSN 0730-7829. doi: 10.1525/mp.2007.25.2.135. URL https://doi.org/10.1525/mp.2007.25.2.135.

[498] Benny J Tang, Angie Boggust, and Arvind Satyanarayan. VisText: A Benchmark for Semantically Rich Chart Captioning. *arXiv preprint arXiv:2307.05356*, 2023.

[499] Yujin Tang, Yingtao Tian, and David Ha. EvoJAX: Hardware-Accelerated Neuroevolution. *arXiv preprint arXiv:2202.05008*, 2022.

[500] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[501] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[502] Paul Théberge. *Any sound you can imagine: Making music/consuming technology*. Wesleyan University Press, 1997.

[503] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[504] Yingtao Tian and David Ha. Modern evolution strategies for creativity: Fitting concrete images and abstract concepts. In *International conference on computational intelligence in music, sound, art and design (part of evostar)*, pages 275–291. Springer, 2022.

[505] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[506] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[507] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. *arXiv preprint arXiv:2312.17742*, 2023.

[508] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024.

[509] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

[510] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023.

[511] James Traer and Josh H McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

[512] Sandra E. Trehub. The developmental origins of musicality. *Nature Neuroscience*, 6(7):669–673, July 2003. ISSN 1546-1726. doi: 10.1038/nn1084. URL https://www.nature.com/articles/nn1084. Number: 7 Publisher: Nature Publishing Group.

[513] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4475–4485, 2021.

[514] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On Mutual Information Maximization for Representation Learning. In *International Conference on Learning Representations*, 2019.

[515] Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1):1–13, 2020.

[516] Joseph Turian, Jordie Shier, George Tzanetakis, Kirk McNally, and Max Henry. One billion audio sounds from GPU-enabled modular synthesis. In *2021 24th International Conference on Digital Audio Effects (DAFx)*, pages 222–229. IEEE, 2021.

[517] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al. HEAR 2021: Holistic Evaluation of Audio Representations. *arXiv preprint arXiv:2203.03022*, 2022.

[518] Alan M. Turing. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460,

1950.

[519] Mark Turner and Gilles Fauconnier. A mechanism of creativity. *Alternation*, 6(2): 273–292, 1999.

[520] George Tzanetakis. GTZAN Music/Speech Collection. URL http://marsyas.info/index.html, 1999.

[521] George Tzanetakis and Perry Cook. Multifeature audio segmentation for browsing and annotation. In *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No. 99TH8452)*, pages 103–106. IEEE, 1999.

[522] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020.

[523] Vesa Valimaki, Julian D Parker, Lauri Savioja, Julius O Smith, and Jonathan S Abel. Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1421–1448, 2012.

[524] Jörgen Valk and Tanel Alumäe. VOXLINGUA107: A Dataset for Spoken Language Recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658, 2021. doi: 10.1109/SLT48900.2021.9383459.

[525] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *Arxiv*, 2016. URL https://arxiv.org/abs/1609.03499.

[526] Akito van Troyer. Constellation: a tool for creative dialog between audience and composer. In *10th International Symposium on Computer Music Multidisciplinary Research*, 2013. URL https://vantroyer.com/lib/doc/Constellation/Constellation.pdf.

[527] Akito van Troyer. Repertoire Remix in the Context of Festival City. In Damián Keller, Victor Lazzarini, and Marcelo S. Pimenta, editors, *Ubiquitous Music*, Computational Music Science, pages 51–63. Springer International Publishing, Cham, 2014. ISBN 978-3-319-11152-0. doi: 10.1007/978-3-319-11152-0_3. URL https://doi.org/10

.1007/978-3-319-11152-0_3.

[528] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017.

[529] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

[530] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

[531] Oliver Vitouch. When Your Ear Sets the Stage: Musical Context Effects in Film Perception. *Psychology of Music*, 29(1):70–83, April 2001. ISSN 0305-7356. doi: 10.1177/0305735601291005. URL https://doi.org/10.1177/0305735601291005. Publisher: SAGE Publications Ltd.

[532] Lev Vygotsky. Interaction Between Learning and Development. In *Mind and Society*, pages 79–91. Harvard University Press, 1978.

[533] Jianrong Wang, Jinyu Liu, Longxuan Zhao, Shanyu Wang, Ruiguo Yu, and Li Liu. Acoustic-to-articulatory inversion based on speech decomposition and auxiliary feature. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4808–4812. IEEE, 2022.

[534] Luyu Wang and Aaron van den Oord. Multi-format contrastive learning of audio representations. *arXiv preprint arXiv:2103.06508*, 2021.

[535] Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord. Multimodal self-supervised learning of general audio representations. *arXiv preprint arXiv:2104.12807*, 2021.

[536] Luyu Wang, Pauline Luc, Yan Wu, Adria Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle, Jean-Baptiste Alayrac, Sander Dieleman, Joao Carreira, et al. Towards learning universal audio representations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4593–4597. IEEE, 2022.

[537] Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023.

[538] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[539] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[540] Zilin Wang, Peng Liu, Jun Chen, Sipan Li, Jinfeng Bai, Gang He, Zhiyong Wu, and Helen Meng. A Synthetic Corpus Generation Method for Neural Vocoder Training. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[541] Megan Ward. Victorian Fictions of Computational Creativity. In Stephen Cave, Kanta Dihal, and Sarah Dillon, editors, *AI Narratives: A history of imaginative thinking about intelligent machines*, page 144. Oxford University Press, USA, 2020.

[542] Thomas B Ward and E Thomas Lawson. Creative cognition in science fiction and fantasy writing. 2009.

[543] Thomas B Ward, Steven M Smith, and Ronald A Finke. Creative cognition. *Handbook of creativity*, 189:212, 1999.

[544] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[545] Max Weber. *Max Weber on the methodology of the social sciences*. Free Press, 1949.

[546] Anna Wiener. Holly Herndon's Infinite Art. *The New Yorker*, November 2023. ISSN 0028-792X. URL https://www.newyorker.com/magazine/2023/11/20/holly-herndons-infinite-art. Section: onward and upward with technology.

[547] Candace Williams, Lilian de Greef, Ed Harris III, Leah Findlater, Amy Pavel, and Cynthia Bennett. Toward supporting quality alt text in computing publications. In *Proceedings of the 19th International Web for All Conference*, pages 1–12, 2022.

[548] Edward J Wisniewski. When concepts combine. *Psychonomic bulletin & review*, 4(2): 167–183, 1997.

[549] Earl Woodruff, Carl Bereiter, and Marlene Scardamalia. On the road to computer assisted compositions. *Journal of Educational Technology Systems*, 10(2):133–148, 1981.

[550] Amanda Woodward and Karen Hoyne. Infants' learning about words and sounds in relation to objects. *Child development*, 70(1):65–77, 1999.

[551] Chao-Yuan Wu and Philipp Krähenbühl. Towards Long-Form Video Understanding. In *CVPR*, 2021.

[552] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[553] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022.

[554] Peter Wu, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala K Anumanchipalli. Deep speech synthesis from articulatory representations. *arXiv preprint arXiv:2209.06337*, 2022.

[555] Peter Wu, Li-Wei Chen, Cheol Jun Cho, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala K Anumanchipalli. Speaker-independent acoustic-to-articulatory speech inversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[556] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1180–1192, 2017.

[557] Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan. Music ControlNet: Multiple Time-varying Controls for Music Generation, November 2023. URL http://arxiv.org/abs/2311.07069. arXiv:2311.07069 [cs, eess].

[558] Tung-Yu Wu, Tsu-Yuan Hsu, Chen-An Li, Tzu-Han Lin, and Hung-yi Lee. The efficacy of self-supervised speech models for audio representations. In *HEAR: Holistic Evaluation of Audio Representations*, pages 90–110. PMLR, 2022.

[559] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[560] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[561] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021.

[562] Xuenan Xu, Zhiling Zhang, Zelin Zhou, Pingyue Zhang, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu. BLAT: Bootstrapping Language-Audio Pre-training based on AudioSet Tag-guided Synthetic Data. *arXiv preprint arXiv:2303.07902*, 2023.

[563] Itai Yanai and Martin J Lercher. It takes two to think. *Nature Biotechnology*, 42(1): 18–19, 2024.

[564] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. *Advances in Neural Information Processing Systems*, 34:9378–9390, 2021.

[565] Karren Yang, Bryan Russell, and Justin Salamon. Telling Left From Right: Learning Spatial Correspondence of Sight and Sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020.

[566] Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*, 2023.

[567] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt

for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

[568] Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. SciCap+: A Knowledge Augmented Dataset to Study the Challenges of Scientific Figure Captioning. *arXiv preprint arXiv:2306.03491*, 2023.

[569] Matthew John Yee-King, Leon Fedden, and Mark d'Inverno. Automatic programming of VST sound synthesizers using deep networks and other techniques. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):150–159, 2018.

[570] Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models. *arXiv preprint arXiv:2305.14985*, 2023.

[571] Halley Young, Maxwell Du, and Osbert Bastani. Neurosymbolic Deep Generative Models for Sequence Data with Relational Constraints. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 37254–37266. Curran Associates, Inc., 2022.

[572] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852, 2022.

[573] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[574] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*, 2023.

[575] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021.

[576] Junhong Zhao, Hua Yuan, Wai-Kim Leung, Helen Meng, Jia Liu, and Shanhong Xia. Audiovisual synthesis of exaggerated speech for corrective feedback in computer-assisted pronunciation training. In *2013 IEEE International Conference on Acoustics,*

Speech and Signal Processing, pages 8218–8222. IEEE, 2013.

[577] Nan Zhao, Asaph Azaria, and Joseph A Paradiso. Mediated atmospheres: A multimodal mediated work environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–23, 2017.

[578] Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678. IEEE, 2021.

[579] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in Neural Information Processing Systems*, 1(January):487–495, 2014. ISSN 10495258.

[580] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*, 2023.

[581] Minghao Zhu, Xiao Lin, Ronghao Dang, Chengju Liu, and Qijun Chen. Fine-Grained Spatiotemporal Motion Alignment for Contrastive Video Representation Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4725–4736, 2023.

[582] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502, 2007.