TEXT RECAPTIONING FOR AUDIO DIFFUSION MODELS

BY

EVAN M. MATTHEWS

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2025

Urbana, Illinois

Adviser:

Professor Paris Smaragdis

ABSTRACT

Textual Inversion (TI) is one of the most recent discoveries in generative audio models that allows for prompt recaptioning of a pretrained model's concept understanding. In this thesis, we explore a rudimentary method and TI for audio sample outputs, specifically focusing on a custom-trained TI model that embeds new concepts into pretrained image models. The TI recaptioned samples are compared with baseline samples, and we find a significant difference in variation between baseline samples. We believe that this comparison will provide a strong foundation and inspiration to future works related to prompt recaptions and analysis on generative audio models.

To my grandfather, Larry Matthews, whom I dedicate my passion for music and motivation to make the world a better place.

ACKNOWLEDGMENTS

First and foremost, this thesis and additional research would not have happened without the overwhelming support I've had, both as an undergrad in CS + Music and a graduate student. I would like to thank my advisor- Professor Paris Smaragdis- for all his guidance in research and audio processing. I certainly wouldn't be the researcher I am today without your wisdom, and the lessons I've learned from our talks and efforts have strengthened my passion for the field and will carry on through my entire career. I'd also like to thank Professor Minje Kim as a lab mentor; while I'm only here for your first year teaching at UIUC, I appreciate that you spent so much time getting to know me, learn about my research interests, and provide helpful insight. And to my audio lab mates and colleagues- Krishna, Dimitris, Chris, Cameron, Jocelyn, Yutong, Jackie, Tsun-An, Jaesung, Riccardo, Jayeon, Quinn, Yiting and Yurii- you're all so incredibly kind, talented, and thoughtful. It has been such a pleasure to bounce ideas around, collaborate on projects, hang out, grab meals, and take trips with *quite literally* the most fun lab at UIUC.

Next, I'd like to thank all the friends I've made during my time in college. I am very lucky to have gone to school so close to home while meeting many passionate, creative, and talented people from all over the world. A special thanks to the Association for Computing Machinery student chapter- keep being yourself.

Last, but certainly not least, I want to thank my close friends and family for all their love and support over the years. My parents, my brother Caden, and Tegan; you've been with me through all of my highs and lows, reminding me to always take life "one step at a time." And I would not have it any other way.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 RELATED WORK 4
2.1 Diffusion Models 4
2.2 Prompt Analysis 5
2.3 Textual Inversion 6

CHAPTER 3 LITERATURE REVIEW 7
3.1 *Investigating Personalization Methods in Text to Music Generation* 7
3.2 *Applying Textual Inversion to Control and Personalize Text-to-Music Models* 9
3.3 *OpenSep: Leveraging Large Language Models with Textual Inversion for Open World Audio Separation* 10
3.4 *Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion* 12

CHAPTER 4 EXPERIMENTS AND RESEARCH METHODS 14
4.1 Research Question 14
4.2 Data 15
4.3 Prompt Recaptioning 17
4.4 Textual Inversion 18
4.5 Diffusion Generation 19

CHAPTER 5 RESULTS 21
5.1 Prompt Recaptioning 21
5.2 Textual Inversion 22
5.3 Manual Evaluations 23

CHAPTER 6 LIMITATIONS AND FUTURE WORK 28
6.1 Limitations 28
6.2 Future Work 29

CHAPTER 7 CONCLUSIONS 30

REFERENCES 31

APPENDIX A LISTENING TO AUDIO SAMPLES 33

CHAPTER 1: INTRODUCTION

In the overarching field of generative audio, one paradigm has stayed consistent despite innovations in generative models: to what extend can we *control* the output of our generative audio models? Large Language Models (LLMs) accomplish this task through training on massive text datasets to upwards of billions of parameters. In turn, recent models from ChatGPT, Claude, LLaMA and Gemini have demonstrated their ability to handle personalized tasks while maintaining a strong general conceptualization of language. This incredible feat is ultimately due to efficient storage of textual information, both as readable text and as embeddings. The same cannot be said for generative models for other modalities, such as images, audio and video.

Despite recent advancements in generative image and audio models, the extent to which we can investigate improvements is still limited. In particular, such models take exponentially longer to train than LLMs, given the amount of data in a single image or audio sample, and the amount of samples needed for a model to learn the underlying structure of the data. Likewise, training a state-of-the-art generative model in reasonable time requires an excessive amount of compute power, which is typically not available to the average researcher or model user. Given these limitations, we are left with a few options:

- Accept existing pretrained models as an acceptable solution for our generative tasks,
- Train or retrain our own models, often with nearly unrestrained time and compute power, or
- Devise methods for fine-tuning existing models to our own tasks, especially in cases where the model already fails in performance.

As the field of computer vision has shown, the third option is the most promising, as it allows for researchers to explore new tasks despite having compute limitations. Textual Inversion (TI) is one such discovery for diffusion models that has shown promise towards all generative tasks, having originally been developed to embed new concepts into pretrained image models. The result allows users to contribute new concepts to a model, conditioning a prompt to add unseen objects to a scene, or even to change the style of an image to something new. Its use as an embedding method has made it a popular generalized addition to generative models, leading to a surge in research on personalization and similar topics.

In this thesis, we choose to also delve into fine-tuning methods, specifically focusing on a rudimentary method and Textual Inversion (TI) to vary generative audio sample outputs. While it is understood that TI is capable of personalized generative tasks, we instead focus

on the extent to which we can vary the output of a generative model through text prompt recaptioning. In turn, we explore basic prompt recaptioning as a baseline and evaluation of a pretrained model's concept understanding, and conclude with TI recaptioning to find a difference in variation. We believe that this comparison will provide a strong foundation and inspiration to future works related to prompt recaptioning and analysis on generative audio models.

In chapter 2, we explore the current state of generative audio models and the technical aspects of Textual Inversion that allows for prompt recaptioning. This is by no means a comprehensive overview of generative modeling, (we choose to begin at Diffusion models), although it is highly recommended to read works on single-frame, context-dependent models that paved the way for Diffusion.

Chapter 3 continues with an in-depth literature review regarding the current state of Textual Inversion for generative audio models: Plitsis et al. [1] are the earliest to approach Textual Inversion for audio generation, and their work makes initial promises on its effectiveness for audio personalization. Thomé et al. [2], (alongside Thomé's thesis work [3]), further explores Textual Inversion for audio generation- albeit with a custom-trained TI model. Mahmud et al. [4], differing from music generation for personalization, explores the use of Textual Inversion for audio separation. Lastly, Manor and Michaeli [5], while not directly using Textual Inversion, develop a similar inversion method for editing audio samples, leaning towards an audio production perspective.

Chapter 4 details our experiments and methods for prompt recaptioning, focusing on an understanding of how much variation is possible through such methods. In particular, we explore two methods of recaptioning- basic rewording with synonymous terms TI conditioning through audio sample embeddings. We wholly expect the former method to be moderately effective, while the latter method should yield some level of variation while potentially maintaining time-based features of the sample embeddings.

Chapter 5, in turn, describes the results of our methods, both quantitative and qualitative in nature. The effectiveness of baseline and recaptioned samples are measured through embedding metrics, where a Vanilla Autoencoder (VAE) is used to encode the audio samples into a latent space. with cosine similarity and L2 distance computed on these embeddings to determine levels of variation between sample pairs. These metrics provide a general overview of our methods' effectiveness, but they don't tell the whole story. In turn, we also provide a qualitative analysis of four sets of samples, analyzing the more humanly describable features of the samples and how they change across recaptioning methods.

Chapter 6 discusses the limitations of our work, especially points of interest that we encountered during our experiments. We also note some additional perspectives from which

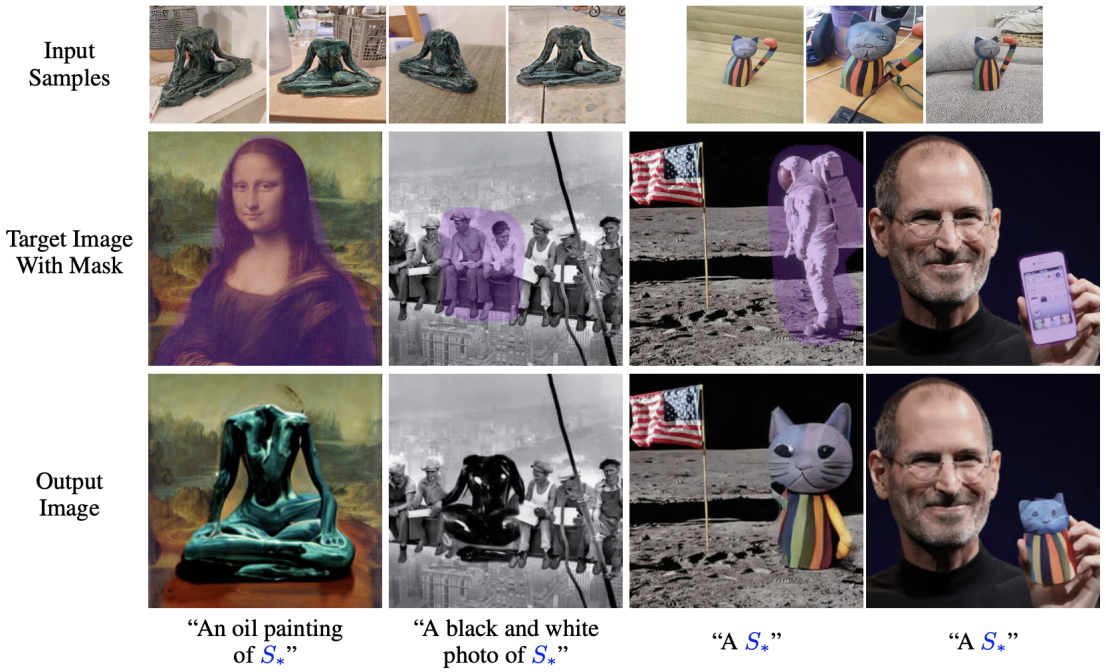

Figure 1.1: Gal et al. [6] demonstrating the use of Textual Inversion to embed completely new concepts into generative output.

our methods and results could be used for future work. As such, we seek to provide a starting point for future works to explore the depth to which variation is possible through prompt recaptioning. It should also be noted that, given the origins of Textual Inversion, our methods are not solely limited to audio generation. We hope to see works on other modalities see our efforts as an additional perspective on tasks like content personalization and editing.

CHAPTER 2: RELATED WORK

The following sections will discuss the related work that built the foundation for this project. A number of advancements in artificial intelligence- including diffusion models, textual inversion, prompt analysis, and audio generation- have shown great promise in the past five years. In particular, audio generation has seen a surge in interest, especially for the purpose of creating personalized music or speech based on vocal context. This spike in interest has also led to collaboration among researchers of similar fields to create new audio-based methods. For instance, Computer Vision is often credited as audio processing researchers borrow techniques to handle time-series data. Textual Inversion [6] is one such method that is reviewed and implemented in this project.

2.1 DIFFUSION MODELS

While multiple methods for signal generation have recently become prevalent, diffusion-based approaches still remain compelling. These models iteratively refine samples through repeated denoising steps, enabling high-quality audio generation that can capture complex temporal structures. Works starting from Ho, Jain and Abbeel [7] and onward have shown that diffusion models can be used to generate high-quality samples among complex distributions.

The diffusion model is a generative model that iteratively refines a sample through a series of denoising steps. In particular, while a complex mathematical structure can be destroyed through successive noise additions, the process is shown to be reversible given its representation as a Markov Chain of successive Gaussian noise additions.

2.1.1 Forward Step

The forward or *diffusion* step of the diffusion model is the process of adding Gaussian noise to a current sample. Suppose an initial sample $\mathbf{x}_0$ is placed through T steps of Gaussian noise-adding process q. Then

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2.1}$$

where β_t is a variance schedule of scalars that control the amount of noise added at each step.

2.1.2 Reverse Step

The reverse or *denoising* step of the diffusion model is the process of removing the added noise from a sample. This step is representative of the model's ability to generate samples from pure Gaussian noise as the process of reversing successive noise additions can be learned through training p_θ:

$$p_\theta(\mathbf{x}_{0:T}) \triangleq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \triangleq \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{2.2}$$

In this case, $p_\theta(\mathbf{x}_{0:T})$ represents a joint distribution starting from $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$, a standard Gaussian distribution, and proceeding through T steps of denoising.

The process of denoising Gaussian distributions in this manner is optimized by training the model on parameter θ via maximum likelihood estimation, (or minimizing the negative log-likelihood of the data):

$$\mathcal{L}(\theta) \triangleq \mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q \left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] \tag{2.3}$$

2.1.3 Latent Diffusion

Latent diffusion models (LDMs) [8, 9] are a recent advancement in diffusion models that allow for the generation of high-quality samples in a lower-dimensional latent space. This is achieved by using a variational autoencoder (VAE) to encode the input data into a lower-dimensional latent space, and then applying the diffusion process in this latent space. The LDM architecture consists of three main components: an encoder, a decoder, and a diffusion model.

The encoder maps the input data into a lower-dimensional latent space, while the decoder maps the latent representation back to the original data space. The diffusion model is then applied to the latent representation, allowing for efficient generation of high-quality samples.

2.2 PROMPT ANALYSIS

Prompt analysis, or prompt engineering [10], is a method of analyzing the effectiveness of text prompts in conditioning generative models. This set of methods has become increasingly popular in the past few years, alongside the rise of large language models (LLMs) and their ability to handle complex task through text input.

While the field of Natural Language Processing (NLP) initially focused on text prompting for LLMs, the domain has expanded to include models of other media such as images and audio. That is, instead of solving a range of tasks, prompt analysis has been used to analyze the effectiveness of generating a non-textual medium given text input. Image generation models were the first to expand into this domain, with models such as DALL-E and Stable Diffusion [8] leading the charge. One of the major benefits of text-to-image generation is its innate relationship between text and imagery. Human language and vocabulary is well suited to handle visual concepts- color, texture, style, shape, to name a few- and in turn, T2I models are handle complex prompts by learning the direct relationship between a word and its visual representation. The same cannot be said for audio generation, where the relationship between text and sound is not as clear.

2.3 TEXTUAL INVERSION

Textual inversion [6] is also a recent advancement in diffusion with the purpose of conditioning generation with text embeddings. This method has shown to be effective in image generation tasks, where text embeddings learned from a subset of data can be used to condition the generation of new images.

$$v_* = \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))||_2^2\right] \tag{2.4}$$

Likewise, textual inversion has been shown to be effective in audio generation tasks, where text embeddings learned from a subset of audio data can be used to condition the generation of new audio samples. The change in medium, while not trivial, is theoretically sound given the strict structure and ordering of data in both cases.

CHAPTER 3: LITERATURE REVIEW

Recent works in diffusion-based audio generation have shown the effectiveness of textual inversion as a potential staple in personalization and general output improvement. As such, their methods and results are worth investigating in the context of this project.

3.1 *INVESTIGATING PERSONALIZATION METHODS IN TEXT TO MUSIC GENERATION*

Plitsis et al. [1] investigates personalization approaches in text-to-music generative models, addressing the challenge of fine-tuning diffusion-based audio generators to capture user-specific musical styles or unique instruments. Existing text-to-music diffusion models often struggle to produce outputs that faithfully represent specific musical concepts or personalized playing styles solely through text prompts. This limitation motivates the adaptation of personalization methods that have recently proven successful in image-generation contexts, namely Textual Inversion (TI) and DreamBooth (DB), to the audio domain.

Text-to-Audio Latent diffusion models (LDMs) constitute the backbone of the authors' approach. In LDMs, a latent representation of audio is progressively denoised from random noise, guided by conditioning on textual embeddings. This denoising process is mathematically captured by the following loss function:

$$L_{\text{LDM}} = \mathbb{E}_{z\sim\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,I),y,t}\left[\|\epsilon - \hat{\epsilon}_\phi(z_t, t, c_\tau(y))\|_2^2\right] \tag{3.1}$$

Here, z_t is the latent representation of an audio sample x noised at timestep t, ϵ represents the sample noise drawn from a Gaussian distribution, $\hat{\epsilon}_\phi$ denotes the denoising network with parameters ϕ, and $c_\tau(y)$ encodes the conditioning provided by the textual prompt y, encoded by $c_\tau(y)$.

To implement personalization, the authors modify the parameter space of the LDM. Specifically, they introduce a novel embedding vector v^* corresponding to a placeholder token S^* representing some new user-specific musical concept. For Textual Inversion, only the new embedding is optimized, while the rest of the model parameters are frozen. DreamBooth, in contrast, fine-tunes all weights of the denoising network ϕ.

Diving deeper into audio personalization, Plitsis et al. additionally explore personalized style transfer. This approach involves partially noising a latent representation z_t derived from a given source audio clip x_{in} followed by conditioning the reverse diffusion process on the personalized concept. Mathematically, this personalized style transfer is defined by

$$p_{\theta'}(z_{0:t}|c(y)) = p(z_t) \prod_{n=1}^{t} p_{\theta'}(z_{n-1}|z_n, c(y)) \tag{3.2}$$

where $c(y)$ encodes the personalized concept from text, and the timestep t denotes the "transfer strength" of the concept onto x_{in}

To evaluate the proposed methods, the authors create a new dataset comprising 32 distinct musical concepts, each represented by five audio clips spanning percussion instruments, melodic solos, and multi-instrument pieces across diverse musical traditions. Evaluation involves both quantitative and qualitative measures. For quantitative evaluation, embedding-based similarity metrics like CLAP-A (audio similarity), CLAP-T (text similarity) [11], and FAD (Fréchet Audio Distance) [12] are used. Music-specific metrics measuring rhythmic similarity (BPM), harmonic properties (key and scale), and loudness (EBU R128 scale) are also employed, calculated using the Essentia audio toolkit.

Experimental results indicate that DreamBooth significantly outperforms Textual Inversion in terms of audio reconstruction accuracy (higher CLAP-A, lower FAD) and textual editability (higher CLAP-T). Training on multiple audio samples (three versus one) notably improves performance. Data augmentation—specifically mixing training audio with environmental noises—further enhances DreamBooth performance but provides limited improvement for Textual Inversion. Interestingly, the authors observe a trade-off resembling a "Pareto front," indicating a tension between accurate audio reconstruction and textual manipulation capability.

A human preference study reinforces these quantitative results, showing clear preferences for the methods decided over default generation. Specifically, 58% of participants preferred DreamBooth's reconstructions, with 24% favoring Textual Inversion. Regarding textual editability, DreamBooth still led (37%) compared to Textual Inversion (31%), but a significant portion (24%) expressed no clear preference, underscoring ongoing challenges in providing subjectively improved audio across a large audience.

In further music-specific evaluations, it's revealed that DreamBooth effectively captures rhythmic elements (BPM) but struggles with precise harmonic reconstruction and matching loudness to reference samples. Textual Inversion, in turn, generally underperforms across musical dimensions, notably rhythmic features.

Lastly, personalized style-transfer experiments demonstrate that adjusting transfer strength allows the generated audio to gradually adopt characteristics of the target personalized concept, identifying an optimal transfer range $t \in [0.4, 0.6]$ for balancing source retention with stylistic transformation.

3.2 *APPLYING TEXTUAL INVERSION TO CONTROL AND PERSONALIZE TEXT-TO-MUSIC MODELS*

Thomé et al. [2] explore an innovative approach to enhancing the controllability and personalization capabilities of Text-to-Music (TTM) models. With the rapid evolution of models such as MusicGen, MusicLM, AudioLDM, and Stable Audio, synthesizing realistic musical audio directly from textual descriptions has become increasingly feasible. Despite this, existing models typically struggle with accurately producing concepts they have not explicitly encountered during training, presenting significant limitations for creative and personalized applications. To address this, the authors propose applying Textual Inversion to MusicGen [13].

Textual Inversion (TI) facilitates the introduction of novel concepts into pretrained models without exhaustive retraining or modification of existing parameters, thereby reducing computational overhead and avoiding catastrophic forgetting. More specifically, the authors utilize TI by augmenting the text embedding matrix of MusicGen's pretrained T5-based text encoder with additional concept embeddings. These embeddings are trained using gradient descent optimization, guided by synthetic text prompts referencing the newly introduced concepts paired with corresponding audio examples.

Formally, the authors define the optimization for MusicGen in terms of minimizing categorical cross-entropy loss, written as

$$v^* = \arg\min_v \sum_{k=1}^{K} C_k(\sigma(A(x)), M(A(x), T(y; v))) \tag{3.3}$$

where v^* represents the optimized embedding parameters for the new concept, C_k denotes the cross-entropy loss for the kth codebook from MusicGen's RVQ-VAE, σ symbolizes the autoregressive token shifting operation, and $T(y; v)$ encodes the textual prompts containing the novel embeddings. The loss function captures the alignment between predicted and actual audio tokens, ensuring the model accurately integrates the new concept.

Additionally, the authors detail the training process involving stochastic gradient descent with the AdamW optimizer and an exponential moving average applied post-training to stabilize embedding weights. The process effectively restricts the model update to only newly-introduced embedding parameters, thus maintaining the integrity of the pre-existing learned representations and avoiding having to fine-tune a large-scale model.

Evaluation of TI effectiveness combines both objective metrics and subjective human judgment. The authors use CLAP embeddings to quantify similarity scores between synthesized

audio outputs and both their textual prompts and original audio references. According to their reported results, MusicGen demonstrates promising performance, particularly excelling in the "editability" metric which reflects on the model's ability to blend novel concepts effectively into diverse musical contexts.

Following objective metrics, Thomé et al. complement quantitative evaluation with a human listening study involving twenty six participants who evaluated synthesized audio clips based on how accurately they matched provided textual descriptions. Participants frequently favored MusicGen-generated outputs in editability scenarios—with prompts like "A disco song with a S*"— while preferring outputs from diffusion-based methods like AudioLDM for direct concept reconstruction tasks (e.g., "A recording of S*"). These findings indicate that MusicGen's autoregressive structure promotes diverse musical interpretations, whereas diffusion-based models tend to strictly adhere to reference audio structures.

The authors also combine their observed performance differences with theoretical reasoning. MusicGen predicts audio tokens in an autoregressive manner, conditioned via cross-attention with textual embeddings, inherently introducing variation and spontaneity in musical generation. By contrast, diffusion-based models refine entire audio sequences iteratively, ensuring strict adherence to reference audio characteristics. Thomé et al. highlight this distinction by referring to the underlying mathematics of diffusion methods (originally shown by Ruiz, et al. [14]) versus their previous autoregressive training, suggesting that each approach can handle specific music generation tasks more effectively.

Despite their promising results, the authors acknowledge some limitations in their methodology. The scope of the evaluation remains preliminary, given the modest sample size and relatively narrow set of test concepts. They propose extending future investigations to include broader musical styles, diverse instrumentation, and more comprehensive user evaluations to generalize findings robustly. They also advocate for a standardized benchmark and evaluation framework within the TTM research community, noting the current challenges due to limited model accessibility and varied evaluation methods across studies.

3.3 *OPENSEP: LEVERAGING LARGE LANGUAGE MODELS WITH TEXTUAL INVERSION FOR OPEN WORLD AUDIO SEPARATION*

Mahmud, et. al. [4] proposed a method for audio separation using textual inversion. Their method, *OpenSep*, was able to separate audio sources with around a 90% accuracy rate, with performance drops of only 8% and 15%.

The process of audio source separation in real-world scenarios is confronted with signif-

icant challenges due to the inherent complexity of mixtures, which frequently comprise a variable and unknown number of sound sources. Conventional audio separation techniques face limitations such as over-separation, where additional irrelevant audio sources are extracted; under-separation, resulting in incomplete extraction; and heavy dependence on pre-established training sources. These limitations severely constrain the adaptability and efficacy of existing methods in practical, real-world scenarios.

In addressing these substantial challenges, the authors introduce a novel and sophisticated framework named *OpenSep*. This innovative method leverages the advanced capabilities of large language models (LLMs) to achieve automated audio separation without human intervention or the restriction to predefined audio classes. Unlike prior conditional audio separators which necessitate manually crafted prompts, OpenSep's approach capitalizes on textual inversion, which employs an off-the-shelf audio captioning model, specifically ms-CLAP [15, 16], to generate meaningful text captions from audio mixtures. These generated captions serve as comprehensive semantic representations, capturing the salient features of audio mixtures and facilitating the subsequent semantic parsing of individual sound sources.

The textual inversion component in *OpenSep* notably simplifies the otherwise complex challenge of parsing multiple audio sources from noisy mixtures. It transforms the acoustic information into descriptive textual forms, thus bridging audio and textual domains. The semantic parsing of these textual captions is executed through few-shot prompting with instruction-tuned large language models, such as LLaMA-3-8b. These models parse each caption into distinct sources, extracting comprehensive audio characteristics crucial for effective separation. This enriched semantic parsing goes beyond conventional single-class identification, encompassing detailed properties such as frequency range, amplitude dynamics, spectral shape, and duration—attributes that significantly enhance the granularity and specificity of audio separation tasks.

Subsequently, the audio separation itself is performed by a text-conditioned audio separator. This separator employs detailed textual information parsed from the captions to condition the separation process. Unlike traditional conditional separators reliant solely on simple class labels or manual text inputs, this model utilizes enriched prompts to greatly enhance performance. This is a particular advantage when handling unseen audio sources or complex, noisy audio mixtures.

A central innovation of *OpenSep* is its advanced multi-level extension of the standard mix-and-separate training methodology. This technique simultaneously tackles both lower-order mixtures—such as pairs of sound sources—and individual sources derived from higher-order mixtures, employing a two-level separation training objective. The training is optimized through an L1 loss function

$$\mathcal{L}_{\text{sep}} = \sum_i \|\hat{X}_i - X_i\|_1 \tag{3.4}$$

where $\hat{X}_i$ represents the predicted magnitude spectrogram and X_i denotes the corresponding ground truth spectrogram. This sophisticated training approach fosters tighter alignment between the extracted audio components and their corresponding textual representations, significantly boosting the model's robustness and precision in handling complex separation tasks.

Empirical validation of *OpenSep* reveals notable performance enhancements over contemporary state-of-the-art audio separation methodologies. Specifically, experimental evaluations demonstrate substantial improvements measured through Signal-to-Distortion Ratio (SDR) and Signal-to-Interference Ratio (SIR)—metrics that respectively quantify overall separation quality and residual interference among extracted audio sources. *OpenSep* surpasses baseline approaches, achieving remarkable SDR improvements of approximately 64% on unseen classes within the MUSIC dataset and up to 180% on the VGGSound dataset. Such advancements underscore the efficacy and potential real-world applicability of the method's framework.

User-based qualitative evaluations reinforce quantitative findings, showcasing the clear superiority of *OpenSep* in handling natural, realistic audio mixtures even without manual prompts. These evaluations validate the method's practical utility and demonstrate how Textual Inversion can be leveraged to enhance the performance of other tasks.

Despite its impressive performance, *OpenSep* has several limitations: it has a heavy reliance on accurate audio captioning models; it is computationally complex, which comes from having to integrate multiple advanced neural models; and issues still arise when attempting to separate brief or subtle sounds from highly noisy mixtures. Nonetheless, the authors suggest feasible paths for optimization, including the adoption of lighter-weight mobile-friendly LLMs such as Phi-3-mini or Gemma-2b, to enhance efficiency and reduce computational demands.

3.4 *ZERO-SHOT UNSUPERVISED AND TEXT-BASED AUDIO EDITING USING DDPM INVERSION*

Manor and Michaeli [5] present an innovative study on zero-shot audio editing, employing denoising diffusion probabilistic models (DDPM) [7]. Their work introduces two core methodologies: the ZEro-shot Text-based Audio (ZETA) editing approach, adapted from image editing, and the novel Zero-shot UnSupervised (ZEUS) editing method. Both meth-

ods allow detailed editing of audio signals without requiring retraining or supervised data, marking significant progress in audio manipulation, particularly music editing.

The technical basis of their approach involves DDPM inversion. DDPMs generate samples through iterative denoising starting from Gaussian noise $x_T \sim \mathcal{N}(0, I)$, progressively denoising through:

$$x_{t-1} = \mu_t(x_t) + \sigma_t z_t, \quad t = T, \ldots, 1 \tag{3.5}$$

where z_t are standard Gaussian vectors, and $\mu_t(x_t)$ is derived from the MSE-optimal prediction $\hat{x}_{0|t}$. The inversion extracts noise vectors $\{x_T, z_T, \ldots, z_1\}$ via:

$$z_t = \frac{x_{t-1} - \mu_t(x_t)}{\sigma_t}, \quad t = T, \ldots, 1 \tag{3.6}$$

These vectors retain the global audio structure, serving as the foundation for subsequent editing.

In the ZETA method, editing leverages text prompts. After extracting noise vectors using DDPM inversion with a source prompt, a different text prompt guides the subsequent denoising process. The method balances fidelity and target adherence via classifier-free guidance strength and starting timestep T_{start}.

The ZEUS approach brilliantly identifies semantic editing directions through principal components (PCs) of the posterior covariance matrix $\mathrm{Cov}[x_0|x_t]$. These PCs, $\{v_i|_{t'}\}$, are efficiently computed via a subspace iteration method. Perturbations are introduced as:

$$x_{t-1} = \mu_t(x_t) + \gamma c_t \lambda_{i|t}^{1/2} v_{i|t'} + \sigma_t z_t, \quad t = T, \ldots, 1 \tag{3.7}$$

where γ controls perturbation strength, and c_t is a timestep-dependent correction factor. These perturbations facilitate semantically meaningful audio variations, such as melody improvisations or instrument emphasis shifts, without textual input. Experimental validation using metrics like CLAP [11, 15, 16], LPAPS, and Fréchet Audio Distance (FAD) [12] confirms that their methods consistently outperform baselines (e.g., MusicGen [13], SDEdit, DDIM inversion). User studies further reinforce the effectiveness and preference for these methods in achieving semantic edits while preserving the original audio essence.

Overall, this study represents a substantial leap in audio editing research by successfully translating DDPM inversion methods to audio, along with pioneering an unsupervised editing strategy. It sets the stage for future development in highly expressive audio editing tools, beneficial in areas from music production to broader multimedia applications.

CHAPTER 4: EXPERIMENTS AND RESEARCH METHODS

The following chapter discusses the experiments and research methods for evaluating text prompts and corresponding audio outputs from Text-to-Audio (TTA) models. As a whole, this focus relies on the use of prompt analysis for text and metric evaluation for audio akin to sample-based sound similarity or "sound variance." That is, should methods which vary the structure of text prompts lead to significant variance in audio outputs, then it can be stated that a generative audio model is capable of producing a larger variety of outputs than expected. In return, this conclusion works as evidence towards further analysis into the connection between text prompts and audio outputs.

Text prompts are objectively preferred methods for conditioning audio generation with current models. As an ordered set of words or tokens, this input represents the most closest representation of human ideas. Likewise, sound is large difficult to describe in typical formats, especially when meaning and influence in audio is subjective and/or impacted by outside context and experience. The use of text prompts as input is expected to remain common practice, and in turn, it is important to continue to analyze the current effectiveness of text prompts in conditioning for generative audio samples.

Additionally, it should be noted that subjective evaluation of audio files is a common method of evaluation when the overall quality of sound is in question. While we recognize this evaluation as a valid approach, the experiments described will not prioritize user studies for results. Research in audio processing often relies on subjective evaluation, especially in cases where the data used is narrowed into a specific domain, (music or speech, genre, instruments used, etc.) However, with our experiments including a large domain of audio samples and being focused on quantitative differences in generative output, quality metrics cannot be seen as the sole source of evaluation. In turn, our results rely on a combination of features: batch-computed metrics for general analysis, and a handful of manual evaluation to more finely review specific cases of prompt variety.

4.1 RESEARCH QUESTION

For the purpose of collecting various experimental methods into a single conclusive response, our research question is:

Can methods of text recaptioning be used to make
significant changes to audio generation in diffusion models?

The goal of this question is to determine if the text prompts used in generative audio models are capable of producing a large variety of outputs, and if so, how much variance can be expected from slight changes in the text prompts. This question sets the stage for a combination of varied methods for evaluation, including basic prompt rewording, LLM text recaptioning, and textual inversion in the same context. Results reliant for answering this question will be based on measured "distances" between audio samples generated from the same diffusion model under different text prompts, as well as between the text prompts themselves.

The following sections will discuss details about the data used, the methods used to generate the audio samples, and the evaluation metrics used to analyze the results. The first section will discuss the data used in our experiments, including the parameters used in audio and text prompt generation. The second section will discuss our method of basic prompt recaptioning, which is later used as an overall comparison towards the effectiveness of Textual Inversion (TI). Likewise, the next section will discuss the use of TI for generating varied audio samples. Lastly, the final section will discuss the diffusion model used for generating audio samples, including its limitations and expected results.

4.2 DATA

The data used in the aforementioned experiments consists of large sets of generated text prompts along with their corresponding audio samples.

The text prompts are generated using a large language model (LLM), specifically Claude 3.7 Sonnet [17], which is queried to generate a list of 1000 unique text prompts[1]. The prompts themselves are generated with a focus on their usage as input for a generative audio model, and as such, provide descriptively vague notions of the context to be generated. More specifically, mentions of real-world locations, objects, people, and other named entities are avoid to improve generation quality.

All of the audio samples evaluated in this project are generated using Stable Audio Open 1.0 [18], a high-quality TTA latent diffusion model.

Lastly, prompt generation also included a couple unexpected but welcome attributes- a defined "category" for each prompt (Figure 4.1), and a "complexity" rating (one of "low", "medium", or "high"), (Figure 4.2). While this data is not important to the evaluation of the audio samples, it does provide an interesting insight into Claude 3.7 Sonnet's training data. More specifically, the model appears to frequently engage in more generalized categories

[1]ChatGPT models 4o and 4.5 were also queried for prompt generation, but their outputs ran into hallucination issues, (a large subset of prompts would repeat around every 150 prompts.)

Prompt ID	**Baseline**	**Reception, Batch 1**	**Reception, Batch 2**
164	Purring cat sleeping on a windowsill during rain.	Purring cat sleeping above a windowsill amid rain.	Droning tabby slumbering above a window edge in the middle of rain.
341	Distant ambulance siren approaching.	Detached ambulance siren approaching.	Detached ambulance siren advancing.
677	Piano being tuned with key hits.	Piano transforming into tuned comprising key hits.	Piano transforming into tuned comprising key hits.
695	Pouring cereal into empty bowl.	Pouring cereal moving to empty bowl.	Pouring cereal moving mellow to empty bowl.

Table 4.1: Selection of text prompts generated and reworded by Claude 3.7 Sonnet

of sound, whereas more specific categories aren't common. This consequence of prompt generation is noted for future work, as the following pattern may also be present in Stable Audio Open 1.0 in the form of bias towards "better learned" categories of sound.

For the purpose of generalizing our methods, a large domain of audio samples will be accepted, including but not limited to:

- Environmental sounds (e.g., ocean waves, rain, forest sounds)
- Music (various genres, instruments, etc.)
- Other sound effects (e.g., mechanical sounds, animal noises)

Human speech samples are strictly excluded from the dataset, given that the audio generative model used lacks support for this case, (see 4.4).

The research question posed also allows for a broad range of methods as candidates for evaluation. In our case, we evaluate large batches of generated audio samples divided into three groups: a baseline or "ground truth" set, a "recaptioned" set, and a "textually inverted" set. The latter two sets are more specifically defined by their respective methods of generation, provided later in this chapter. Lastly, the baseline set is generated from a set of initial LLM-generated text prompts, with a focus on having unique and varied prompts between each prompt in the set.

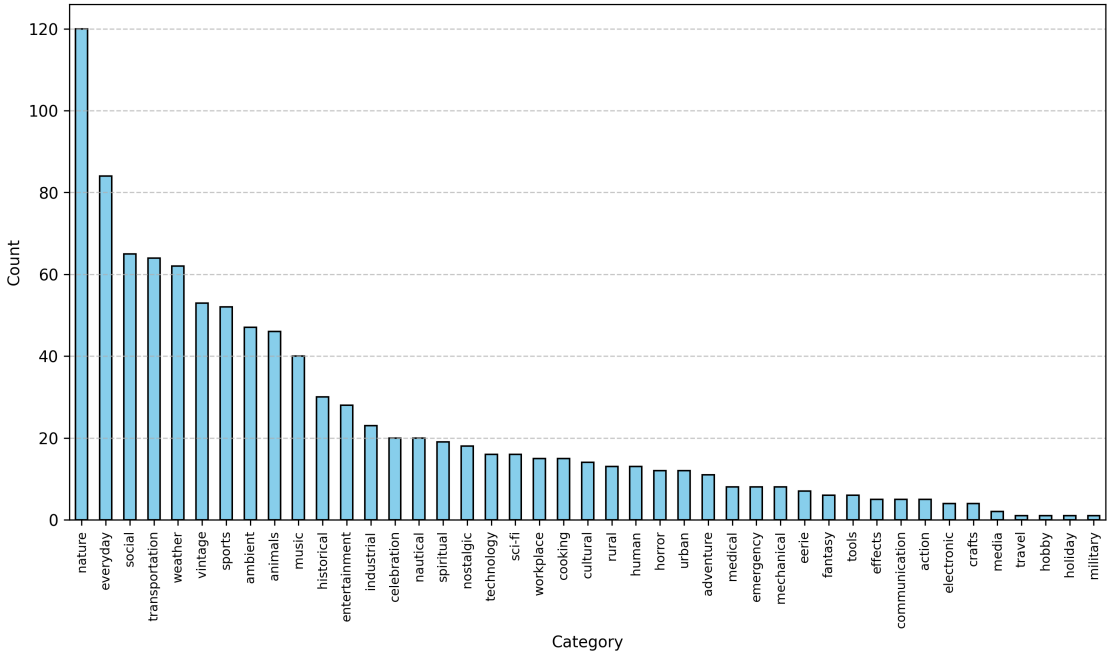

Figure 4.1: Distribution of text prompt categories

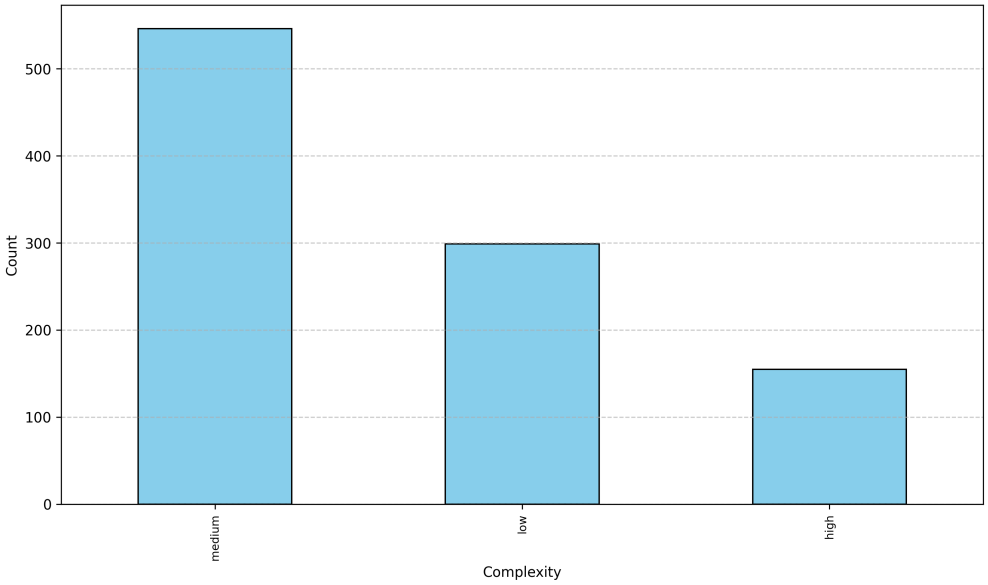

Figure 4.2: Distribution of text prompt complexities

4.3 PROMPT RECAPTIONING

One of the simplest methods to evaluate the effectiveness of text prompts is to analyze the variance in audio outputs from moderate rewording of the original text prompt. In order to

recaption a large quantity of text prompts, we use Claude 3.7 Sonnet [17], a large language model (LLM), to generate a set of unique text prompts with respect to the baseline prompts and the new set.

4.3.1 Methodology

The methodology for our prompt rewording experiment abides by the following steps:

- Select a base text prompt (e.g., "A calm ocean wave sound").
- Create multiple variations of this prompt (e.g., "A serene sound of ocean waves", "Gentle ocean wave noises").
- Generate multiple audio samples from each prompt.
- Evaluate the audio samples:
  - Between samples of the same prompt (*generation variance*).
  - Between samples of different prompts (*prompt variance*).

4.3.2 Evaluation Metrics

In order to evaluate the effectiveness of basic prompt recaptioning, we use a combination of audio and text similarity metrics. Both modalities are evaluated by encoding each into embeddings, where Stable Audio Open's internal Vanilla Autoencoder (VAE) is used for encoding to the same latent space. For the audio samples, we use a combination of cosine similarity and L2 distance to measure the "distance" between sample embeddings. For the purpose of this experiment, cosine similarity is expected to decrease as the samples become more dissimilar, while L2 distance is expected to increase. Likewise, the text prompts are also encoded and evaluated using cosine similarity in order to measure the embedding distance between the text prompts. In this case, a significant distance between prompt embeddings indicates a move away from the baseline prompt's original context.

4.4 TEXTUAL INVERSION

Textual inversion [6] is a powerful method for conditioning diffusion models on text embeddings, allowing for the generation of personalized audio samples based on learned embeddings from a small set of data. This section will discuss our use of textual inversion for prompt

recaptioning with respect to extracted embeddings, thus allowing for a more personalized or improved variance of audio outputs.

4.4.1 Methodology

- Train a diffusion model on a small set of audio samples with corresponding text prompts.
- Use textual inversion, a feature extraction method, to learn an embedding for a specific concept (e.g., "ocean wave sound").
- Generate new audio samples using the learned embedding
- Evaluate the audio samples:
  - when the input is identical (reconstruction error)
  - when the input is varied (generation variance)

4.4.2 Evaluation Metrics

Similar to the basic prompt recaptioning method, the textual inversion method is evaluated by computing embedding distances between pairs of audio sample embeddings. Cosine similarity and L2 distance are computed on three batches of audio pairs, again using an internal VAE to encode the audio samples into a latent space. Additionally, to verify the cause of variation in our method, we compute the reconstruction error from inputting three batches of audio samples. This error is computed using Mean-squared error (MSE) between the original audio sample and the reconstructed audio sample. In our case, the reconstruction error is expected to be significantly low, meaning that the resulting sample variance is not due to the model's autoencoder.

4.5 DIFFUSION GENERATION

The diffusion model used for this experiment is Stable Audio Open 1.0 [18], an advanced generative audio model designed to produce high-quality stereo audio clips directly from textual descriptions. Capable of generating audio segments of up to 47 seconds at a 44.1 kHz sampling rate, its architecture integrates three core components. First, it employs an autoencoder to efficiently compress audio waveforms into manageable latent representations, using convolutional blocks, ResNet-like layers, and Snake activation functions. Textual

prompts are converted into guiding embeddings through a T5-based model, which interfaces seamlessly with the autoencoder. Audio synthesis itself is performed by a transformer-based diffusion model (DiT), iteratively refining audio output in the autoencoder's latent space to produce coherent and realistic audio samples.

Stable Audio Open 1.0 was trained on a meticulously curated dataset comprising nearly 500,000 audio files sourced primarily from Freesound and the Free Music Archive, all licensed under Creative Commons. Rigorous verification methods, including audio classifiers and content detection services, ensured the absence of copyrighted materials. Consequently, the model excels in generating realistic sound effects, ambient soundscapes, and musical samples suitable for film, television, and music production.

However, Stable Audio Open 1.0 does have several limitations. It notably struggles with intelligible speech synthesis and realistic vocal generation, making it less suitable for tasks requiring nuanced voice synthesis. Complex textual prompts, especially those involving multiple elements or actions described using conjunctions such as "and" or "while," frequently result in incomplete or inaccurate audio representations. Additionally, the model primarily supports English, limiting performance in non-English prompts. Its musical composition capabilities, while effective for short loops and riffs, are not robust enough to handle full compositions or sophisticated musical structures. Finally, given the nature of its training data, the model may not equally represent all musical styles and cultural contexts.

CHAPTER 5: RESULTS

The following chapter provides results with respect to aforementioned experiments and research methods. As mentioned in Section 4.2, the results consist of evaluations on cosine similarity and L2 distance between relative pairs of audio samples, as well as mean-squared error (MSE) and L2 distance on reconstructions resulting from our Textual Inversion (TI) generative method. For the purpose of bounding our results, we designate cosine similarities below 1.0 and decreasing, and L2 distances above 0 and increasing, as *differing* from the baseline. Likewise, decreasing L2 distances and MSE values towards 0.0 are considered *similar* to the baseline when evaluating reconstructions from TI.

5.1 PROMPT RECAPTIONING

The results of the basic prompt recaptioning experiment, (Table 5.1) show that resulting samples produce reasonably different results from the baseline samples. In particular, batch 2 of the recaptioned samples demonstrate a greater amount of diversity than batch 1. This trend appears to originate from the LLM-based recaptioning process, as some recaptioned batch 1 prompts appear near-exact to the baseline prompts, (see Section 7.1.1) Regardless, a non-zero level of diversity is expected and subsequently observed across all recaptioned batches with respect to our synonym-based recaptioning method.

5.1.1 Audio Sample Similarity

The cosine similarity and L2 distance between the baseline and recaptioned batches are shown in Table 5.1. These results, calculated from the raw audio samples, indicate that the recaptioned batches are more similar to each other than to the baseline samples, as expected.

Cosine Similarity (↓)	**Average**	**Std. Deviation**	**Variance**
Baseline $\stackrel{?}{=}$ Batch 1	0.844	0.192	0.037
Baseline $\stackrel{?}{=}$ Batch 2	0.744	0.216	0.046
L2 Distance (↑)			
Baseline $\stackrel{?}{=}$ Batch 1	0.430	0.355	0.126
Baseline $\stackrel{?}{=}$ Batch 2	0.650	0.299	0.090

Table 5.1: Cosine Similarity and L2 Distances between baseline and recaptioned batches.

5.1.2 Text Prompt Similarity

Similar to the sample similarity calculations, we also calculated the cosine similarity and L2 distance with respect to the embeddings of a pair of text prompts. This was done to evaluate the effectiveness of our recaptioning method, as well as to provide a more general understanding of the differences between the baseline and recaptioned prompts. The results of this experiment are shown in Table 5.2, where an average level of similarity is expected and observed between the baseline and recaptioned prompts.

Cosine Similarity (↓)	**Average**	**Std. Deviation**	**Variance**
Baseline $\stackrel{?}{=}$ Batch 1	0.916	0.079	0.006
Baseline $\stackrel{?}{=}$ Batch 2	0.845	0.090	0.006

Table 5.2: Cosine similarity between baseline and recaptioned text prompts.

In our case, the recaptioned prompts have a small amount of variance, indicating a level of diversity in prompt recaptioning that, through manual evaluation, doesn't impact the prompt's expected concept.

5.2 TEXTUAL INVERSION

Similar to the basic recaptioning experiment, comparisons are made between the baseline and recaptioned batches of audio samples. In contrast, this method is evaluated on three batches, (baseline, recaptioned batch 1, and recaptioned batch 2 prompts), and additional evaluations are performed to determine the reconstruction error from the TI method. The results of the cosine similarity and L2 distance between the baseline and recaptioned batches are shown in Table 5.3, and in comparison to Section 5.1, we observe a significant decrease in cosine similarity and increase in L2 distance between batches.

In considering that the text prompts stay the same while including conditional embeddings, the following results demonstrate the overall effectiveness of the TI method in generating varied audio samples on a concept.

5.2.1 Reconstruction Error

In order to confirm that diversity in generated samples results from TI conditioning, we also evaluate the method's ability to unconditionally reconstruct input samples. This is done by generating audio samples from the TI model using the same text prompt batches, but

Cosine Similarity (↓)	**Average**	**Std. Deviation**	**Variance**
Batch 1	0.538	0.202	0.041
Batch 2	0.520	0.208	0.043
Batch 3	0.507	0.209	0.044
L2 Distance (↑)			
Batch 1	0.937	0.215	0.046
Batch 2	0.955	0.220	0.048
Batch 3	0.968	0.219	0.048

Table 5.3: Cosine Similarity and L2 Distances between baseline and TI-recaptioned batches.

the conditional embeddings are removed. The results of this experiment are shown in Table 5.4, where we observe a relatively low reconstruction error.

Reconstruction Average	**MSE** (↓)	**L2 Distance** (↑)
Baseline	0.005	46.180
Reword, Batch 1	0.005	46.398
Reword, Batch 2	0.004	42.720

Table 5.4: Reconstruction error (MSE) and L2 distance on generated audio samples from Section 5.1.

While this finding is expected, we curiously note that the L2 distances between the reconstructed samples and the baseline samples are significantly high in comparison to other evaluations. It is believed that this difference comes from a formatting discrepancy between the compared samples, as the MSE is verified through manual evaluation of the reconstructed samples.

5.3 MANUAL EVALUATIONS

While results of the aforementioned methods dictate a quantitative effectiveness of prompt rewording, subjective evaluation still remains as a means of demonstrating finer impact and benefits from a listener's perspective. As such, we performed manual evaluations on recaptioned samples to determine the qualitative effectiveness of our experiments. The results of this experiment are shown in Figures 5.1, 5.2, 5.3, and 5.4, where a significant amount of diversity in the recaptioned samples is observed with respect to the computed averages. Each prompt is labeled "Prompt X-Y", where X is the prompt ID ($X \in [0, 1000)$) and Y is the version number ($Y \in [0, 3)$). As an example, "Prompt 135-2" refers to the

2nd version of the 135th prompt (zero-indexed) in the dataset. We also provide prompts with a "TI" prefix, indicating that the samples were generated in the Textual Inversion experiment. Lastly, the text prompts for each recaptioned sample are also provided below the spectrograms of each sample, and the spectrograms have been clipped to 100-150 bins[1] to make key features more visible.

5.3.1 Prompt 30-0

In considering qualitative differences between the baseline and recaptioned samples, specific features immediately stand out in particular cases. The first sample (Figure 5.1) is a good initial example containing both consistent and inconsistent features directly related to prompt words. Every sample of this set contains a similar "bubbling brook"- a delicate stream of water flowing over rocks. Each version features near-exact non-rhythmic bubbling sounds, mind for the baseline's brook being decently louder.

On the other hand, the "birds singing in the background" incur distinct differences from recaptioning: the baseline features the birds chirping at a close proximity; the first recaptioned sample perceives the birds making shorter chirps from farther away; and the second recaptioned sample makes the birds nearly inaudible. This sound transition with respect to wording demonstrates a learned perception of distance in the prompt's context, given how the sound's perceived distance correlates with the description of the birds and their actions becoming more vague. This feature is also observable in the spectrograms, the chirp sounds in the higher frequencies fade away or disappear entirely in the recaptioned samples.

5.3.2 Prompt 1-1

Prompt 1-1, in comparison to the previous example, demonstrates a more drastic change in samples due from the recaptioning process. The overall sound can be split into two features: transportation sounds, and city street ambiance. For the first sound, the distinct car horns are replaced by recaptioning with "conveyance [2] warning sounds," indicating a contextual change to public transportation as opposed to personal vehicles. In particular, the speaking voice in the recaptioned samples has aesthetic characteristics of a clear-toned narrator, providing instructions and information to passengers. This difference clearly draws from synonymous meanings for crowded, noisy urban environments, and it appears that the

[1] each spectrogram is generated with 512 rows or "bins" in total.

[2] Conveyance is defined as the action or process of transporting someone or something from one place to another by Oxford Languages.

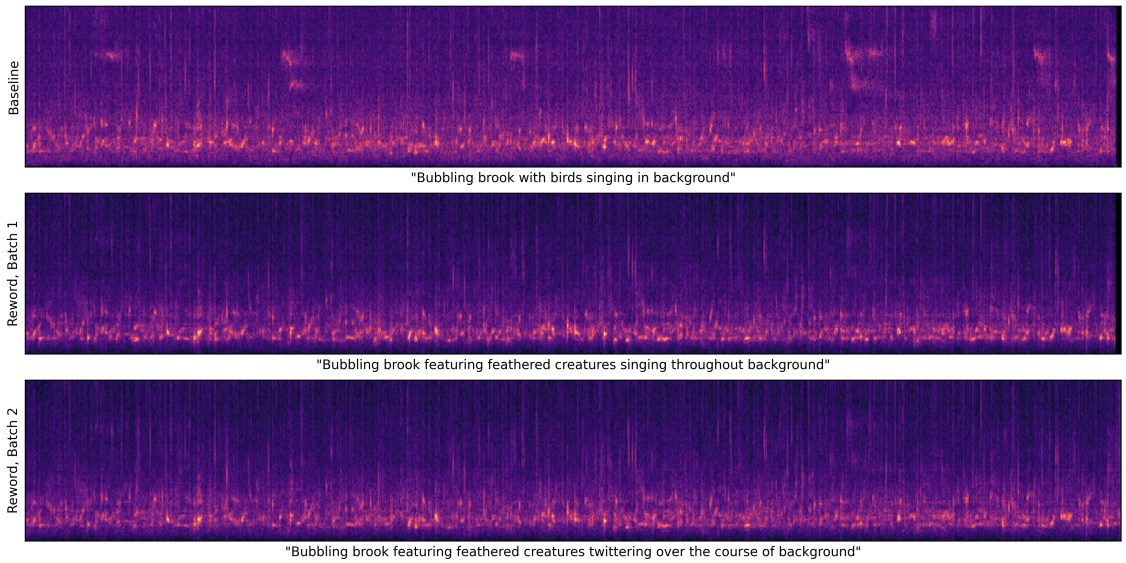

Figure 5.1: Basic recaptioning results, prompt 30-0.

recaptioning identified the wording as representing a far denser city environment than the baseline.

The second feature, the ambiance, also articulates the change to a more crowded and technologically dependent environment. As with the previous feature, the recaptioned samples contain more pronounced industrial sounds that are prominent in modern cities. The sounds of people moving about are also more pronounced, (although this may also be audible in the baseline given quieter car horns.)

Lastly, while some form of "people talking" is present in each prompt, the generative audio model appeared to disregard this feature in favor of louder sounds. Our firm belief is that, in lieu of comprehensible speech, the model fails to generate this feature when its not the most prominent sound in the prompt. And although this feature fails to be generated in the recaptioned samples, the results are still more diverse and interesting soundscapes than the baseline.

5.3.3 TI Prompt 8-2

Introducing results from our Textual Inversion experiment, we observe a trend of changes to fine details while leaving the overall sound intact. TI Prompt 8-2 is an immediate example of this point, as key features line up across all samples, but specific details are altered. Specifically, every remains relatively similar except for the "squeaky floorboards" which

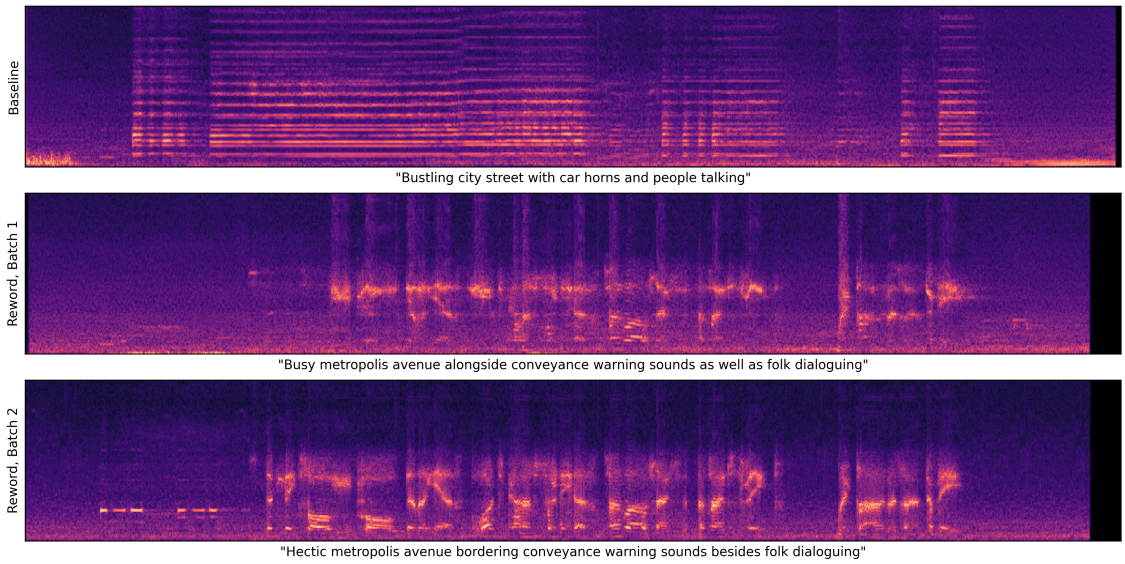

Figure 5.2: Basic recaptioning results, prompt 1-1.

present three separate versions. The recaptioning of the "floorboards" feature is solely responsible for these differences; each version, (creaking floorboards, creaking wooden floors, squeaking timber floor surface), results in a different set of frequencies and slightly different timbres. Visually, this claim is supported by the varied position of bright lines in the spectrograms, an indication of the sound's varied frequency and relative harmonics. The other feature, the "old abandoned house," while appearing unused in the samples, might potentially support the floorboard sounds in generation.

5.3.4 TI Prompt 4-1

In our final example, we observe a peculiar edge case- the recaptioned samples appear *drastically* different from the baseline. The first two samples are relatively similar, but the third sample is of a completely different musical sound. An immediate point of interest is the recaptioning of "chimes" to "song tubes" and "swaying/wavering" to "vibrating." While a context could exist where these words are synonymous and interchangeable, we find the following to be the result of mistranslation between the prompt-generating LLM and the generative model. As such, the baseline an first recaptioned samples correctly replicate the sound of chimes, but the third sample appears to take "song tubes" to be a more tonal instrument. Despite this not being the intended outcome, the resulting sounds demonstrate the volatility of generative audio with respect to small changes in wording.

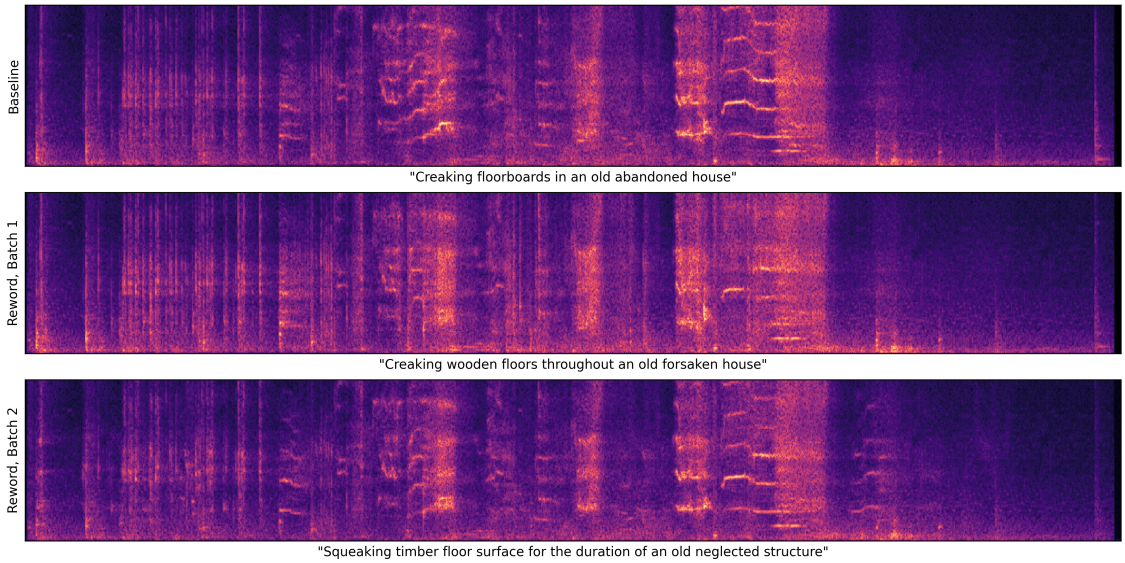

Figure 5.3: Textual Inversion recaptioning results, TI prompt 8-2.

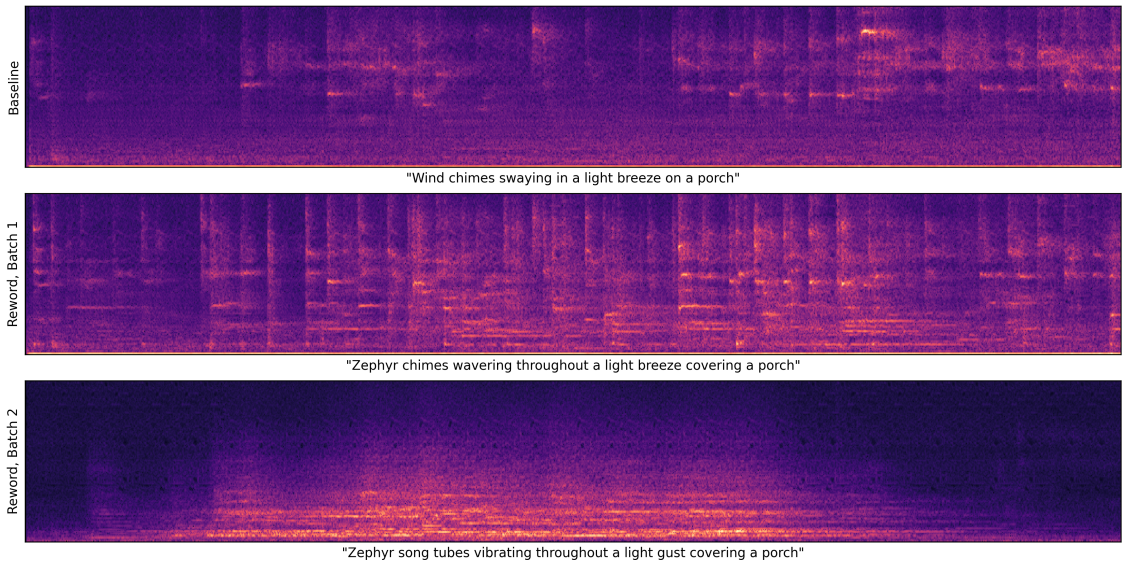

Figure 5.4: Textual Inversion recaptioning results, TI prompt 4-1.

CHAPTER 6: LIMITATIONS AND FUTURE WORK

The following chapter discusses limitations to our experimentation and potential avenues for future research with respect to text prompting on audio generative models.

6.1 LIMITATIONS

While our experiments have yielded promising results, there are a few limitations to consider for deeper evaluation.

6.1.1 Prompt Complexity

Given the scope of our baseline and recaptioned text prompts, we have effectively used Claude 3.7 Sonnet [17] to collect a large amount of text grounded in real-world data. However, the complexity and quantity of prompts is quite limited to this LLM's ability to generate text that is both coherent and relevant to the audio samples. This limitation is particularly evident in a small percentage of text prompts, where the recaptioned text is an exact match to the baseline text, (equivalent to a cosine similarity of 1.0 and L2 distance of 0.0). This hiccup in prompt generation suggests that, while current LLMs are at the point of providing a generalized variety of coherent text, they are still capable of failing on a larger number of tasks. Likewise, it is important to note that, following some number of prompts, the LLM returned to previous concepts with different wording, (just as we intended to recaption each concept!) The aforementioned hiccup, while not a detriment to our findings, does suggest that sample size should be considered when building generated datasets.

6.1.2 Sample Length

In a similar vein to the limitations of prompt complexity, we also note that generated sample length can play a more significant role in the effectiveness of our experiments. While we believe that 10-second samples proved valuable for our results, it is understood that longer and equally reliable samples can be achieved from the same model. In fact, an evaluation between 1-10 seconds and 10-20 seconds of an audio sample may provide further insight into the consistency of concepts during inference. Nevertheless, like with all generative models, there is a trade-off between sample length and quality, as longer samples run the risk of data hallucination. But our work has certainly left room for methods on the basis of sample length to be explored.

6.2 FUTURE WORK

Given the results and limitations of our experiments, we can additionally suggest several avenues for projects to build upon our work.

6.2.1 "Concept Building"

The results of our experiments suggest that generative audio models are capable of producing a wider variety of outputs than expected given specific prompting. This point is particularly evident as we observe concepts stay consistent despite being generated from prompts using different words. While generalization across prompts is expected, the results of our experiments raise a important question: what else can be done with these concepts? One potential avenue for model-generalized concepts is the idea of building a new concept from pieces that a model is already familiar with. For instance, if a generative audio model is trained on how two distinct objects sound, how effective would it be to generate the sound of both objects colliding? It is generally understood that generative models are capable of combining concepts in a way that combines their sounds, but whether or not this method is capable of essentially predicting interactions between objects in sound space is still an open question.

6.2.2 Comparisons Between Generative Audio Models

While our experiments have focused on the capabilities of a single generative audio model, it is important to note that there are several other models that are capable of generating audio samples. For instance, models such as Jukebox, MusicLM, and AudioLDM are capable of generating music samples from text prompts. While these models are not directly comparable to the model used in our experiments, it would be interesting to see how they perform on our recaptioning methods and whether or not they are capable of meeting or exceeding results.

CHAPTER 7: CONCLUSIONS

As a whole, our work demonstrates the extent to which text prompts can be recaptioned to make distinct changes, controlled and uncontrolled, to generated audio samples. Results following our research question and review of prompt-controlled audio generation verify that a level of variation is possible through basic recaptioning and conditioning on existing sample embeddings.

Audio Diffusion models have become relatively powerful and popular in recent years, especially given their ability to generate coherent samples in much less time than previous models. However, full control of sample generation is still a challenge as the embedding space obscures how concepts are represented. Recent work into generative audio personalization and the development of Textual Inversion have recently defied existing limitations, allowing for the generation of samples that are more closely aligned with the user's intent. On the other hand, no specific work has been done to explore how much variation is possible through prompt recaptioning. In turn, we demonstrate how pretrained generative models can have a level of flexibility given synonymous wording and embedding conditioning in a rigid context.

Resulting quantitative and qualitative analysis of our experiments verify the effectiveness of our recaptioning methods to the extent that a pretrained model is trained. However, with the use of a single pretrained generative model, a number of limitations are present and must be considered. The selected model, Stable Audio Open 1.0, is not generalizable to all types of sound and by no means perfectly represents learned concepts. Likewise, future experiments of a similar nature should consider different sample rates and sound lengths to determine a higher bound for prompt recaptioning. The use of a single model, while effective for sample-to-sample comparisons, does not provide a full picture of the capabilities of all generative audio models for our methods.

Nevertheless, our work shows promise for the future of generative audio models and how we choose to handle concept generalization. While training new models on larger datasets also assists in personalization, prompt recaptioning not only proves that other methods are possible, but also that they can do so without additional training. In turn, we hope to see future work pursue this idea as a means of fine-tuning generative models, as opposed to extensive training and retraining new models.

REFERENCES

[1] M. Plitsis, T. Kouzelis, G. Paraskevopoulos, V. Katsouros, and Y. Panagakis, "Investigating Personalization Methods in Text to Music Generation," Sep. 2023, arXiv:2309.11140 [cs]. [Online]. Available: http://arxiv.org/abs/2309.11140

[2] C. Thomé, J. Pertoft, and N. Jonason, "Applying textual inversion to control and personalize text-to-music models," *Proc. 15th Int. Workshop on Machine Learning and Music, 2024*, Nov. 2024. [Online]. Available: https://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A1912384&dswid=9661

[3] C. Thomé, "Applying textual inversion to control and personalize text-to-music models by audio reference," M.S. thesis, KTH Royal Institute of Technology, 2024.

[4] T. Mahmud and D. Marculescu, "OpenSep: Leveraging Large Language Models with Textual Inversion for Open World Audio Separation," Sep. 2024, arXiv:2409.19270 [cs]. [Online]. Available: http://arxiv.org/abs/2409.19270

[5] H. Manor and T. Michaeli, "Zero-shot unsupervised and text-based audio editing using DDPM inversion," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 2024. [Online]. Available: https://proceedings.mlr.press/v235/manor24a.html pp. 34 603–34 629.

[6] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022. [Online]. Available: https://arxiv.org/abs/2208.01618

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[9] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2301.12503

[10] S. Vatsal and H. Dubey, "A survey of prompt engineering methods in large language models for different nlp tasks," 2024. [Online]. Available: https://arxiv.org/abs/2407.12994

[11] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," 2024. [Online]. Available: https://arxiv.org/abs/2211.06687

[12] S. B. Azalea Gui, Hannes Gamper, "Adapting frechet audio distance for generative music evaluation," in *Proc. IEEE ICASSP 2024*, 2024. [Online]. Available: https://arxiv.org/abs/2311.01616

[13] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," 2024. [Online]. Available: https://arxiv.org/abs/2306.05284

[14] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," 2023. [Online]. Available: https://arxiv.org/abs/2208.12242

[15] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," 2023. [Online]. Available: https://arxiv.org/abs/2309.05767

[17] Anthropic, "Claude 3.7 sonnet, claude.ai," 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-7-sonnet

[18] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," 2024. [Online]. Available: https://arxiv.org/abs/2407.14358

APPENDIX A: LISTENING TO AUDIO SAMPLES

The medium of audio is understood to be difficult to convey through academic writing alone. As such, we provide a weblink to listening demos of the generated audio samples from our experiments. Inclusion of these demos are intended to provide readers with their own subjective understanding of results, and as such, are optional to our research conclusion. Further details on the audio samples and project development can be found on `https://ematth.dev`.