Audio-Driven Image Generation and Editing with Pretrained Diffusion Models

by

Burak Can Biner

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Science and Engineering

KOÇ ÜNİVERSİTESİ

January 02, 2024

Audio-Driven Image Generation and Editing with Pretrained Diffusion Models

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Burak Can Biner

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Aykut Erdem (Advisor)

_____

Prof. Erkut Erdem

_____

Prof. Engin Erzin

Date: _____

To all the failed attempts along the way

ABSTRACT

Audio-Driven Image Generation and Editing with Pretrained Diffusion Models

Burak Can Biner

Master of Science in Computer Science and Engineering

January 02, 2024

We are witnessing a revolution in conditional image synthesis with the recent success of large scale text-to-image generation methods. This success also opens up new opportunities in controlling the generation and editing process using multi-modal input. While spatial control using cues such as depth, sketch, and other images has attracted a lot of research, we argue that another equally effective modality is audio since sound and sight are two main components of human perception. Hence, in this thesis we propose SonicDiffusion to enable audio-conditioning in large scale image diffusion models. Our method first maps features obtained from audio clips to tokens that can be injected into the diffusion model in a fashion similar to text tokens. We introduce additional audio-image cross attention layers which we finetune while freezing the weights of the original layers of the diffusion model. In addition to audio conditioned image generation, our method can also be utilized in conjuction with diffusion based editing methods to enable audio conditioned image editing. We demonstrate our method on a wide range of audio and image datasets. We perform extensive comparisons with recent methods and show favorable performance.

ÖZETÇE

Önceden Eğitilmiş Yayınım Modelleri ile Ses Tabanlı Görüntü Oluşturma ve Düzenleme
Burak Can Biner
Bilgisayar Bilimleri ve Mühendisliği, Yüksek Lisans
January 02, 2024

Son zamanlarda büyük ölçekli metinden görüntü üretim yöntemlerinin ses getiren başarısıyla koşullu görüntü üretimi alanında bir devrime şahitlik etmekteyiz. Bu başarılar aynı zamanda farklı çok kipli koşullanmış görüntü üretimleri için de yeni fırsatlar sunmakta. Uzamsal kontrolü belirleyen görüntü, derinlik, eskiz gibi kipin araştırma konusu olarak ilgi çekerken ses ve görüntünün insan algısının iki ana bileşeni olmasından ötürü eşit miktarda etkili olabilecek bir diğer kipin ses olduğunu savunmaktayız. Bu sebeple, bu tezde büyük ölçekli yayınım modellerinde sese koşullanmış görüntü üretimini mümkün kılan SonicDiffusion adını verdiğimiz bir yöntem önermekteyiz. Yöntemimiz öncelikle ses kliplerinden elde edilen öznitelikleri metin belirteçlerine benzer şekilde yayınım modeline enjekte edilebilen belirteçlere dönüştürmekte. Ardından yayınım modellerinin orijinal değişkenlerini dondurarak, ekstra ses-görüntü çapraz dikkat katmanları sunmakta ve bu katmanları ince ayarlamakta. Yöntemimiz ses koşullandırılmış görüntü üretiminin yanı sıra, görüntü düzenlemesi için yayınım modelleri tabanlı görüntü düzenleme yöntemleriyle birlikte kullanılabilmekte. Yöntemimizin performansını çeşitli ses ve resim veri kümeleri üzerinde göstermekteyiz. Yakın dönemde öne çıkan diğer yöntemlerle de kapsamlı bir kıyaslama yapmakta ve lehimize sonuçlar göstermekteyiz.

ACKNOWLEDGMENTS

First and foremost, I express profound gratitude to my advisor, Assoc. Prof. Aykut Erdem, for their unwavering support and priceless guidance. I am also immensely thankful to Prof. Erkut Erdem and Duygu Ceylan for their enlightening insights and invaluable feedback.

A special acknowledgment goes to Farrin Marouf Sofian and Umur Berkay Karakaş, fellow members of my research group, for their exceptional contributions to our research studies.

My sincere thanks are extended to the members of the KUIS AI Lab. I appreciate Canberk's consistent presence, Ece's wonderful friendship and support, Mert's fruitful research discussions, Adil Kaan's valuable research advice, and the inventive paper ideas that may never see publication. I acknowledge Batur for contributing to discussions regardless of his focus on NLP, Sadra for the pleasant backgammon breaks, and countless others who have turned the lab into a vibrant environment beyond a mere study space. Lastly, I express gratitude to Selçuk, Alihan, and Burak for making Istanbul feel like Ankara.

In closing, I extend a special thanks to my parents and sister for their continuous support throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

SD	Stable Diffusion
LDM	Latent Diffusion Models
MSE	Mean Squared Error
GAN	Generative Adversarial Networks
PnP	Plug and Play
GPT	Generative Pre-training Transformer
NLP	Natural Language Processing
CLIP	Contrastive Language-Image Pre-Training
CLAP	Contrastive Language-Audio Pre-Training
BLIP	Bootsraping Language-Image Pre-Training
DDPM	Denoising Diffusion Probabilistic Models
AIS	Audio-Image Similarity
AIIS	Image-Image Similarity
AIC	Audio-Image Content
FID	Fréchet Inception Distance
ItW	Into the Wild

Chapter 1

INTRODUCTION

1.1 Problem Definition and Motivation

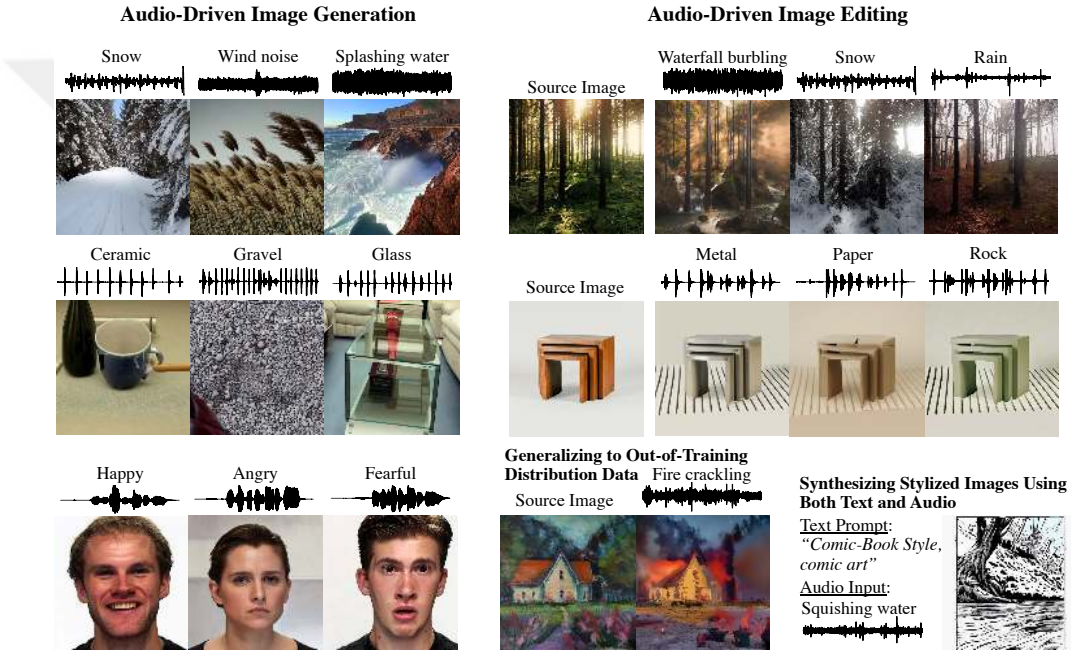


Figure 1.1: **Audio-Driven Image Generation and Editing**. In this thesis, we introduce a novel approach where audio inputs guide image generation and editing. We are able synthesis images from landscape audios, sound emanating from several objects, and emotional speeches. Beyond this, our method can combine audio with textual information for richer image sampling.

In the era of artificial intelligence, generative models have emerged as a new technique that have quickly captured attention of the entire world. They have transformed tasks previously seemed impossible into everyday routines in just a few years, becoming irreplaceable tools for engineers, creators, scientists, and more. Par-

ticularly the visual content generation models, have demonstrated an extraordinary growth. These models have the capability of creating images, videos, 3D models, and avatars in diverse styles with mesmerizing perceptual quality. Their application span from various industrial fields to research based efforts. Entertainment content, advertisement ideas, creative artistic works, generation of synthetic data, enhancement of system efficiencies are only a small subset of tasks that these models are useful. Introduction of [Goodfellow et al., 2014] was one of the first footsteps of this upcoming trend. There were plenty of GAN-based approaches [Karras et al., 2019] followed this trend and demonstrated promising visual generations. However, a distinct technique, apart from GANs, became the catalyst for this trend to spread globally.

Diffusion models have revolutionized visual content synthesis and manipulation [Ramesh et al., 2022, Rombach et al., 2022, Saharia et al., 2022]. These models have surpassed previous deep generative approaches like GANS, Vaes in photorealism, sparking a wave of creative approaches to text-based image generation and semantic image editing. From generating landscapes that stretch the bounds of imagination to crafting portraits with striking realism, these models have demonstrated a remarkable ability to capture and recreate the complexity of visual data. Several works [Podell et al., 2023, Zhao et al., 2023, Ye et al., 2023, Zhang et al., 2023a] demonstrated generation of this complex visual data can be steered by multiple modalities such as text and image. Despite these advancements in other modalities, the integration of audio cues into the image generation process remains a relatively uncharted domain.

Our approach, aims to explore this undiscovered domain by proposing a novel method to enhance image generation models with the ability to process audio cues. Our study is inspired by the profound connection between sound and sight that characterizes human perception. The rustling of leaves, the bustling of a city, or the tranquility of a flowing river, each sound paints a vivid picture in the mind's eye. This synergy is not merely a poetic intersection but a tangible aspect of cognition, capable of creating more engaging and contextually rich visual content. By exploring this interplay, our work aims to push the boundaries of generative models to a new

level.

Exploring audio-to-image generation is an emerging line of research. Prior GAN-based works [Lee et al., 2022b, Lee et al., 2023a, Li et al., 2022] have shown how audio can guide image generation, yet they often fall short in sample quality and in capturing the audio's semantic nuances. More recent methods [Qin et al., 2023, Yariv et al., 2023b] have used the pre-trained text-to-image Stable Diffusion model [Rombach et al., 2022] to integrate audio information into image generation. These approaches align cues with the text conditioning space of Stable Diffusion but do not modify the denoising network, limiting the model's adaptability to new modalities. Multi-modal generative networks [Tang et al., 2023, Girdhar et al., 2023] have also investigated image generation from audio information. However, they train large scale and expensive models with huge datasets and do not specifically focus on audio-driven generations. To address these challenges, we propose SonicDiffusion, a new diffusion model that can generate images semantically aligned with accompanying audio.

1.2 Proposed Method and Contributions

SonicDiffusion, steers the process of image generation and editing using auditory inputs. Central to our approach is the adaptation of the Stable Diffusion framework, now repurposed to interpret and interact with audio signals. We achieve this by integrating audio-image cross-attention layers, which act like a bridge to translate auditory information into visual features. This integration preserves the model's inherent capacity for high-quality image generation while introducing a new modality with minimal additional training required. Furthermore, we extend our method to edit real images, performing modifications that reflect the characteristics of the audio input through feature injection.

We validate the efficacy of our approach through rigorous testing on three diverse datasets. Our results show that our model surpasses current existing methods for audio-to-image generation and audio-driven image editing, both in qualitative and quantitative terms. Notably, our model exhibits exceptional ability in capturing in-

tricate high-frequency details in landscape generation, accurately rendering human facial features influenced by sound, and precisely mapping specific sounds to corresponding visual elements that reflect the material properties of depicted objects.

To sum up, our contributions are threefold: (i) We introduce a new diffusion model that extends the capabilities of the pre-trained Stable Diffusion model, enabling sound-guided image generation. (ii) Our approach presents audio-image cross-attention layers, strategically designed to require a minimal set of trainable parameters. (iii) We demonstrate the versatility of our model in not just generating, but also editing images to match auditory inputs.

1.3 Organization of the Thesis

The structure of the thesis is organized as follows:

Chapter 2 discusses the advances in diffusion-based image generation, involvement of adapters in generative networks and audio-guided image and video generation.

Chapter 3 elaborates the methodology that we followed in this thesis for both audio-driven image generation and editing.

Chapter 4 provides information on datasets, evaluation metrics and competing approaches.

Chapter 5 demonstrates capabilities of our approach, and discusses importance of each component on our results.

Chapter 6 concludes the thesis and mention further research directions.

Chapter 2

RELATED WORK

2.1 Diffusion-based Image Generation and Editing

Image generation and editing have evolved significantly with diffusion models like Imagen [Saharia et al., 2022], DALL·E 2 [Ramesh et al., 2022], and Latent Diffusion [Rombach et al., 2022], which generate realistic images from text prompts. These models are also central in text-based image editing [Meng et al., 2021, Nichol et al., 2022]. Various approaches have been developed since then. PnP [Tumanyan et al., 2023] injects image features and manipulates self-attention weights during text guided editing. Imagic [Kawar et al., 2023] optimizes text embeddings and fine-tunes the model for targeted edits. Prompt-to-Prompt [Hertz et al., 2022] adjusts cross-attention maps to ensure spatial consistency during editing. InstructPix2Pix [Brooks et al., 2023] employs GPT-3 [Brown et al., 2020] to generate editing instructions which are then used to train a diffusion model to enable instruction based editing. Pix2pix-zero [Parmar et al., 2023] calculates edit directions from the text embeddings of original and edited image pairs. Lastly, [Mokady et al., 2023] focus on optimizing null-text embeddings to facilitate better inversion and editing performance. All these methods, however, focus on text guided generation and editing while our goal is to unlock audio guidance.

2.2 Adapters

Adapters have proven to be highly effective for transfer learning within large pre-trained NLP models, achieving results that closely approach the state-of-the-art [Houlsby et al., 2019]. For instance, adapters have recently been utilized in [Zhang et al., 2023b] to tune LLaMa [Zhang et al., 2023b] into an instruction model. In the text-to-image diffusion models, adapters have also been instrumental. T2I-Adapter [Mou

et al., 2023] illustrates that adapters can offer a straightforward, cost-effective, and flexible means of guiding pre-trained Stable Diffusion models while preserving their core structure. ControlNet [Zhang et al., 2023a] and its' successors like [Zhao et al., 2023] utilizes a larger encoder to preserve spatial structures with great accuracy. Gligen [Li et al., 2023b] utilizes trainable self-attention layers that can inject grounding information in to text-to-image generation. IP-Adapter [Ye et al., 2023] introduces an adapter architecture with a decoupled cross-attention mechanism, facilitating multimodal image generation.

2.3 Audio-Driven Image Generation and Editing

Incorporating audio into image generation and editing has recently gained traction. Early studies predominantly utilized GANs [Goodfellow et al., 2014]. SGSIM [Lee et al., 2022b] uses a two-step approach that augments the CLIP [Radford et al., 2021] embedding space with audio embeddings via InfoNCE loss [Alayrac et al., 2020, van den Oord et al., 2019], combining audio, image, and text. It then manipulates StyleGAN's [Karras et al., 2019] latent codes to produce images compatible with the audio inputs. Figure 2.1 summarizes this two-step approach. A follow-up work named Robust-SGSIM enhances this method by adding a KL divergence term to its loss function to better preserve image structure and reduce bias. Following a similar idea, Sound2Scene [Sung-Bin et al., 2023] utilizes InfoNCE loss to align image and audio encoders and combines it with BigGAN [Brock et al., 2018] for synthesis.

AVStyle [Li et al., 2022] offers a different viewpoint by introducing audio-visual adversarial losses. In this approach, one discriminator is dedicated to enhancing the texture stylization from the given audio signal and the other one is dedicated to preserving the structure via patch-wise contrastive learning. They make editing the image via audio information possible with this methodology. Figure 2.2 shows their approach.

Diffusion-based methods have gained popularity recently. GlueGen [Qin et al., 2023] is a flexible model handling multimodal inputs, merging linguistic contexts via XLM-R [Conneau et al., 2020] and auditory elements through AudioCLIP [Guzhov

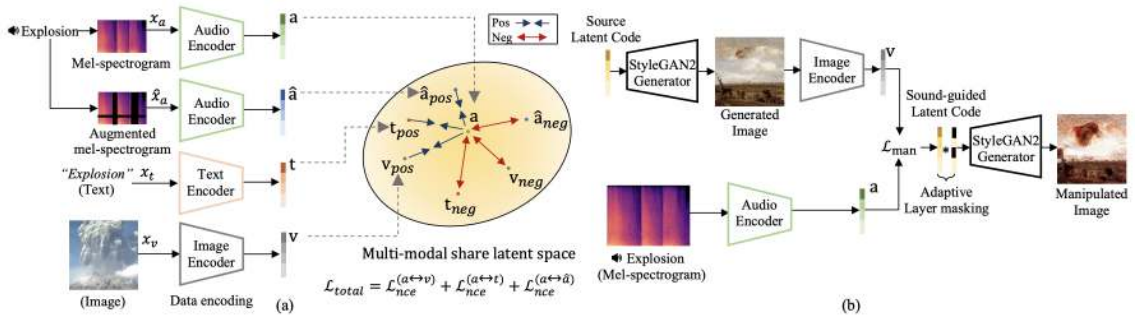


Figure 2.1: **Training pipeline of SGSIM**. SGSIM follows a two step training approach. Initially a joint encoder utilizes a contrastive loss to bring positive text, image and audio embeddings closer and pushes negative embeddings away from. In the second stage audio embeddings are injected into the StyleGAN2 generation process. This figure is taken from [Lee et al., 2022b].

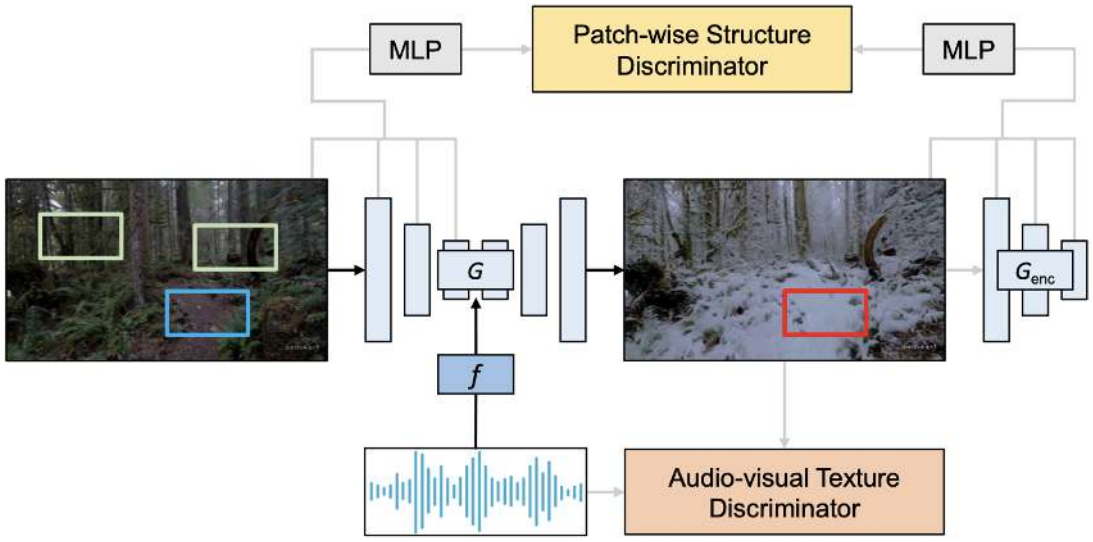


Figure 2.2: **Training pipeline of AVStyle**. AVStyle leverages one discriminator to preserve initial source image structure and another discriminator manipulate images with the corresponding audio information. This figure is taken from [Li et al., 2022].

et al., 2022], aligning representations within the CLIP framework to enable conditional image generation via [Rombach et al., 2022]. This is one of the first works that incorporates audio into Stable-Diffusion architecture. The overview of their approach is demonstrated in Figure 2.3

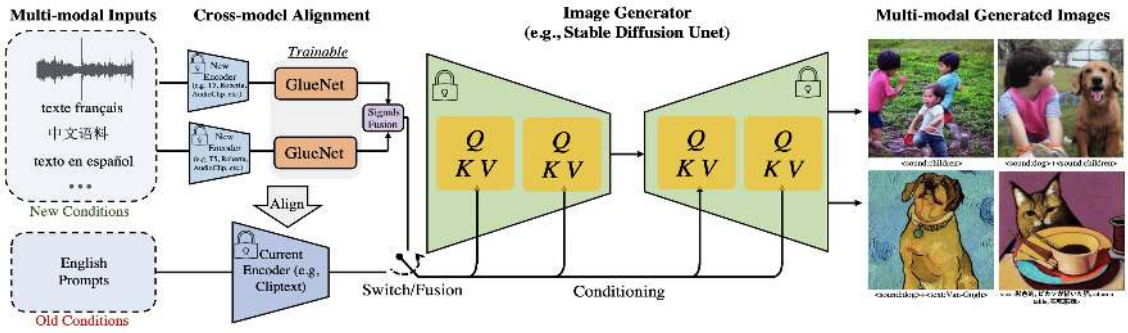


Figure 2.3: **Overview of GlueGen**. GluGen leverages the strong pre-trained Stable Diffusion model. They introduce a GlueNet model which aligns audio information that is extracted from AudioCLIP[Guzhov et al., 2022] with original text conditioining space of the Stable Diffusion model. This figure is taken from [Qin et al., 2023].

AudioToken [Yariv et al., 2023b] presents an audio embedder that replaces a specific token's CLIP embedding, enabling audio-guided image generation via pre-trained Stable-Diffusion model. Diffferent from Gluegen's conditioining space alignment strategy, AudioToken embeds the audio signal information on the token level. Which once again does not change generative capability of the pre-trained text-to-image model. Their approach resembles the personalizaiton methods like Textual-Inversion [Gal et al., 2022] but they do not learn a mapping for a small set of audios. They propose learning a general purpose framework to embed audio information into a token. Figure 2.4 demonstrates the overview of this learned embedder network.

There are some other recent works follows an optimization approach. [Lee et al., 2023b] introduce a two-stage method, initially using an audio captioning transformer for generating attention map, followed by direct sound optimization to generate new images from the initial latent embeddings. Figure 2.5 summarizes their approach. Another optimization based method is 'Align, Adapt, and Inject'(AAI) [Yang et al., 2023]. AAAI leverages a three-phased strategy to perform generation, stylization and image editing via audio conditioning. First they train a multi-modal encoder to map audio to a conditioning space. Secondly they optimize an audio representation from a small set of images of a specific audio class. This approach is inspired from the personalization methods like Textual-Inversion [Gal et al., 2022]. Finally they inject

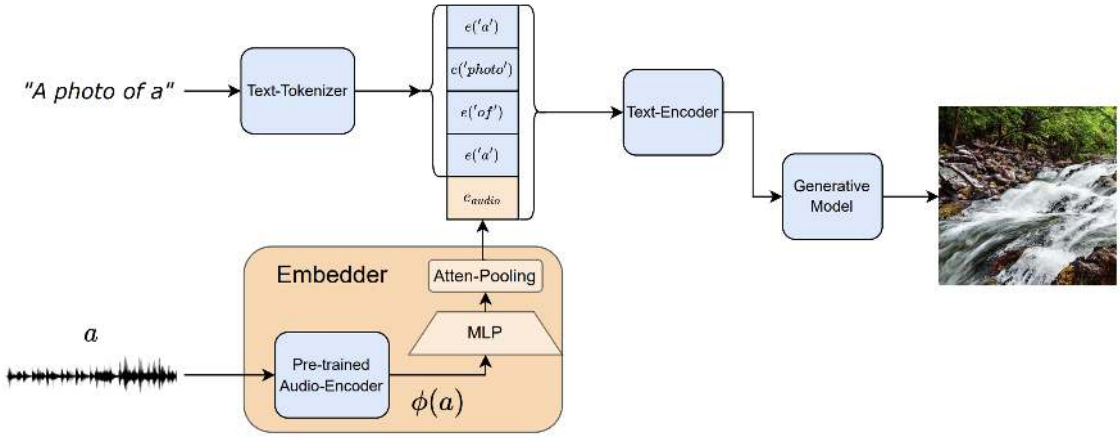


Figure 2.4: **Overview of AudioToken**. Audiotoken learns an embedder network that takes audio information and yields a token which necessary audio information. This figure is taken from [Yariv et al., 2023b].

features into the Diffusion model similar to PNP [Tumanyan et al., 2023] to edit appearances or styles of the images. Figure 2.6 summarizes their overal framework.

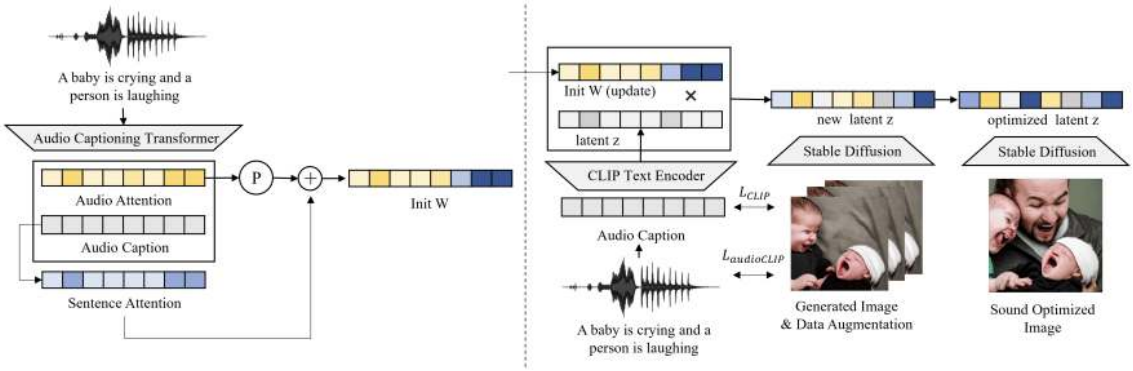


Figure 2.5: **Overview of Generating Realistic Images from In-the-wild Sounds**. Two staged approach of [Lee et al., 2023b]. First vector W is initialized. Then latent is optimized during the second stage. This figure is taken from [Lee et al., 2023b].

ImageBind [Girdhar et al., 2023] combines audio, image, and text modalities in one embedding space using InfoNCE, also enabling audio-driven image synthesis. Overview of this unified embedding space approach is shown in Figure 2.7. Another multi-modal work named Composable Diffusion (CoDi) [Tang et al., 2023], enables generation between multiple modalities such as audio, text, image and video. Fig-

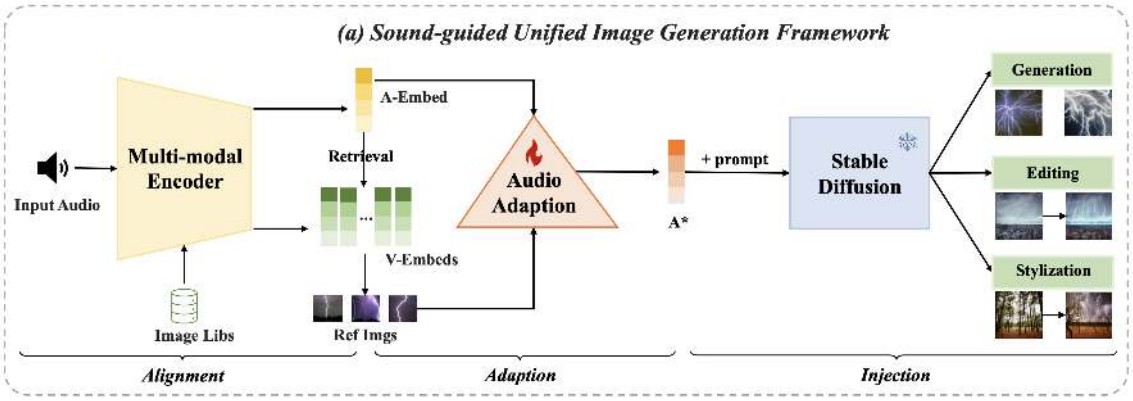


Figure 2.6: **Overview of Adapt, Align and Inject**. A learned multi-modal encoder aligns audio modality with others. Aligned embedding is optimized for a small set of reference images. Finally, adapted new prompt representation is fed into the SD. This figure taken from [Yang et al., 2023].

ure 2.8 summarizes overall training and inference schemes of this framework. These multi-modal generative models follow diffusion based approaches for image generation, however, they perform large scale trainings instead of building lightweight modules on top of strong pre-trained models.

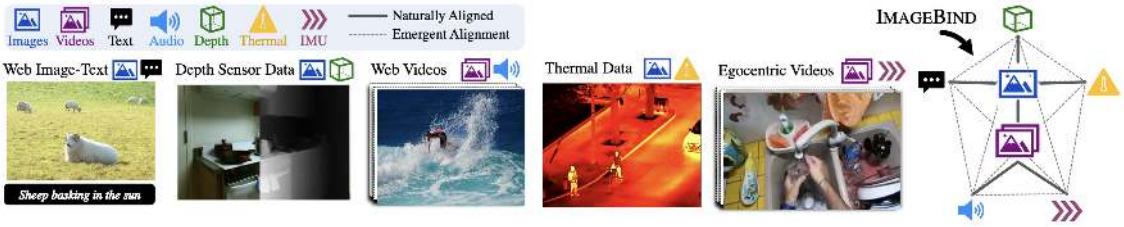


Figure 2.7: **Overview of Imagebind**. Taking image as the reference modality, Imagebind creates a unified space for several modalities. Here audio modality binded into this space via video datasets. This figure is taken from [Girdhar et al., 2023].

Our model distinguishes itself from these methods by adaptively tuning the denoising process of the Stable Diffusion. In particular, we incorporate audio-image cross-attention layers to enhance the semantic and visual alignment between the synthesized images and the input audio. We enjoy the benefits that diffusion models have over GAN models like high perceptual quality and lower chance of mode

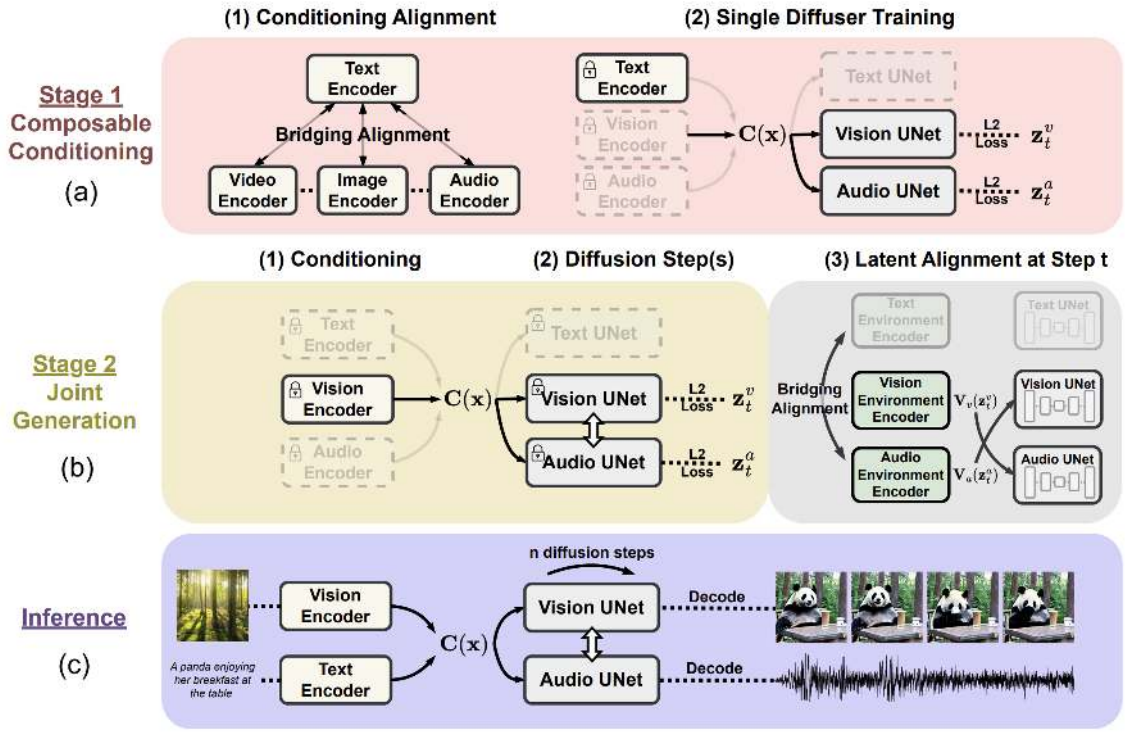


Figure 2.8: **Overview of Composable Diffusion**. Separately trained modalities aligned with the text encoder which serves as a bridging element. Then a second stage jointly trains separate generative networks where models attend to each other. Here an image can synthesize from a given audio. This figure is taken from [Tang et al., 2023].

collapse issue. We built our work on top of strong pre-trained text models. Thus, we do not need to fully train a large model on a very large scale.

2.4 Audio-Driven Video Generation

Recent years have seen growing interest in generating videos from audio. Building on their earlier work [Lee et al., 2022b], [Lee et al., 2022a] developed a method combining a sound inversion encoder with a StyleGAN generator, using recurrent blocks for fine texture variations in video frames. TPoS [Jeong et al., 2023] similarly employs an audio encoder with recurrent blocks, adding a temporal attention module for better audio temporal understanding. This model uses the multimodal CLIP space from [Lee et al., 2022b] for image synthesis. Another recent approach is TempoTokens [Yariv et al., 2023a], which proposes to use specific tokens generated

from audio using an audio mapper. The model leverages an audio-conditioned temporal attention mechanism and facilitates the generation of videos by integrating these tokens into a pre-trained text-to-video diffusion model.

Chapter 3

METHOD

In this chapter, we introduce SonicDiffusion, an approach that steers the process of image generation and editing using auditory inputs. As depicted in Fig. 3.3, our proposed approach has two principal components. The first module, termed the Audio Projector, is designed to transform features extracted from an audio clip into a series of inner space tokens. These tokens are subsequently integrated into the image generation model through newly incorporated audio-image cross-attention layers. Crucially, we maintain the original configuration of the image generation model by freezing its existing layer weights. This positions the added cross-attention layers as adapters, serving as a parameter-efficient way to fuse the audio and visual modalities. The overall training pipeline of SonicDiffusion is illustrated in Figure 3.3. Our model employs a two-stage training process, each stage serving a specific purpose in achieving audio-driven image synthesis. The initial stage is dedicated to training the audio projector module. The second stage of training focuses on integrating gated cross-attention layers within the Stable Diffusion (SD) model framework. Once the training is done, these cross-attention layers guide the image generation with audio information. We further employ a feature injection strategy to be able to edit images with audio. Figure 3.4 presents overview of the inference pipeline to generate images or editing an image that already exists.

Below, we first give background knowledge on diffusion models (Sec.3.1), and present data pre-processing steps (Sec. 3.2). We then detail the specifics of the audio projector (Sec. 3.3) and discuss the methodology employed for the integration of audio tokens into the SD framework (Sec. 3.4). Furthermore, we describe how we can extend our approach to sound-guided image editing through feature injection (Sec. 3.5).

3.1 Diffusion Preliminaries

Diffusion models are generative models that can perform unconditional or conditional image generation. Denoising Diffusion Probabilistic Models (DDPM) introduces diffusion probabilistic models as parameterized Markov chains. The defined Markov chain slowly injects noise into the original image and then learns to denoise it back. The noise injection part is named the forward diffusion process. Forward diffusion process, randomly injects noise in each of the timesteps according to a variance schedule. In the reverse diffusion process, a network learns how to obtain the denoised image from the noised one to construct images from the noise. Figure 3.1 demonstrates noise injection and denoising schemes via Markov chains.

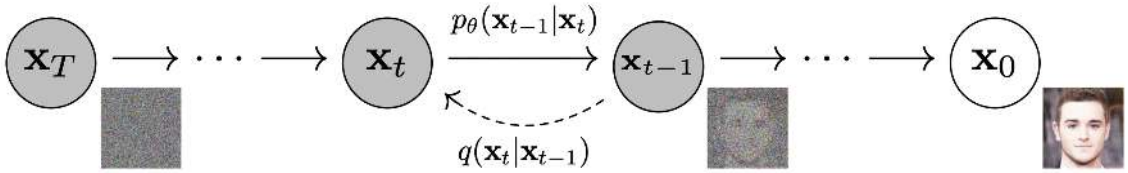


Figure 3.1: **Denoising Diffusion Probabilistic Models Markov Chains.** Original image is perturbed by noise injection in each timestep. Then reverse diffusion process learns to denoise this added noise.

Initial diffusion models manipulate the pixel-space by adding and removing noise directly on the image. Stable Diffusion (SD) emerged as the predominant approach, to prove that such operations can be done in the latent space. This shift introduced by SD, from pixel space to latent space significantly reduces computational complexity by avoiding redundancy inherent in pixel space. SD is a latent text-to-image diffusion model (LDM) [Rombach et al., 2022], integrating a variational autoencoder to project an input image $\mathbf{x}$ into a reduced latent space $\mathbf{z} = E(\mathbf{x})$ via its encoder E. The corresponding decoder D reconstructs an image $\hat{\mathbf{x}} = D(E(\mathbf{x}))$ from any given latent $\mathbf{z}$. The denoising network that operates on the latent level is an UNet [Ronneberger et al., 2015] architecture. UNet backbone operates in this latent space to progressively denoise input latents at each diffusion step. Figure 3.2 exhibits the forward diffusion process and conditional denoising of the latent via UNet backbone.

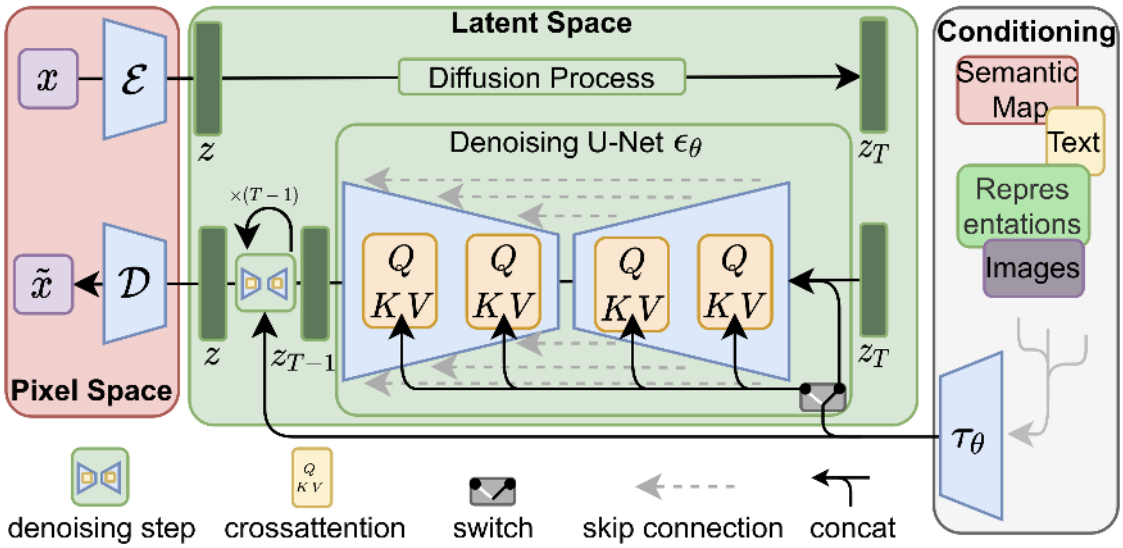


Figure 3.2: **Overview of the Stable Diffusion architecutre.** An encoder E maps images from pixel space to latent space z. In this latent space forward diffusion process adds noise to latent and UNet architecture denoises it back. τ_θ encodes conditoining information into tokens that can be injected into UNet via cross-attention layers. A decoder D converts latents to pixel space.

An unconditional UNet ϵ_θ is trained with the established denoising loss introduced by Denoising Diffusion Probabilistic Models (DDPM) [Ho et al., 2020]:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z}\sim E(\mathbf{x}),\epsilon\sim N(0,1),t}\left[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t)||_2^2\right] \tag{3.1}$$

where $\mathbf{z}_t$ represents the noise corrupted latent at time step t, and $\mathbf{z}_0 = E(\mathbf{x})$.

The conditional variant of LDM incorporates a cross-attention mechanism, enabling it to be conditioned with different modalities. In this cross-attention mechanism query projections come from the latent that is denoised and key and value projections come from the conditioning modality such as images, semantic maps, texts. Attention of conditioning space to latent features successfully guides the image generation towards target conditioning. Specifically, SD leverages text conditioning through embeddings generated by the CLIP [Radford et al., 2021] text encoder. Consequently, the denoising loss is adapted to accommodate this conditioning, as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z}\sim E(\mathbf{x}),\epsilon\sim N(0,1),t,\mathbf{y}}\left[||\epsilon - \epsilon_\theta(\mathbf{x}, t, c_\phi((\mathbf{y}))||_2^2\right] \tag{3.2}$$

where $\mathbf{y}$ is the conditioning text prompt, and c_ϕ is the text encoder transforming $\mathbf{y}$ into an intermediate representation.

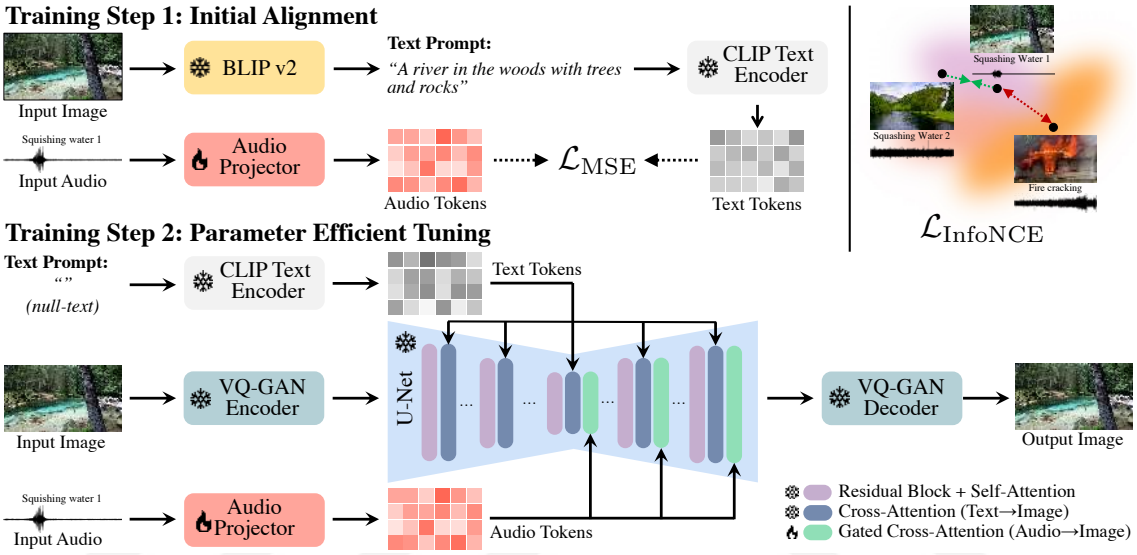


Figure 3.3: **Training pipeline of SonicDiffusion** comprises two distinct stages: (1) aligning audio features with CLIP's semantic space, and (2) sound-driven parameter-efficient tuning of the Stable Diffusion (SD) model. In stage (1), the audio projector module is trained to transform audio clips into semantically rich tokens, employing MSE and contrastive loss functions. Stage (2) involves integrating gated cross-attention layers, which facilitate interaction between image and audio modalities, into the existing SD framework. Only these newly added layers, alongside the audio projector module, are adjusted to enable audio-conditioned image synthesis.

3.2 Preprocessing

In SonicDiffusion, we employ the CLAP model [Elizalde et al., 2023] as audio feature extractor. CLAP effectively merges audio and language into a single shared embedding space, showing strong performance with robust generalization capabilities across various downstream tasks, including sound event detection, acoustic scene

classification and speech-based emotion recognition. We preprocess the audio input in accordance with CLAP. Specifically, we generate log-mel spectrograms with a hop size of 320 seconds, a window size of 1024 seconds, and 64 mel bins. Consistent with the training of Stable Diffusion v1.4, our model processes images of size 512×512 pixels. We apply random horizontal flipping as our data augmentation technique. We conducted additional experiments with color space and geometric augmentations, such as color jitter, grayscale, random perspective, and random rotation, but observed that these led to degraded performance. Apart from horizontal flipping, the only data augmentation found to be beneficial is randomly cropping images instead of applying center cropping. To facilitate text-audio mapping as part of our initial training loss, we generated the caption of each image in training set using BLIP v2 [Li et al., 2023a]. We also tested audio captioning models in [Xu et al., 2022, Mei et al., 2021] but found that BLIP v2 provided captions which are much better aligned with the audio and image pairs.

3.3 Audio Projector

In training the Stable Diffusion model, text-encoding space of the CLIP[Radford et al., 2021] model is utilized to guide image generation with text-descriptions. CLIP is a widely recognized text-encoder that uses a large text-image pair dataset to understand image and text in a unified way. Likewise, Stable Diffusion employs a sizable dataset containing image and text pairs to learn an alignment between CLIP's text-encoding space and images. To achieve a similar semantic alignment between the audio and image domains, we introduce the concept of an *Audio Projector* module. This module is designed to transform audio clips into semantically rich tokens. Just like Stable Diffusion utilizes CLIP, we utilize the semantically rich tokens of the Audio Projector as a conditioning-space from audio to image. The architecture of the Audio Projector is depicted in Fig. 3.5, converts audio clips into tokens for additional conditioning input. It utilizes the CLAP model [Elizalde et al., 2023], a state-of-the-art audio encoder, to generate one-dimensional embeddings from audio inputs. These embeddings are then transformed by an initial mapper with 1D con-

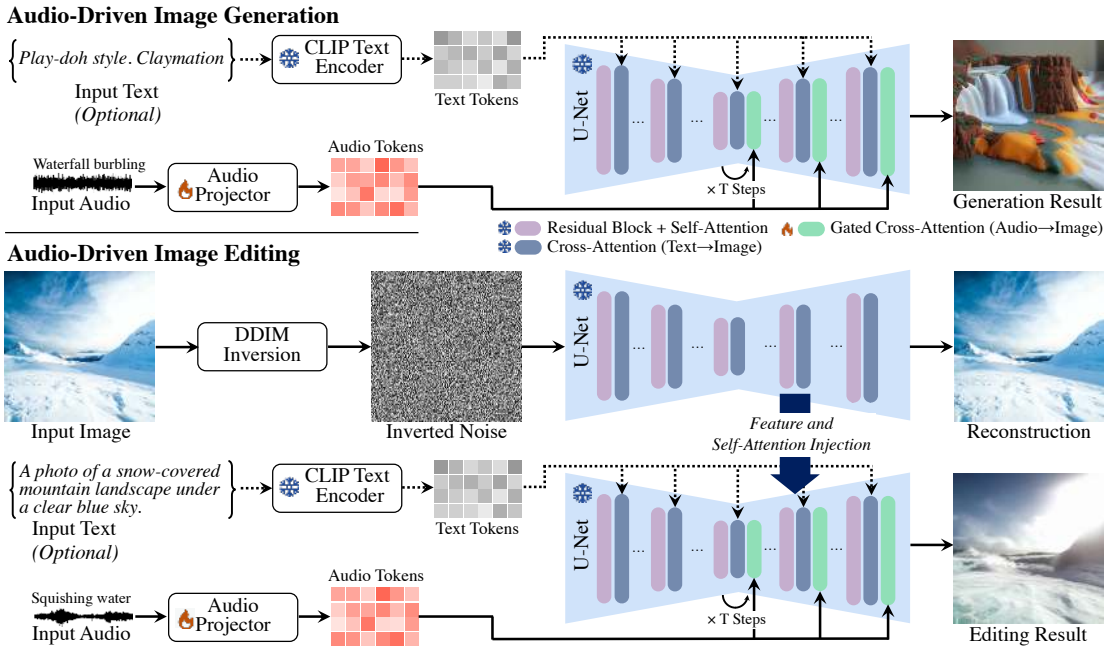


Figure 3.4: **Inference overview of our proposed SonicDiffusion model**. Our framework allows for two core functionalities: (1) audio-driven image generation, and (2) audio-guided image editing. In (1), both sound inputs and optional text prompts are tokenized, guiding the denoising process via text-to-image and audio-to-image cross-attention layers. For (2), the process begins with the inversion of the input image using DDIM. Subsequently, extracted spatial features and self-attention maps are integrated into the generation process, complemented by audio-conditioned cross-attention maps to obtain the desired changes.

volutional and deconvolutional layers into K tokens of C channels. These tokens are further refined by four self-attention blocks to produce final representations, which are compatible with the cross-attention layers of the SD model. Since we choose $K = 77$ and $C = 768$, our audio projection space dimensions matches with the text-conditioned embedding space of Stable Diffusion.

The training of the Audio Projector includes a joint strategy involving both Mean Squared Error (MSE) and contrastive loss functions. Specifically, for the contrastive loss, we employ the InfoNCE [Alayrac et al., 2020] loss formula, treating audio clips from the same class as positive pairs, and those from different classes as negative pairs. Within each training batch, two audio clips, $\mathbf{a}_0$ and $\mathbf{a}_1$, are randomly selected

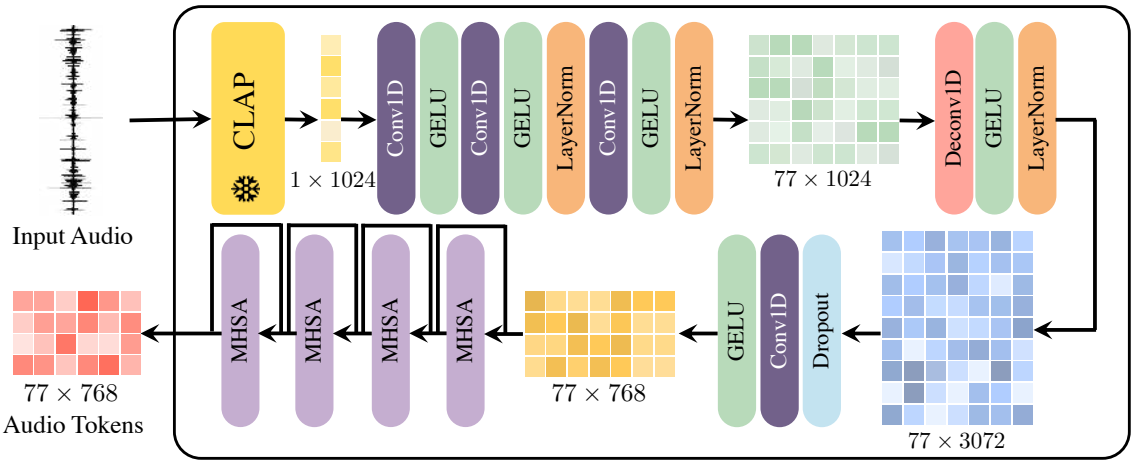


Figure 3.5: **Audio Projector** utilizes CLAP audio encoder [Elizalde et al., 2023] for initial feature extraction, followed by a mapper with 1D convolutions and deconvolutions. Four self-attention layers are then integrated to enhance learning effectiveness of the projector.

from the same class, alongside N audio clips, denoted as $\{\mathbf{a}_i\}_{i=2}^{N+1}$, from other classes. The contrastive loss objective is thus formulated as:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\langle \mathbf{a}_0, \mathbf{a}_1 \rangle)}{\sum_{i=1}^{N+1} \exp(\langle \mathbf{a}_0, \mathbf{a}_i \rangle)} \tag{3.3}$$

where N is set to 128 in our experiments.

This contrastive loss pushes different classes away from each other, resulting in a well-defined semantic space. To enrich this space, we utilize audio-image pairs from training dataset. For each audio clip, captions are extracted from corresponding images using the BLIP v2 [Li et al., 2023a] and their text embeddings are computed using the text encoder of SD. An MSE loss is defined as the squared norm difference between the audio tokens $\mathbf{c}_{\text{audio}}$ and the text tokens $\mathbf{c}_{\text{text}}$:

$$\mathcal{L}_{\text{MSE}} = \|\mathbf{c}_{\text{audio}} - \mathbf{c}_{\text{text}}\|_2^2, \tag{3.4}$$

As discussed earlier, since text and audio projection spaces exactly match in dimensions, we are able to define this straight-forward MSE loss function.

The training of the audio projector combines these two losses with weights $\alpha_1 = 1.0$ and $\alpha_2 = 0.25$, respectively:

$$\mathcal{L}_{\text{AudioProjector}} = \alpha_1 * \mathcal{L}_{\text{InfoNCE}} + \alpha_2 * \mathcal{L}_{\text{MSE}} \tag{3.5}$$

We calculate both the MSE and the contrastive losses for each of the K audio tokens. Drawing inspiration from GlueGen [Qin et al., 2023], which showed that information in CLIP text tokens is not uniformly distributed, instead of averaging the loss across all the tokens, we apply a weighted summation. This is formulated using reverse sigmoid weighting, as given below:

$$\mathcal{L} = \sum_{i=1}^{N} w_i \mathcal{L}_{\text{i}} \ , \quad w_i = \frac{t}{t + \exp(i/t)} \ , \tag{3.6}$$

where $\mathcal{L}_{\text{i}}$ is the total loss calculated for token i, and t is temperature coefficient, which is set to 5 in our experiments.

It is also worth noting is that keeping C at 768 enables us to initialize the weights of our audio-image cross-attention layers from the pre-trained text-image cross-attention layers of Stable Diffusion. This initialization strategy is helpful for faster convergence in the upcoming second-stage of the training.

3.4 Gated Cross-Attention for Audio Conditioning

Various approaches have been explored in the literature for aligning different domains or modalities with the conditioning space of Stable Diffusion (SD). Uni-ControlNet [Zhao et al., 2023] and GlueGen [Qin et al., 2023] append new modality information to text tokens, while AudioToken [Yariv et al., 2023b] learns a specific token representing the unique information provided in the input. These methods use the existing text-image cross attention layers of SD, originally trained for text conditioning, to provide conditioning with respect to the new modalities. Our approach introduces a new trainable cross-attention layer, which allows a new modality to develop its own learned representation during image generation.

The diffusion UNet architecture of SD comprises blocks that include residual, self-attention, and cross-attention layers. The self-attention layers capture fine-grained spatial information, while the cross-attention layers establish semantic relationships between the conditioning text and the generated image. To effectively model the relationship between audio tokens and the image, we introduce new gated cross-attention layers into this architecture (Fig. 3.6). This modification aims to fa-

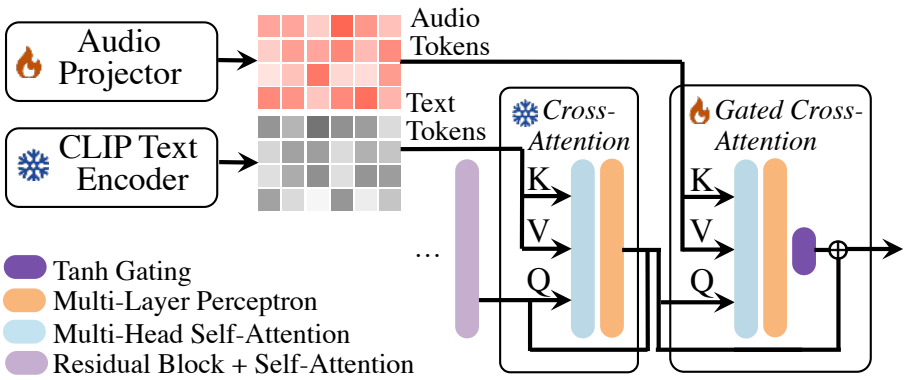


Figure 3.6: **Gated cross-attention module.** Our method guides the diffusion process by gated cross-attention and dense feed-forward layers added after pre-trained text-conditioned layers, using audio tokens as keys/values and ensuring training stability and quality.

cilitate audio-conditioned image generation by enabling image features to interact dynamically with audio tokens through these layers.

We have retained the encoder component of the UNet, while augmenting the decoder with our newly introduced gated cross-attention layers. In this configuration, only the audio projector and these new layers are set as trainable, with the rest of the network remaining frozen. The pre-trained audio projector is further refined through DDPM loss optimization, enhancing the alignment between audio and image pairs.

Within each decoder block of the UNet, we process visual tokens $\mathbf{v}$, conditioning text tokens $\mathbf{c}_{\text{text}}$, and audio tokens $\mathbf{c}_{\text{audio}}$ as follows:

$$\mathbf{v} = \mathbf{v} + \text{Residual\_Block}(\mathbf{v}), \tag{3.7}$$

$$\mathbf{v} = \mathbf{v} + \text{Self\_Attn}(\mathbf{v}), \tag{3.8}$$

$$\mathbf{v} = \mathbf{v} + \text{Cross\_Attn}(\mathbf{v}, \mathbf{c}_{\text{text}}), \tag{3.9}$$

$$\mathbf{v} = \mathbf{v} + \beta * \tanh(\gamma) * \text{Cross\_Attn}(\mathbf{v}, \mathbf{c}_{\text{audio}}) \ . \tag{3.10}$$

A key novelty in our approach is the incorporation of a gating mechanism within these new layers. This mechanism is crucial for maintaining the original denoising

efficiency of the model. We initialize the trainable parameter $\gamma = 0$ and set $\beta = 1$. During training, γ is optimized to foster gradual learning. The conditioned DDPM loss objective is applied at this stage, as defined below:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z}\sim E(\mathbf{x}),\epsilon\sim N(0,1),t,\mathbf{c}_{\text{audio}}}\left[||\epsilon - \epsilon_\theta(\mathbf{x}, t, \mathbf{c}_{\text{audio}})||_2^2\right] \tag{3.11}$$

Note that while null-text is used during tuning, our approach can incorporate an optional text prompt at inference time to augment the generation process.

In the training of the Stable Diffusion model, encoders like CLIP are typically kept frozen. However, in our model's second stage of training, we keep the audio projector module trainable to further enhance its robustness. Initially, the audio projector focuses on class separation and aligning audio features with those of the image captions. By continuing the training in the second stage, we aim to improve its ability to match image and audio features. It should be noted that we perform the second stage training using a lower learning rate of 1e-5 to preserve the initial knowledge learned in the first stage.

In our training scheme, the total number of trainable parameters is 34M, with 8M for the audio projector, and 26M for the gated cross-attention layers. This is considerably low as compared to the original UNet architecture, which has a total of 880M parameters. We further initialize the weights of the new audio-image cross attention layers with those of the text-image cross attention layers. Hence, our model achieves efficient tuning to introduce the ability to condition the network with respect to a new modality. This efficiency extends beyond model parameter storage. Another notable advantage of using adapter layers is their ability to prevent catastrophic forgetting in the pre-trained model. In our experiments, we validate this by demonstrating the use of mixed modality inputs including audio along with text prompts.

To enable classifier-free guidance, we set the CLAP embedding to zero with ten percent probability during the tuning stage. Similar to null-text embedding, we also create a "null audio embedding". Followingly, we define the classifier-free guidance

as follows:

$$\epsilon = w\epsilon_\theta(\mathbf{x}, t, \mathbf{a}) - (1 - w)\epsilon_\theta(\mathbf{x}, t_\emptyset, \mathbf{a}_\emptyset) \tag{3.12}$$

where t is the text, $t_\emptyset$ is the null text, a is the audio, $a_\emptyset$ is the null audio embedding and w is the CFG scale.

3.5 Audio-Guided Image Editing

By extending SD to perform audio conditioned image generation, our method can be combined with existing SD-based image editing methods to enable audio-guided image editing. We demonstrate this by using the Plug-and-Play features method [Tumanyan et al., 2023]. This approach involves injecting residual and self-attention features, derived from inverting a source image, into a subsequent image generation phase. This process ensures the preservation of the structure of the source image while facilitating text-conditioned edits.

In line with this approach, we inject features from the residual layers f_t^4 and attention maps A_t^l, obtained during the inversion step t, into the corresponding editing phase as follows:

$$\mathbf{z}_{t-1} = \epsilon_\theta(\mathbf{x}_t, t, c_\phi(\mathbf{y}), f_t^4, A_t^l) \tag{3.13}$$

Through the injection of self-attention features, our model captures the relationship between spatial features, preserving critical layout and shape details. By incorporating audio tokens as conditioning signals, our approach effectively facilitates audio-guided image editing.

3.6 Implementation Details

Our experiments were conducted on a single NVIDIA V100 GPU. In both stages, the AdamW optimizer with a weight decay of 1e-2 was utilized. During the first stage, the audio projector was trained using a learning rate of 1e-4. For the second stage, the cross-attention layers were trained at a learning rate of 1e-4, while the audio projector was adjusted to a lower rate of 1e-5. We used a batch size of 6. For our

image generation backbone, we used Stable Diffusion version 1.4, the same backbone all our Stable-Diffusion based competitors employ, to ensure a fair comparison in our evaluations. We employed the DDIM sampler with 50 steps and set the classifier-free guidance scale as 8 for our evaluation setup.

Chapter 4

EXPERIMENTAL SETUP

4.1 Datasets

We evaluate our approach on three different audio-visual datasets focus on three different domains (landscapes, materials and objects, and human faces). Our approach requires paired audio-visual data. Therefore, we extract frames from video datasets that correspond to specific audio. We next discuss the details of each dataset.

Landscape + Into the Wild For the landscape scenes, we create a dataset by combining videos from Into the Wild [Li et al., 2022] and High Fidelity Audio-Video Landscape (Landscape) datasets [Lee et al., 2022a], both gathered from YouTube. The Into the Wild dataset contains 94 hiking videos with nature sounds, and the Landscape dataset is composed of 928 high-resolution videos of varying lengths. The collected videos are split into 10 seconds, and a frame is extracted from each interval along with its corresponding 10-second audio. The final dataset contains 22,000 image-audio pairs. The finalized dataset contains 11 categories namely underwater bubbling, wind noise, fire crakcling, squishing water, splashing water, thunder, raining, snow, forest.

Greatest Hits [Owens et al., 2016] Greatest Hits is a video dataset that contains distinctive sounds that come from various objects when a drumstick hits them. Every hit in these videos is annotated by which material they have been hit on. There are 17 material types in total. Some of which are wood, metal, dirt, rock, leaf, plastic, cloth, and paper. In total, there are 977 videos with 46,577 hit actions. Similar to the Landscape dataset, we extract 10 seconds of audio clips and a frame from that interval to have matching audio and image pairs. However, during the recording of this dataset, some materials are being hit by turns in a random

order within a couple of seconds. To have more consecutive audios, we filter actions that have one material consecutively being hit. The final dataset contains 3,385 image-audio pairs.

RAVDESS [Livingstone and Russo, 2018] RAVDESS dataset consists of 7,356 song-video files of 24 male and female actors. The dataset includes speech samples representing emotions like calm, happy, sad, angry, fearful, surprise, and disgust. In our experiments, we took the song-video version of this dataset and extracted a single frame along with its corresponding song from each video since the videos were only 4-5 seconds long. The final dataset contains 1,008 frames with their paired audio.

4.2 Evaluation Metrics

In our study, we conduct a quantitative evaluation of methods based on two critical aspects: (i) the semantic relevance of generated images to the input audio, and (ii) the photorealism of the samples. To assess the semantic relevance, we employ three specific metrics: Audio-Image Similarity (AIS), Image-Image Similarity (IIS), and Audio-Image Content (AIC), as introduced in [Yariv et al., 2023b]. For evaluating the sample quality, we utilize the commonly used FID metric [Heusel et al., 2017]. In the following, we provide the detailed descriptions of these metrics.

Audio-Image Similarity (AIS) is designed to measure the semantic similarity between the audio input and the generated image, determining how well the image reflects the audio content. In particular, this similarity is estimated by employing the Wav2CLIP model [Wu et al., 2022], a model providing a common semantic space for images and audio like CLAP [Elizalde et al., 2023] does for text and audio. As noted in [Yariv et al., 2023b], relying solely on the similarity score between features of the input audio and the features of the generated image can be misleading due to variations in the scales of the similarity scores. Hence, AIS not only compares the generated image with its corresponding input audio but also contrasts it with audio samples from the entire validation set. Specifically, the similarity of the generated

image both with the conditioning audio clip as well as other audio clips in the validation set are computed. Then the validation audio clips which result in a lower similarity then the conditioning audio clip are identified. The ratio of such audio clips to the total number of audio clips is reported as the AIS score. Thus, AIS serves as a reference-based score and provides a more comprehensive and meaningful measure of similarity.

Image-Image Similarity (IIS) is designed to quantify the semantic similarity between a generated image and the corresponding ground truth image linked to the input audio. It serves to evaluate how accurately the generated image captures the semantic content of the target image. The scoring method employed is similar to that of AIS. The similarity score is calculated using the CLIP [Radford et al., 2021] model, comparing the generated image with both its ground truth image and image samples from the validation dataset. The IIS score for a model is then computed as an average of these similarity scores across all entries in the validation set, providing a holistic measure of the model's performance.

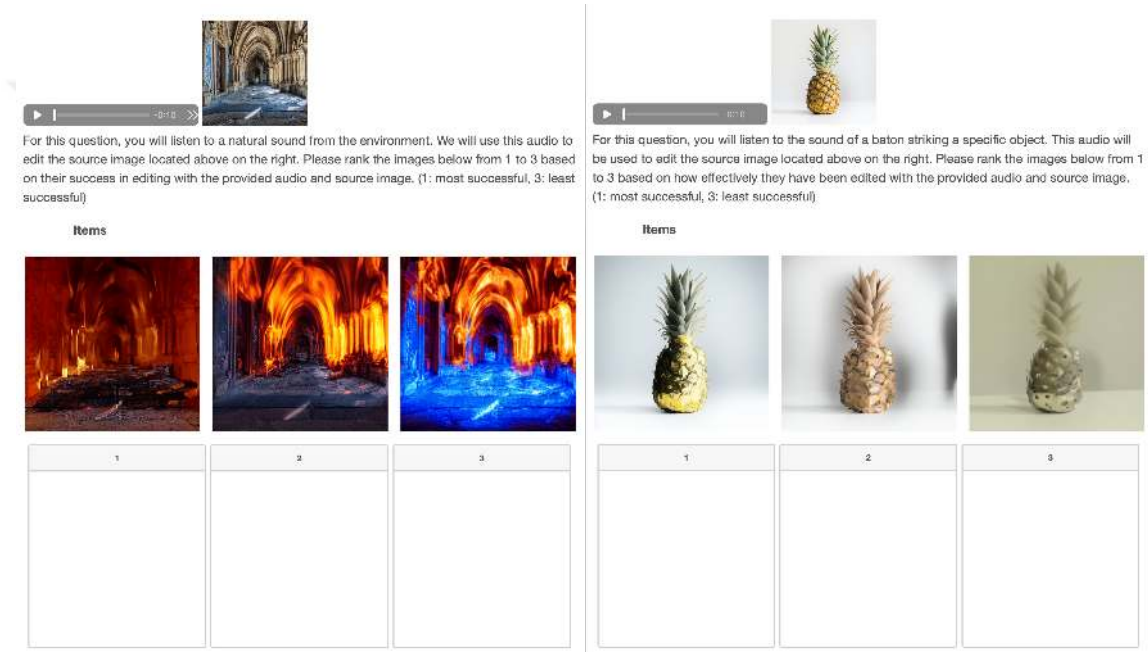
Audio-Image Content (AIC) is designed to evaluate the relevance of the content of a generated image to its corresponding ground-truth audio label. It assesses the alignment between the class predicted by an image classifier and the actual audio label. Specifically, we use CLIP as a zero-shot classifier to compute the probability of the image belonging to each of the class labels in the dataset. An agreement is noted when the ground-truth label achieves the highest probability among these. The AIC score is then calculated as the average of these agreement instances.

Fréchet Inception Distance (FID) is used to assess the perceptual quality and diversity of the generated images. In particular, it measures the photorealism by measuring the difference between the distribution of the real images and that of the generated or manipulated images using deep features.

4.2.1 User Study

To subjectively evaluate the effectiveness of SonicDiffusion, we performed a user study using the Qualtrics platform. We picked a total of 43 source images (20 for

the model trained on Greatest Hits, and 23 for the model trained on Landscape + Into the Wild) from Unsplash website [uns,] and audio inputs to generate the audio-conditioned edited versions of these images using our approach and the competing Glue-Gen [Qin et al., 2023] and AudioToken [Yariv et al., 2023b] models. In each question, the participants are asked to rank the outputs of each model according to editing quality. Screenshots of two sample questions from the user study are given in Figure 4.1.



Landscape + Into the Wild Greatest Hits

Figure 4.1: **User study GUI**. Screenshots of two sample questions from the user study where the participants are asked to rank the model outputs in terms of editing quality with respect to provided audio and the source image.

4.3 Competing Approaches

In our evaluation, we compare our SonicDiffusion model with seven state-of-the-art methods, SGSIM [Lee et al., 2022b], Sound2Scene [Sung-Bin et al., 2023], Glue-Gen [Qin et al., 2023], ImageBind [Girdhar et al., 2023], CoDi [Tang et al., 2023], AudioToken [Yariv et al., 2023b], TempoTokens [Yariv et al., 2023a]. Among these,

SGSIM and Sound2Scene utilize GANs for audio-to-image synthesis, while GlueGen, CoDi and AudioToken integrate Stable Diffusion in their image generation pipeline. ImageBind employs a DALLE·2-based generator [Ramesh et al., 2022]. TempoTokens differs from the aforementioned models as it is designed for generating videos from audio clips, again leveraging Stable Diffusion. Hence, we consider the center frames of the generated videos when evaluating TempoTokens. We finetune GlueGen, AudioToken and Sound2Scene on our datasets. However, for CoDi and ImageBind, we use the pre-trained models provided by the authors due to their large-scale nature. We do not perform finetuning for TempoTokens, as it was already trained on the Landscape dataset. For SGSIM, we do not finetune their audio encoder but train StyleGAN2-ADA [Karras et al., 2020] on our datasets to run inference. To apply AudioToken and GlueGen for image editing, we extend their generative frameworks with PnP injection [Tumanyan et al., 2023] at self-attention layers 4-11 and residual layer 4. By injecting the features, we can successfully preserve the source image structure during the generation process. We also compare our image editing results with the text-based PnP model [Tumanyan et al., 2023], using audio class labels as text prompts. This comparison shows the advantage of injecting audio-derived semantic information into image manipulation, as opposed to relying solely on text labels. Below we provide training details of the aforementioned competing approaches.

SGSIM We first trained a separate StyleGAN2-ADA model for each of the three datasets using adaptive augmentation. For Landscape + Into-the-Wild, the model is trained for around 800K steps. For Greatest Hits and RAVDESS, their corresponding models have converged around 25K steps and 40K steps so we use these checkpoints.

AudioToken We finetuned their embedder and Lora weights with all of our datasets. Although finetuning was not available within their provided codebase, we managed to add this feature. During finetuning, we used the parameters suggested in their GitHub repository for training. For Greatest Hits and RAVDESS, we finetuned for

60K steps rather than 30K as we observed that the models finetuned with 30K steps resulted in worse performance than the models finetuned with 60K steps. We used same parameters during generation and editing and applied injection at the same layers during editing.

Sound2Scene We used default parameters in their training script and finetuned their model for 100 epochs on each dataset. As their pre-trained generator was trained for 128×128 images, we generated 128×128 images for comparison with our model.

GlueGen We finetuned the GlueNet model on audios from all three datasets for 30 epochs, with a learning rate of 2e-5 and kept the rest of the parameters as the suggested defaults.

ImageBind We did not perform finetuning on ImageBind. We demonstrated generation results of ImageBind's shared embedding space using Stable Diffusion UnCLIP[1].

CoDi We did not perform finetuning on CoDi. We used the official repository and checkpoints to have audio to image generation results.

[1]https://github.com/Zeqiang-Lai/Anything2Image

Chapter 5

RESULTS

5.1 Audio-Driven Image Generation.

In Fig. 5.1, we present a visual comparison between our model and the competing methods. Our model demonstrates a superior capability in generating images that are not only visually coherent but also closely aligned with the input audio cues. For example, landscape images synthesized by our model capture the essence of the target scenes more accurately. When processing audio encoding material properties, our model synthesizes images featuring objects or elements that reflect the desired attributes. Additionally, the human face images generated from speech audio effectively mirror the corresponding vocal tonations, capturing the intended emotions.

Table 5.1 supports these qualitative observations with quantitative evidence. Our model consistently outperforms competing approaches in producing photorealistic images, demonstrated by its superior FID scores across all datasets. Notably, our landscape images exhibit a remarkable alignment with the audio semantics. In the Greatest Hits dataset, our model achieves the highest AIC score and the second-highest IIS score, further showing its effectiveness in audio-driven image generation.

Moreover, the versatility of our in handling mixed-modality inputs, blending both text and audio, is illustrated in Fig. 5.2. These results demonstrate our method's proficiency in interpreting and fusing concepts from varied modalities. For example, it successfully merges a *'watercolor painting'* text prompt with a *'squishing water'* audio input, resulting in accurately composed and distinct image.

5.2 Audio-Driven Image Editing.

Our model also excels in editing real images based on audio clips. Fig. 5.3 illustrates various examples of our method's image editing results. These show our model's

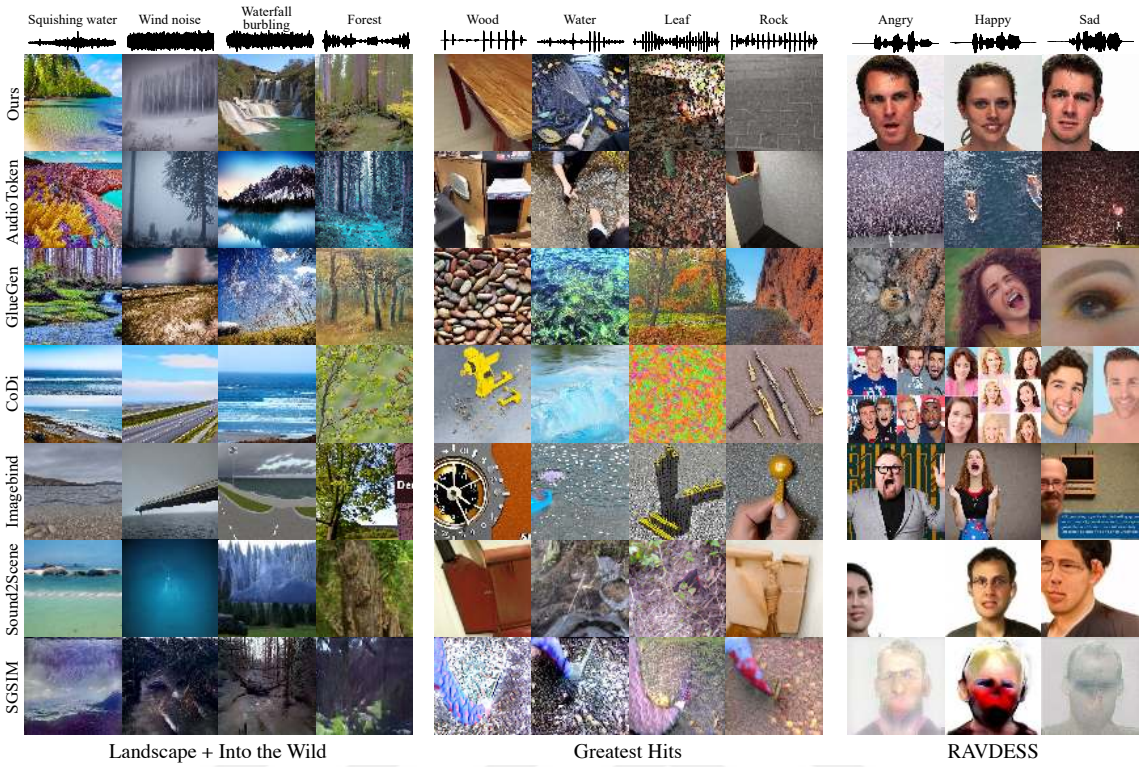


Figure 5.1: **Comparison against the state-of-the-art audio-driven image synthesis methods.** Our model generates images that closely align with the semantics of the input audio clips, surpassing all other existing methods in performance and fidelity.

ability in modifying image content to align with the semantics of the provided audio. For example, when editing a landscape image with the sounds of *'waterfall burbling'*, *'fire cracking'* and *'rain'*, our method makes the necessary changes in the overall look of the input image. In another instance, as shown in the second row, the model successfully modifies the visual appearance of the objects based on the audio inputs reflecting material characteristics like *'plastic bag'*, *'rock'* and *'paper'*.

In Fig. 5.4, we show a comparative analysis of our model's editing capabilities against existing methods, including the text-based image editing PnP model, which serves as a foundational component of our framework. For this comparison, we employ class label information from audio inputs as text prompts in the PnP model. The results show that SonicDiffusion surpasses others in manipulation quality and style transfer accuracy. For instance, in editing the second image, methods like AudioToken and GlueGen not only fall short in achieving convincing manipulations

	Model	AIS↑	AIC↑	IIS↑	FID↓
Landscape + ItW	SGSIM	.7224	.2166	.6898	220.3
	Sound2Scene	.7466	.3672	.8894	122.2
	GlueGen	.6632	.4618	.7357	133.0
	ImageBind	.7209	.4600	.8044	159.1
	CoDi	**.7578**	.3618	.7749	134.6
	AudioToken	.6983	.2851	.8592	141.3
	TempoTokens	.7446	.2242	.6215	258.0
	SonicDiffusion (Ours)	.7390	**.5436**	**.8898**	**118.6**
Greatest Hits	Sound2Scene	.6536	.3125	.6693	143.1
	SGSIM	.5065	.2000	.6612	239.2
	GlueGen	.5047	.5050	.5976	208.9
	ImageBind	.5736	.215	.5889	186.0
	CoDi	.5694	.2325	.5721	218.0
	AudioToken	**.6552**	.2675	**.7541**	123.6
	SonicDiffusion (Ours)	.6237	**.6050**	.7411	**123.5**
RAVDESS	Sound2scene	5053	.1257	5301	140.3
	SGSIM	.5090	.1108	.5094	155.4
	Gluegen	.5052	.2252	.5059	247.1
	Imagebind	**.5802**	.2131	.6332	248.8
	CoDi	.5292	.1396	.5674	229.1
	AudioToken	.5009	.1821	.6409	279.4
	SonicDiffusion (Ours)	.5309	**.2316**	**.8736**	**89.6**

Table 5.1: **Quantitative comparison of our proposed approach with existing audio-conditioned image generation methods**, focusing on AIS, AIC, ISS for semantic consistency, and FID for image quality. Top scores are **bolded**, second best are underlined. Pre-trained large-scale models and audio-to-video generation model are highlighted in blue and yellow, respectively.

but also introduce artifacts and color shifts. Similarly, the text-based PnP model fails to capture the full extent of the modifications implied by the audio inputs, unlike our model.

Quantitative results in Table 5.2 further validate our approach. SonicDiffusion achieves the highest scores in AIS and ISS metrics, and the second-highest in IIS,

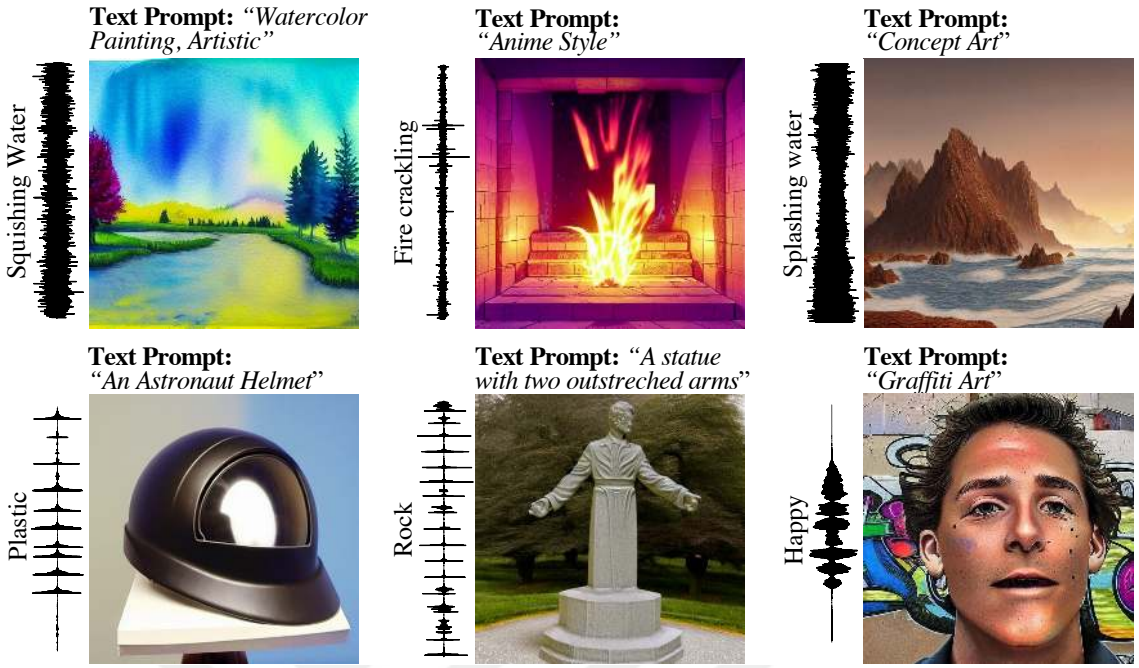


Figure 5.2: **Image generation results using both text and audio.** Our model enables mixing text and audio modalities while synthesizing novel images.

confirming its ability to accurately manipulate images in response to audio cues. It also leads in FID scores, indicating better image quality compared to existing approaches. In a user study, our results are preferred by a majority of participants over competing approaches. Specifically, 56.05% of 19 participants favored our method over GlueGen (17.37%) and AudioToken (26.58%) for the Greatest Hits samples. Similarly, for the Landscape+Into the Wild samples, 64.95% of 16 participants rated our results higher than those by GlueGen (22.83%) and AudioToken (12.23%).

We also show the ability of our model to control image editing outcomes through simple manipulations of input sounds. Our model can blend characteristics from two distinct audio sources by linearly interpolating their audio embeddings, as illustrated in Fig. 5.5(a). This process results in outputs representing seamlessly transitions between features corresponding to each audio input. Additionally, as depicted in Fig. 5.5(b), our model responds to changes in audio volume, allowing for precise manipulation of image content. The degree of change in the image directly corre-

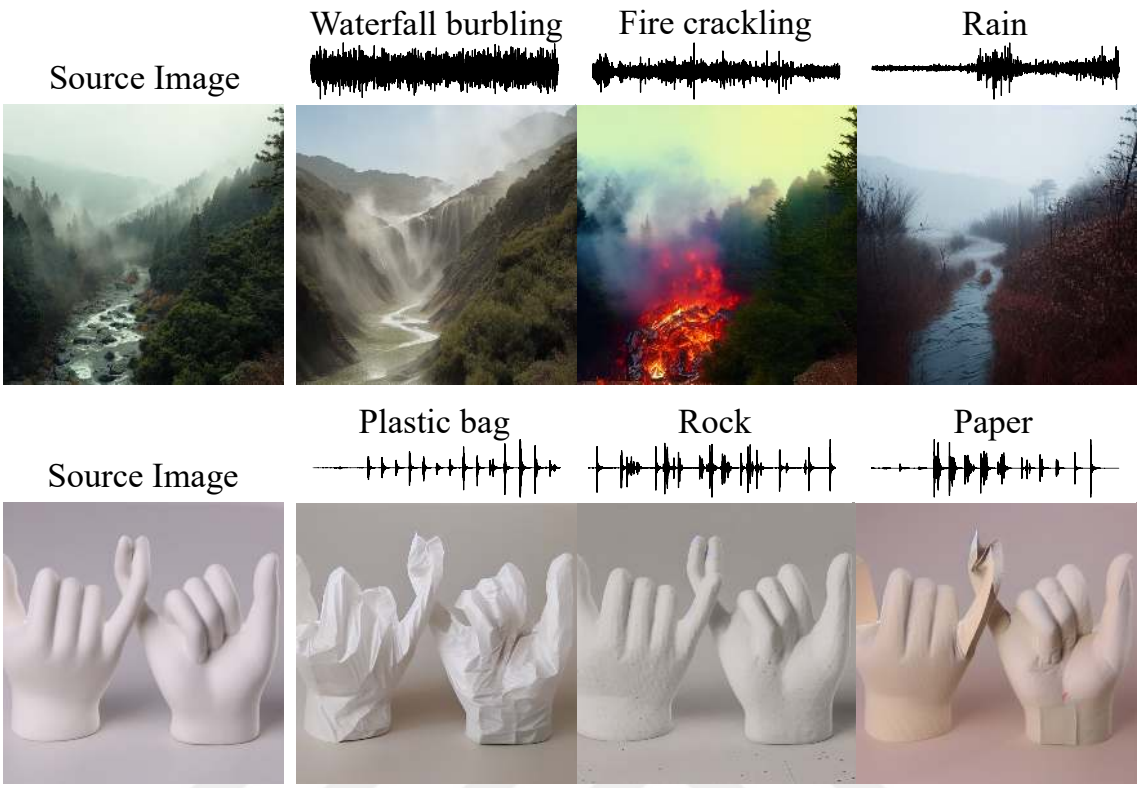


Figure 5.3: **Audio-guided image editing results.** Our model successfully performs manipulations that semantically reflect the audio content while maintaining the content of the source images.

	Model	AIS ↑	AIC ↑	IIS ↑	FID ↓
L + ItW	PnP	-	**.6833**	-	239.2
	GlueGen	.5869	.2254	.6491	250.3
	AudioToken	.6059	.1849	.7405	206.8
	SonicDiffusion (Ours)	**.6496**	.3745	**.7933**	**172.4**
GH	PnP	-	**.6604**	-	199.5
	Gluegen	.4680	.2275	.5425	208.9
	AudioToken	.4749	.1450	.5441	206.8
	SonicDiffusion (Ours)	**.5491**	.3575	**.6255**	**152.7**

Table 5.2: **Quantitative Results for Image Editing**. We evaluate semantic consistency with AIS, AIC, and ISS metrics, and image quality with FID. The highest scores are in bold, with second-best scores underlined. The text-based model is highlighted with red.

Figure 5.4: **Comparison against the state-of-the-art audio-driven image editing methods.** Our model generates images that closely align with the semantics of the input audio clips, surpassing all other existing methods in performance and fidelity.

lates with the volume adjustments, offering users detailed control over the audio's influence on the image.

5.3 Ablation Studies and Further Analysis

5.3.1 Impact of Loss Functions in the Training of SonicDiffusion's Audio Projector

In the first training stage of our model, we employ two loss functions: (i) contrastive loss and (ii) mean squared error (MSE) loss. We conducted ablation studies to understand the effects of omitting one of these losses. Table 5.3 presents the results of these experiments. When the contrastive loss is removed, our network struggles to distinguish between different audio conditionings. On the other hand, omitting the MSE loss results in the model still being able to differentiate between various audio cues, but with a longer convergence time and reduced accuracy. Including

(a) Sound mixing

(b) Volume changes

Figure 5.5: **Controlling Editing with Sound Manipulation:** Our model can (a) blend two sounds via linear interpolation of audio embeddings, and (b) adjust manipulation intensity by varying sound volume.

the MSE loss is particularly crucial as it facilitates the alignment of the audio space with the pre-trained CLIP space, though not through a direct mapping due to the simultaneous application of the contrastive loss. In essence, the contrastive loss adds another layer of complexity to this alignment. After completing the first stage of training, we continue to train the audio projector module but with a reduced learning rate. This step is vital as it enables the network to directly capture the semantic correlations between audio and image, enhancing the overall effectiveness of our SonicDiffusion model.

5.3.2 Impact of Different Training Strategies on the Performance

Our goal is to establish a conditioning space that effectively guides the diffusion process. Leveraging embeddings from a pre-trained, well-represented space is a recognized strategy in diffusion models. Large text encoders such as CLIP [Radford et al., 2021] and T5 [Raffel et al., 2020] are commonly employed for this purpose due to their robust pre-trained capabilities. In our approach, we extract audio fea-

	Landscape				Greatest Hits			
	AIS ↑	AIC ↑	IIS ↑	FID ↓	AIS ↑	AIC ↑	IIS ↑	FID ↓
SonicDiffusion	.7390	.5436	.8898	118.6	.6237	.6050	.7411	123.5
w/o Contrastive Loss	.7245	.3919	.8432	121.6	.6560	.3694	.7264	133.4
w/o MSE Loss	.6827	.4018	.7767	124.9	.6269	.4027	.6826	142.5

Table 5.3: Impact of various loss functions employed in training the Audio Projector module on the performance of our SonicDiffusion model. We report the performance using both contrastive and MSE losses (first row), using only the MSE loss (second row), and using only the contrastive loss (third row).

tures from CLAP [Elizalde et al., 2023] and consider an additional learning phase to map these features from the CLAP embedding domain to the CLIP embedding domain. To assess the impact of various training strategies on our model's performance, we conduct a series of experiments. These experiments focus on evaluating changes in performance under different conditions: representing audio features with a single token, mapping CLAP embeddings to CLIP space solely using diffusion loss (thereby omitting the first stage of training), and freezing the Audio Projector's parameters during the second stage of training. The outcomes of these experiments are summarized in Table 5.4. We found that relying exclusively on the diffusion loss for learning the mapping led to a notable decrease in the model's performance. Notably, the model seems to lose its ability to distinguish between different audio inputs. Freezing the Audio Projector does not consistently improve results and can even lower image quality, as indicated by the FID scores. Additionally, we also observed that representing audio with just a single token, as in AudioToken [Yariv et al., 2023b], leads to worse performance.

5.3.3 Impact of Where to Insert Gated Cross-Attention Layers

Stable-Diffusion v1.4 UNet configuration has sixteen text-image cross attention layers. Six of them lies inside encoder side, one of them lies inside the middle block, and nine of them lies inside the decoder side within blocks of 3 to 11. In SonicDiffusion,

	Landscape + Into the Wild				Greatest Hits			
	AIS ↑	AIC ↑	IIS ↑	FID ↓	AIS ↑	AIC ↑	IIS ↑	FID ↓
SonicDiffusion	.7390	.5436	.8898	118.6	.6237	.6050	.7411	123.5
w/o First Stage Training	.6013	.0600	.5160	116.7	.4376	.0888	.5128	177.6
w/ Audio Projector Frozen in Stage Two	.7488	.6020	.8920	128.3	.5849	.5972	.7188	128.3
w/ Single Token Audio Representation	.7042	.4618	.8281	136.6	.6863	.4625	.7647	119.0

Table 5.4: Impact of various training choices on the performance of our proposed SonicDiffusion model.

we inject gated cross-attention layers only into the middle block and decoder blocks. We perform an analysis on this design choice and investigate the impact of where to insert gated cross-attention layers within our SonicDiffusion model. Table 5.5 summarizes the results of this analysis. As can be seen, injecting gated cross attention layers to only for the sixth to eleventh blocks of the UNet significantly diminishes the model's effectiveness, whereas injecting them only to the encoder blocks gives only slight improvements.

	Landscape + Into the Wild				Greatest Hits			
	AIS ↑	AIC ↑	IIS ↑	FID ↓	AIS ↑	AIC ↑	IIS ↑	FID ↓
SonicDiffusion	.7390	.5436	.8898	118.6	.6237	.6050	.7411	123.5
w/ injection to layers 6-11	.7089	.4242	.8479	145.4	.5968	.4027	.6638	160.5
w/ injection to all layers	.7471	.5676	.8931	125.6	.6148	.6250	.7542	132.5

Table 5.5: Influence of where to inject gated cross-attention layers within our SonicDiffusion model on model performance.

5.4 Inference Time Choices

Diffusion-based image generation offers extensive controllability over various aspects of image synthesis – an attribute our SonicDiffusion model seeks to expand upon. In diffusion models, the classifier-free guidance (CFG) mechanism is used to modulate the influence of the conditioning factor. In our model, we harness CFG to finely

balance the influence of both audio and text inputs. Additionally, we achieve finer control over these two modalities. As detailed in Equation 9, adjusting the β value enables us to modulate the relative importance of audio versus text in the generation process (Figure 5.6). Interestingly, the β value can be varied across different layers of the UNet architecture, providing layer-specific control, which we haven't fully explored.

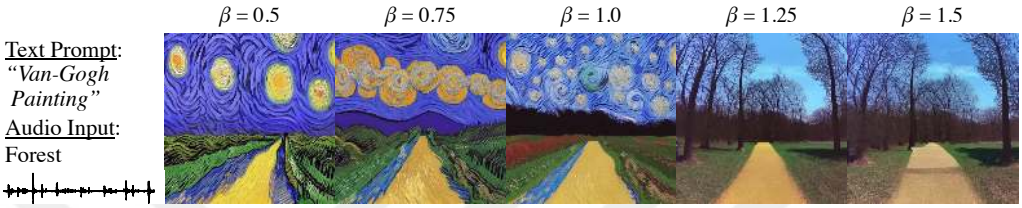


Figure 5.6: **Influence of the β coefficient.** As shown in this example, setting the value of β specifies the strength of how the generated process will be modulated in relation to the audio versus text input.

Previous studies, such as plugandplay, have identified that the earlier layers of the UNet decoder are responsible for low-frequency image information, while the higher layers manage high-frequency details. We observe a similar phenomena for our injected gated cross-attention layers. In Figure 5.7, we plot the norms of our adapter block injected at different decoder layers along the diffusion steps. The norms in the middle block suggest that low-frequency information is predominantly introduced in the early time steps. Conversely, in layers such as 7, 8, and 9, higher norms are observed during the middle time steps, indicating that these layers contribute different types of information. This layer-specific variation indicates that these layers are crucial for conveying different types of information during the generation process.

In the methodology we follow for image editing, the layers that we choose to inject features and the timesteps that we inject features are also design choices. We compare our results with other baselines using self attention layers 4-11 and residual layers 4 to have standard layers selected in the original PnP paper. However, some of the results are better when we use more residual injections like 4-11 self-attention and 4-6 residual blocks. This helps to preserve the original image structure better.

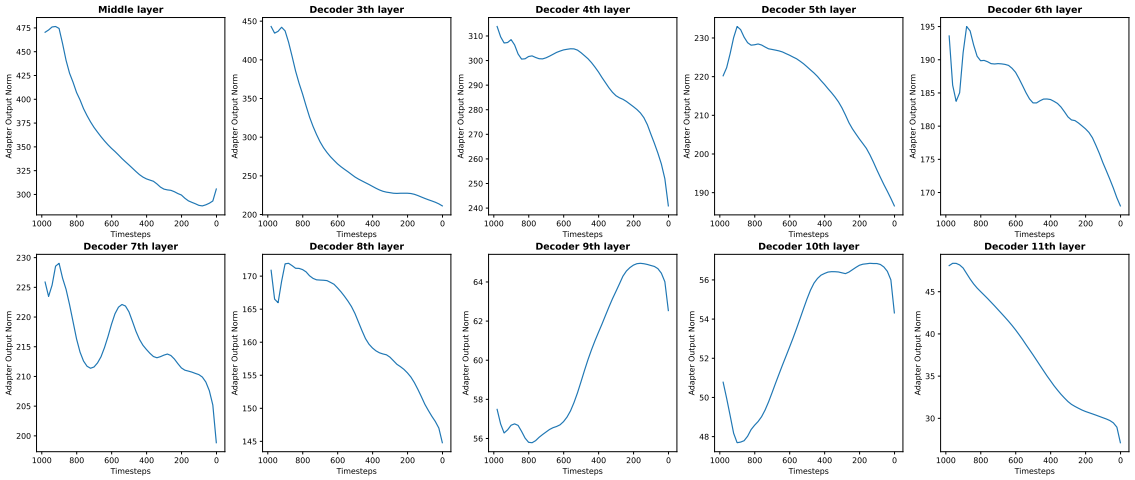


Figure 5.7: **Norms of gated cross attention outputs across UNet layers**. Plots reveals the introduction of low-frequency information in early layers at the initial time steps where as high-frequency details are encoded in higher layers, illustrating the distinct contributions of each layer in the image generation process.

5.5 Additional Results

In this section, we provide additional qualitative results obtained with our proposed SonicDiffusion model and further comparisons against the competing approaches. Figure 5.8, 5.9, 5.10 present audio-driven image generation results using audio inputs from the Landscape + Into the Wild, Greatest Hits and RAVDESS datasets, respectively. Furthermore, we demonstrate the audio-driven image editing capabilities of our SonicDiffusion model in Figure 5.11, 5.12, 5.13, each utilizing audio inputs from the Landscape + Into the Wild, the Greatest Hits, and the RAVDESS datasets, correspondingly. In Figure 5.14 and Figure 5.15, we provide qualitative comparisons of our SonicDiffusion against the state-of-the-art audio-driven image generation and audio-guided image editing models, respectively. Additionally, Figure 5.18 demonstrated the dynamic control our model offers in editing results through simple manipulations of input sounds, such as adjusting audio volume and blending two different audio sources via linear interpolation of their embeddings. Lastly, Figures 5.19 and 5.20 show sample images generated using mixed modality inputs that blend both text and audio, illustrating the versatility of our SonicDiffusion model in handling

multimodal data.

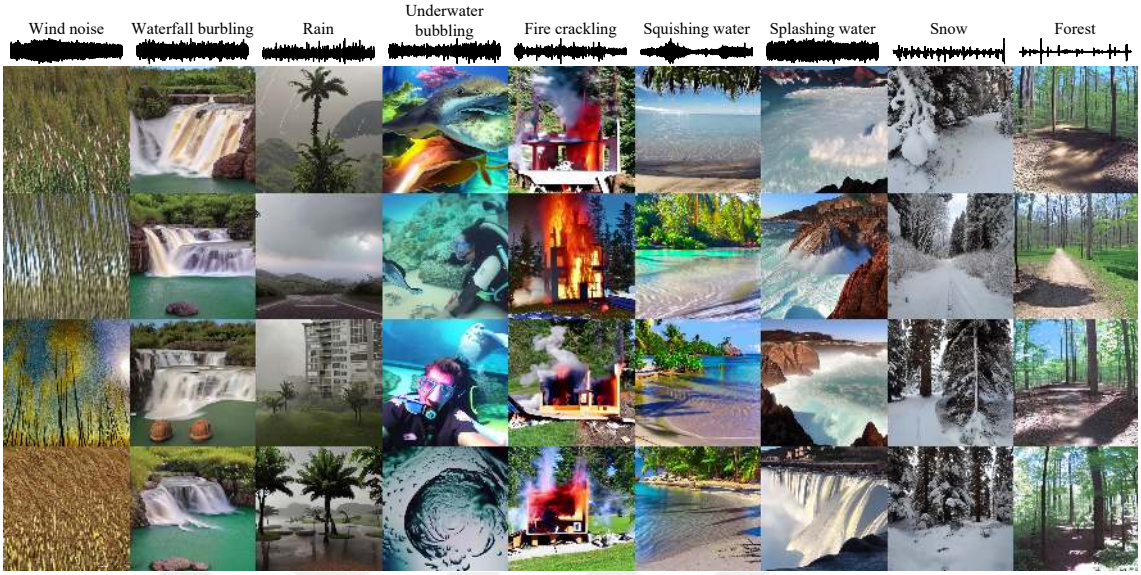


Figure 5.8: Additional audio-driven image generation results obtained with different seeds and audio clips from the Landscape + Into the Wild dataset.

5.6 Limitations and Failure Cases

Our SonicDiffusion model demonstrates an encouraging capacity for image generation that aligns with the semantic context of provided audio input. Nonetheless, as illustrated in Figure 5.21, it is not without its shortcomings. In particular, the model occasionally does not understand the essence of the audio content. One significant source of these suboptimal outcomes is the artifacts intrinsically produced by the underlying Stable Diffusion framework. This is evident when processing the Greatest Hits dataset, where the model's rendition of hands and objects in motion often falls short. Similarly, with the RAVDESS dataset, which primarily features human facial expressions, the model sometimes struggles with accurately rendering facial features, such as teeth, resulting in clearly noticeable anomalies. When it comes to image editing, the shortcomings become more pronounced with the employment of the DDIM inversion method. This technique is liable to obliterate delicate details in intricate scenes or integrating extraneous elements, thereby skewing the image's

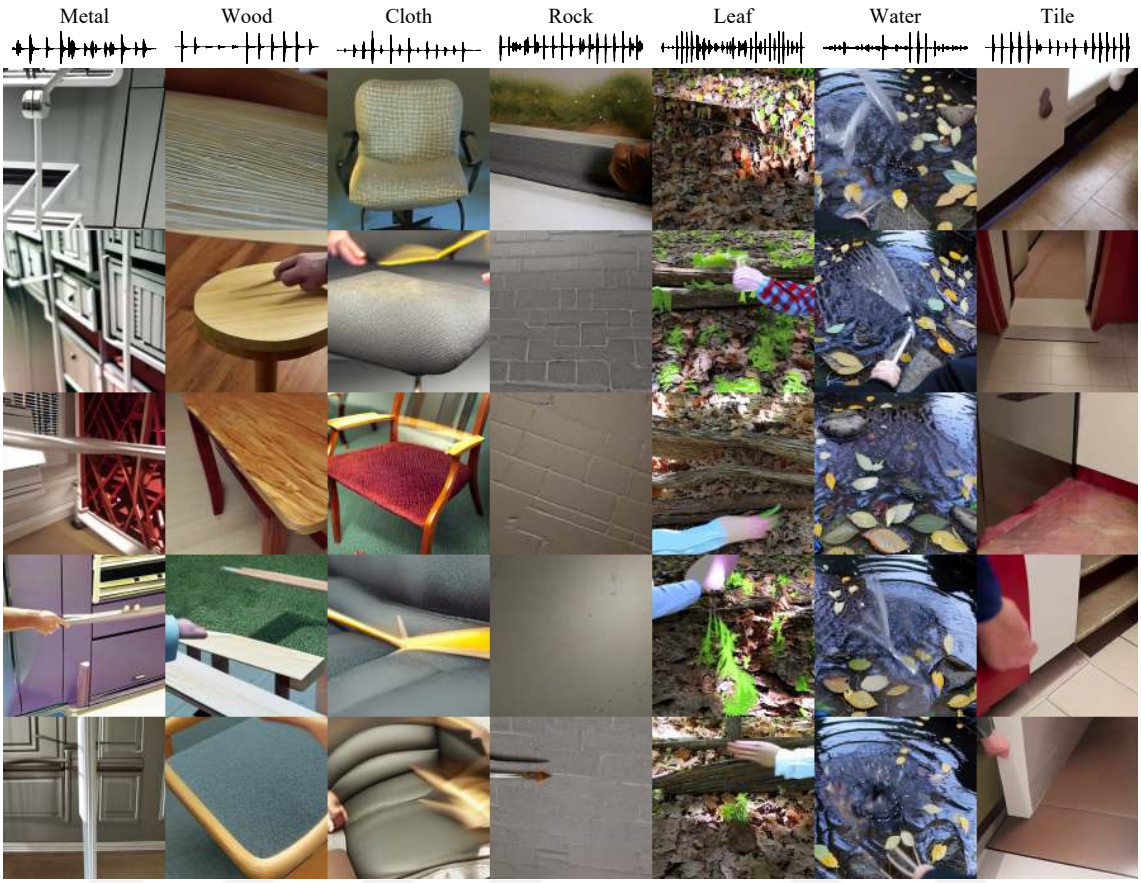


Figure 5.9: Additional audio-driven image generation results obtained with different seeds and audio clips from the Greatest Hits dataset.

authentic structure. Instances of such editing impediments are exemplified in the third row of Figure 5.21.

5.7 Broader Impact

The broader impact of our SonicDiffusion model spans several domains, reflecting a significant step forward in the integration of audio signals for image synthesis and editing. This work opens new possibilities for creating visual content that resonates with the emotional and semantic nuances of sound, offering novel applications in digital art, media production, and virtual reality. For instance, our model can enhance the accessibility of visual media for those with hearing impairments by translating sound into corresponding visual narratives. Additionally, it paves the way for more

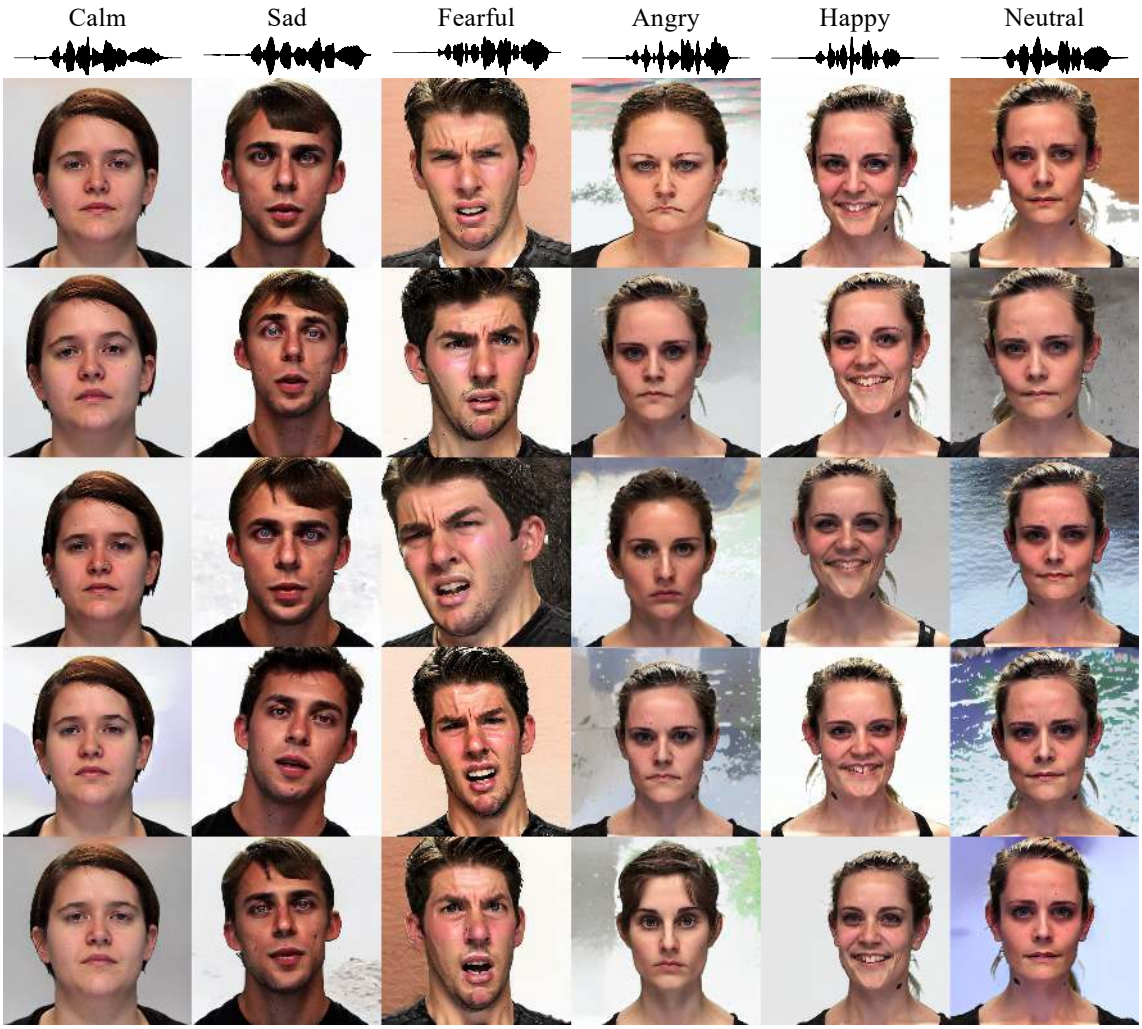


Figure 5.10: Additional audio-driven image generation results obtained with different seeds and audio clips from the RAVDESS dataset.

intuitive human-computer interaction where users can guide visual creation through voice and sound, making technology more accessible to those without expertise in complex image editing software.

Moreover, our method holds potential for advancing the field of automated content generation, where audio tracks can influence the generation of dynamic visual scenarios, such as in video games or immersive simulations. It also contributes to the research community by providing a new direction for multimodal learning, encouraging further exploration into the interplay between different sensory inputs and computational creativity.

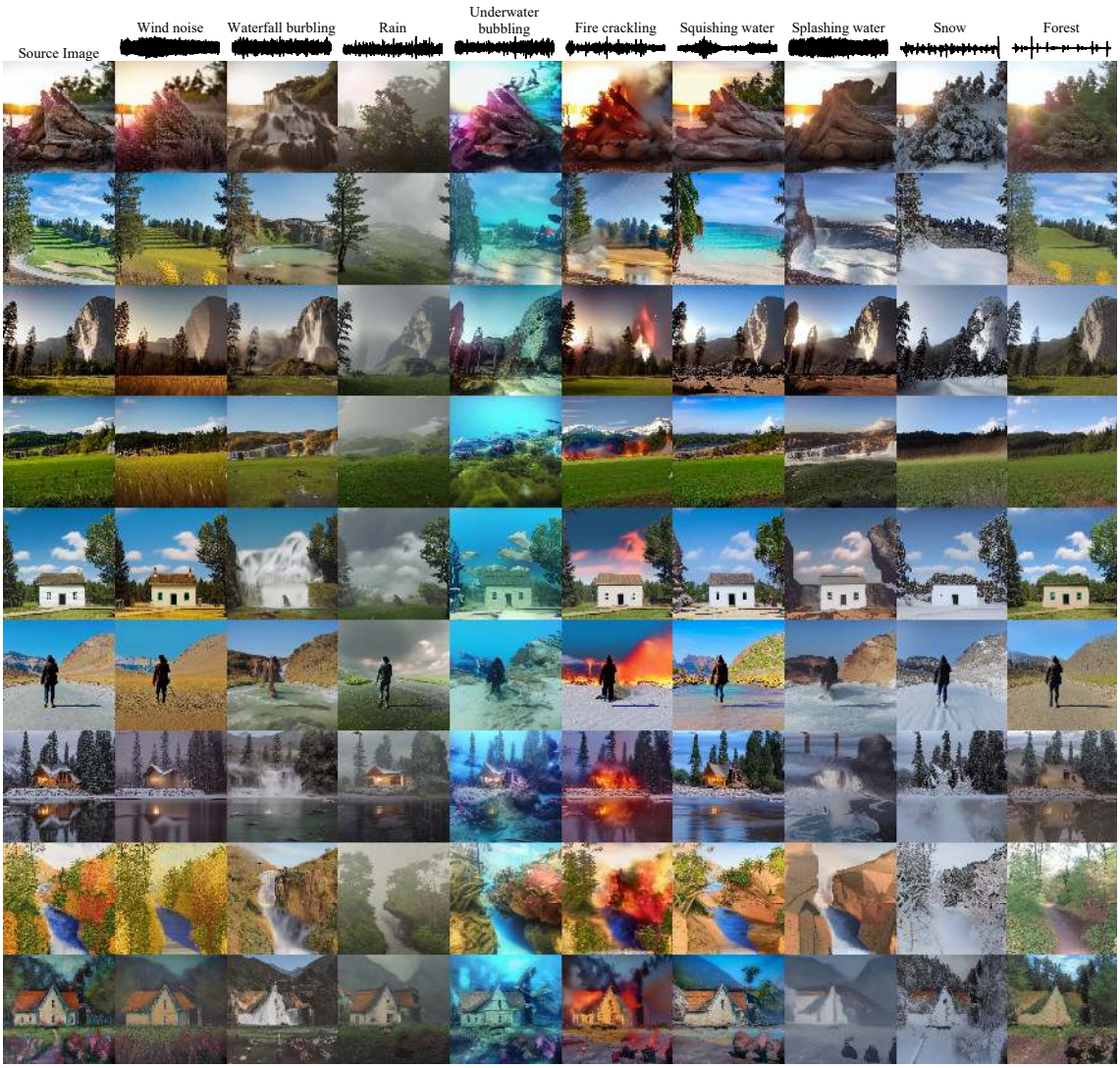


Figure 5.11: Additional audio-guided image editing resulting on a variety of source images influenced by different audio clips from the Landscape + Into the Wild dataset.

However, it is also essential to acknowledge and address potential ethical considerations, such as the use of this technology to create deepfakes or other forms of misleading content. Responsible usage guidelines and further research into detection and prevention mechanisms are crucial to ensure the positive impact of this technology on society.

Figure 5.12: Additional audio-guided image editing results on a variety of source images influenced by different audio clips from the Greatest Hits dataset.

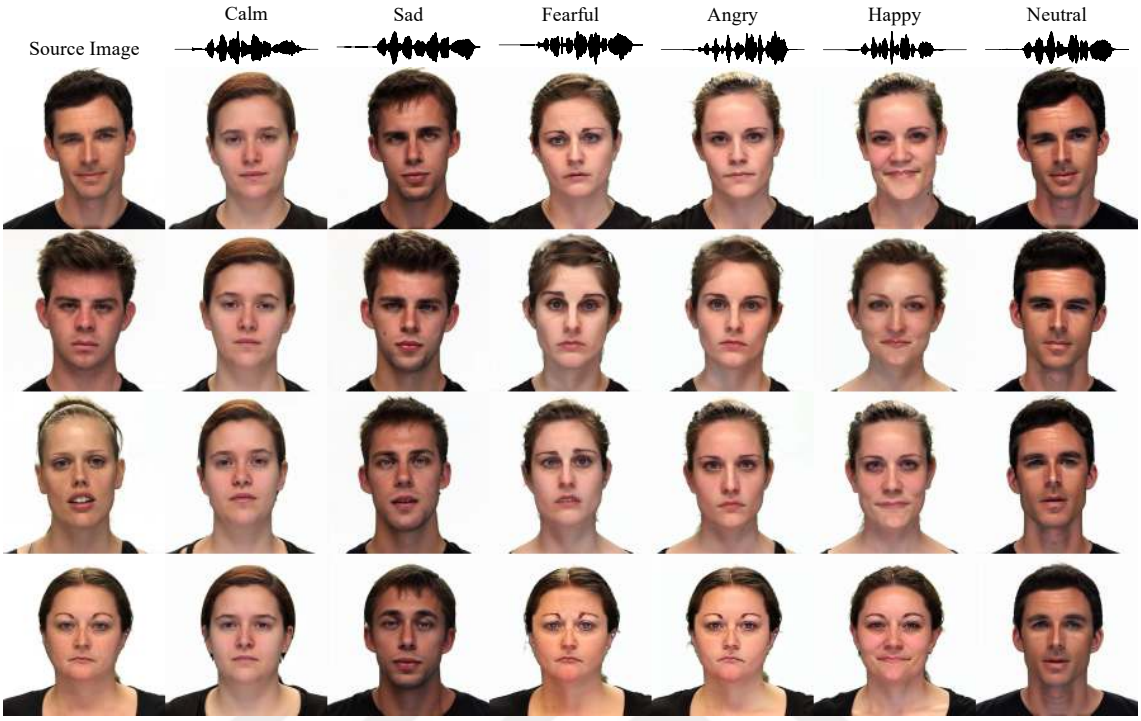


Figure 5.13: Additional audio-guided image editing results on a variety of source images influenced by different audio clips from the RAVDESS dataset.

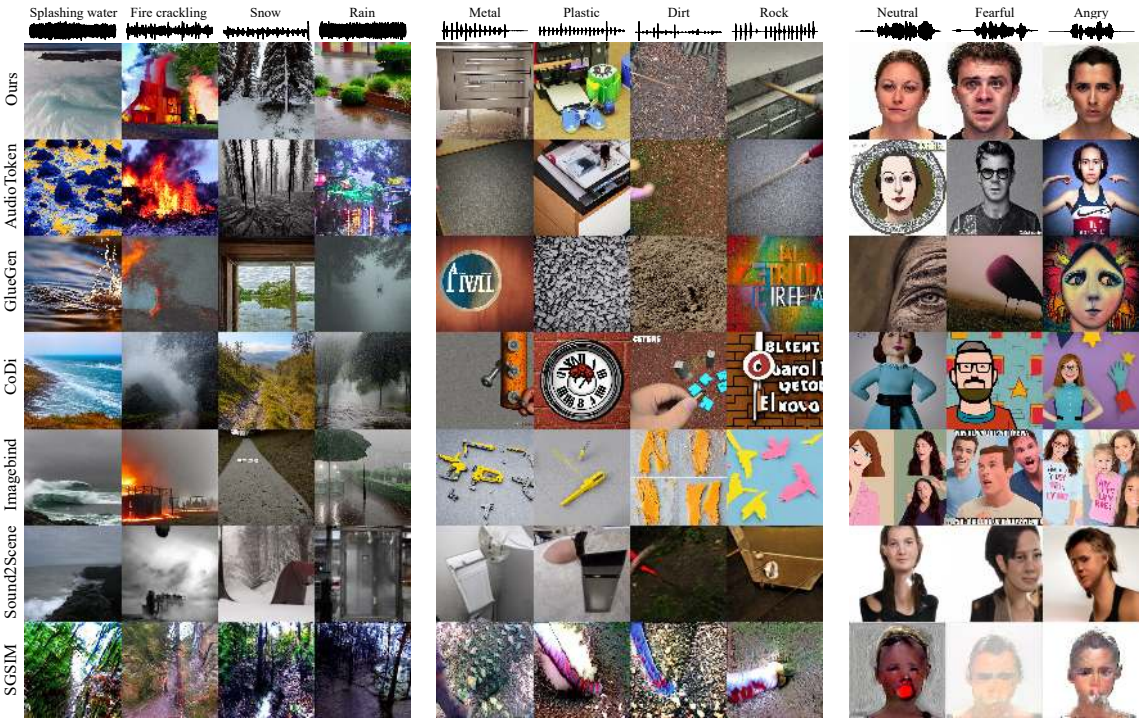


Figure 5.14: Additional comparsions against the current state-of-the-art audio-driven image generation approaches.

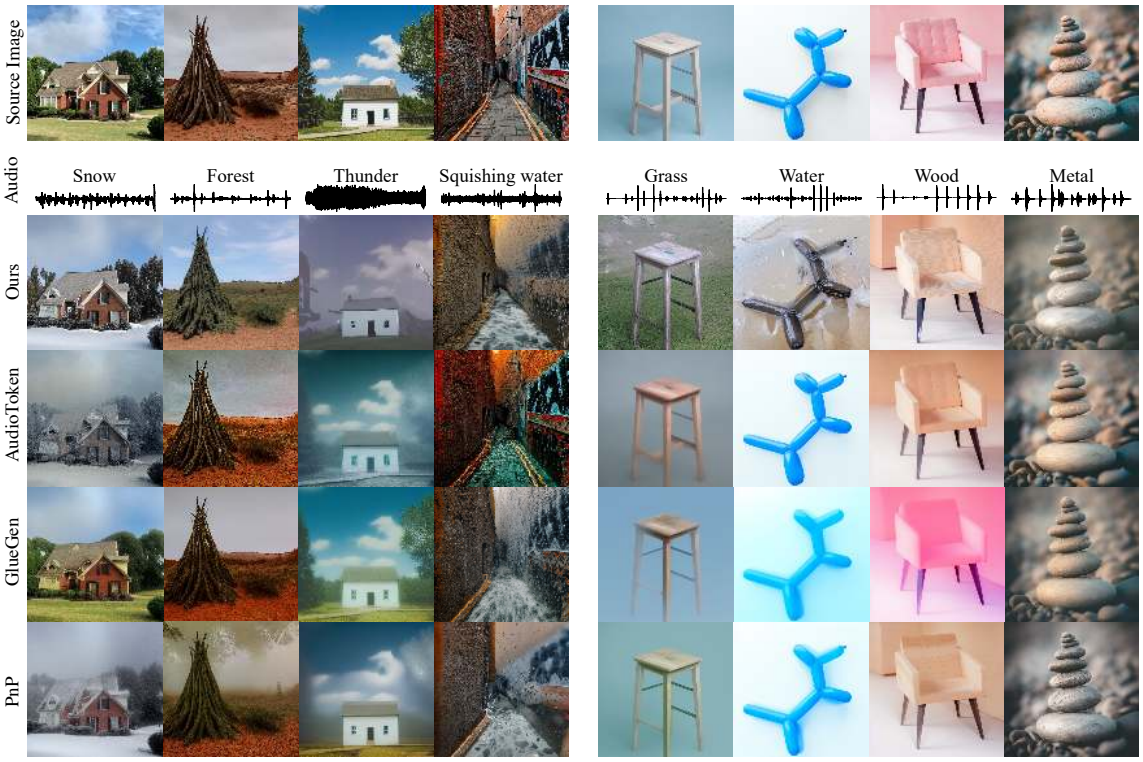


Figure 5.15: Additional comparsions against the current state-of-the-art audio-driven image manipulation approaches.

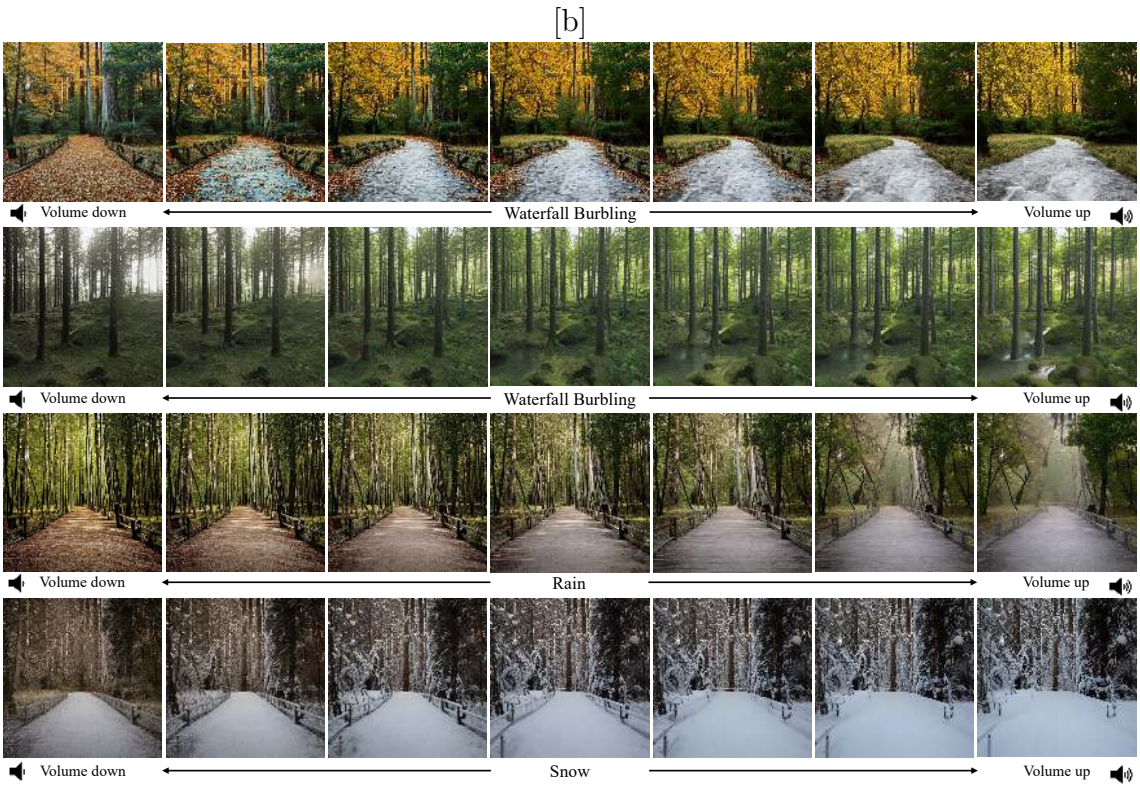


Figure 5.16: Volume Changes

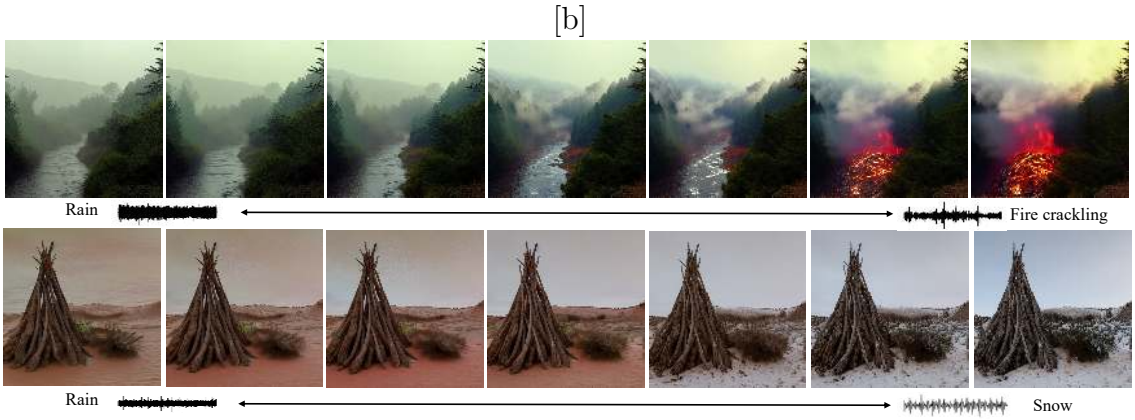


Figure 5.17: Sound Interpolation

Figure 5.18: Audio Interpolation and Mixing Impact on Image Editing: The effectiveness of the editing process is illustrated by (a) applying linear interpolation between two audio tracks, as shown in the figure, and (b) adjusting the intensity of the edits by manipulating the sound volume.

Figure 5.19: The integration of additional text prompts, representing different image styles, changes the overall appearance of the generated images, yet preserves the visual content conveyed by the audio inputs.

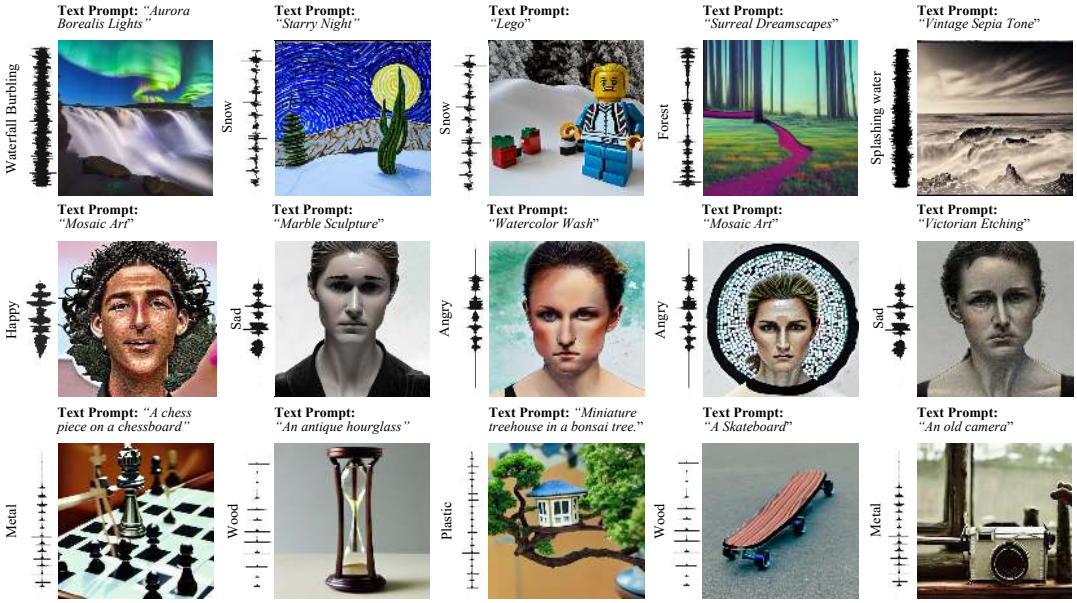


Figure 5.20: Sample outputs of the proposed SonicDiffusion model with mixed modality inputs. These images demonstrate our model's capability to synthesize novel images by harmonizing audio and text inputs, showing its adaptability across various text prompts and audio tracks.

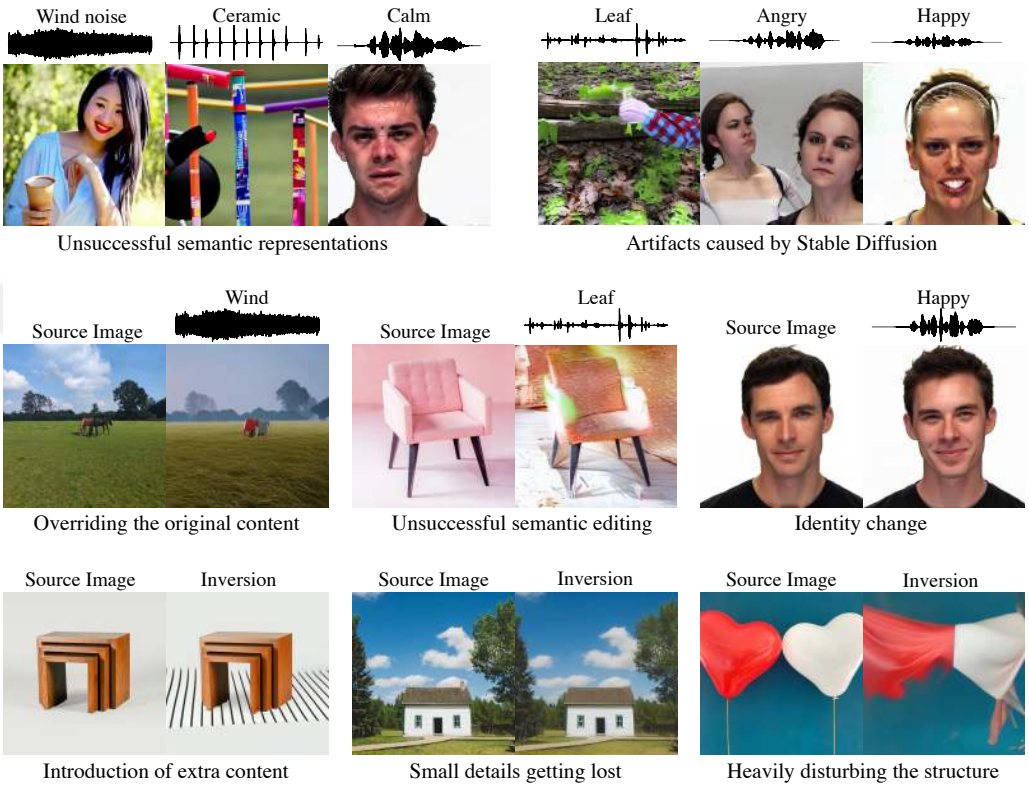


Figure 5.21: **Limitations of the SonicDiffusion model.** The figure illustrates various scenarios where the SonicDiffusion model does not adequately perform in the generation or modification of images based on audio cues. Issues arise from imprecise semantic interpretations of audio signals, as well as from artifacts introduced during the Stable Diffusion process. In the context of image editing, our model may inadvertently replace original content, conduct ineffective semantic modifications, or alter the subject's identity. Notably, some of the editing challenges are attributed to the use of DDIM inversion, which can result in the insertion of extraneous elements, the omission of fine details, or significant structural disruptions.

Chapter 6

CONCLUSION

6.1 Concluding Remarks

In this thesis, we introduced the SonicDiffusion model, an original framework for sound-guided image generation and editing, leveraging the robustness of the SD model. We proposed an audio projector that takes audio clips and creates meaningful tokens. We integrated these tokens seamlessly into the generation process with adapted cross-attention layers in SD. Since the original architecture and weights of the pre-trained model are untouched during this operation, we introduced the minimal possible overhead. The proposed framework allows for incorporating audio context that aligns with the input sound, effectively enhancing image generation and editing tasks.

We showcased our results over three distinct datasets. Results of the landscape dataset prove that our model is capable of utilizing audio clips such as rain, splashing waves, and fire crackling to portray entire scenes. The Greatest Hits dataset illustrates the sensitivity of our approach in recognizing sounds stemming from specific objects being hit. Our model effectively represents the impact of these hit actions by altering the texture of a target image. Moreover, our approach exhibits the ability to capture subtle meanings in emotional representations of human faces. We demonstrated the results of this ability on the RAVDESS dataset. We made extensive qualitative and quantitative comparisons with state-of-the-art generative methods that synthesize images from audio signals. Our results proved that our model can outperform these baselines both qualitatively and quantitatively. While doing all these our approach preserves initial capabilities of the text-to-image model, which enables us to generate images guiding by audio and text together.

We verified our design decisions through various ablation studies for our choices

in loss objectives, training procedures, and architectural choices. We also presented the cases where our model fails to deliver successful results and provide insights into the reasons behind these failure cases. We discussed the broader impact of this work by highlighting the potential use cases of this methodology, and potential ethical considerations.

6.2 Future Directions

The successful integration of audio cues into the image generation process opens up possibilities for more multi-model interactions. A direct progression of this work is combining audio, image, and text conditionings simultaneously. Currently, our model can effectively combine global meanings of text and audio modalities. The potential incorporation of our approach with ControlNet [Zhang et al., 2023a] would enable further spatial guidance, while including IP-Adapter-like [Ye et al., 2023] approaches would allow global guidance from image modality. Exploring the applicability of our model for inpainting tasks is also worth investigating. Furthermore, audio carries temporal information. Thus, the inclusion of our approach with video generation models is a natural extension of our work. Our current model treats audio signals as a unified global information piece. Given the rapid advances in the video generation field, our model would be suitable for diffusion-based video generation models due to its lightweight nature.

BIBLIOGRAPHY

[uns,] Unsplash. `https://unsplash.com`. Accessed: 2023-11-15.

[Alayrac et al., 2020] Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., and Zisserman, A. (2020). Self-supervised multimodal versatile networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 25–37.

[Brock et al., 2018] Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*.

[Brooks et al., 2023] Brooks, T., Holynski, A., and Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

[Conneau et al., 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

[Elizalde et al., 2023] Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. (2023). CLAP: Learning audio concepts from natural language supervision.

In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[Gal et al., 2022] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion.

[Girdhar et al., 2023] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. (2023). ImageBind: One embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27.

[Guzhov et al., 2022] Guzhov, A., Raue, F., Hees, J., and Dengel, A. (2022). AudioCLIP: Extending clip to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

[Hertz et al., 2022] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-or, D. (2022). Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*.

[Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[Ho et al., 2020] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851.

[Houlsby et al., 2019] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, pages 2790–2799.

[Jeong et al., 2023] Jeong, Y., Ryoo, W., Lee, S., Seo, D., Byeon, W., Kim, S., and Kim, J. (2023). The power of sound (TPoS): Audio reactive video generation with stable diffusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7822–7832.

[Karras et al., 2020] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020). Training generative adversarial networks with limited data. In *Proc. NeurIPS*.

[Karras et al., 2019] Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410.

[Kawar et al., 2023] Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. (2023). Imagic: Text-based real image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017.

[Lee et al., 2022a] Lee, S. H., Oh, G., Byeon, W., Kim, C., Ryoo, W. J., Yoon, S. H., Cho, H., Bae, J., Kim, J., and Kim, S. (2022a). Sound-guided semantic video generation. In *European Conference on Computer Vision (ECCV)*, pages 34–50. Springer.

[Lee et al., 2023a] Lee, S. H., Oh, G., Byeon, W., Yoon, S. H., Kim, J., and Kim, S. (2023a). Robust sound-guided image manipulation. *arXiv preprint arXiv:2208.14114*.

[Lee et al., 2022b] Lee, S. H., Roh, W., Byeon, W., Yoon, S. H., Kim, C., Kim, J., and Kim, S. (2022b). Sound-guided semantic image manipulation. In *IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 3377–3386.

[Lee et al., 2023b] Lee, T., Kang, J., Kim, H., and Kim, T. (2023b). Generating realistic images from in-the-wild sounds. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7160–7170.

[Li et al., 2023a] Li, J., Li, D., Savarese, S., and Hoi, S. (2023a). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

[Li et al., 2022] Li, T., Liu, Y., Owens, A., and Zhao, H. (2022). Learning visual styles from audio-visual associations. In *European Conference on Computer Vision (ECCV)*, pages 235–252. Springer.

[Li et al., 2023b] Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. (2023b). Gligen: Open-set grounded text-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22511–22521.

[Livingstone and Russo, 2018] Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess) [data set]. *PLoS ONE*, 13(5):e0196391.

[Mei et al., 2021] Mei, X., Liu, X., Huang, Q., Plumbley, M. D., and Wang, W. (2021). Audio captioning transformer. In *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, pages 211–215, Barcelona, Spain.

[Meng et al., 2021] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. (2021). SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*.

[Mokady et al., 2023] Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. (2023). Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047.

[Mou et al., 2023] Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., and Qie, X. (2023). T2I-Adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

[Nichol et al., 2022] Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen, M. (2022). GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, pages 16784–16804. PMLR.

[Owens et al., 2016] Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., and Freeman, W. T. (2016). Visually indicated sounds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2413.

[Parmar et al., 2023] Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., and Zhu, J.-Y. (2023). Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11.

[Podell et al., 2023] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis.

[Qin et al., 2023] Qin, C., Yu, N., Xing, C., Zhang, S., Chen, Z., Ermon, S., Fu, Y., Xiong, C., and Xu, R. (2023). Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23085–23096.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning

transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763.

[Raffel et al., 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

[Ramesh et al., 2022] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

[Rombach et al., 2022] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241.

[Saharia et al., 2022] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 36479–36494.

[Sung-Bin et al., 2023] Sung-Bin, K., Senocak, A., Ha, H., Owens, A., and Oh, T.-H. (2023). Sound to visual scene generation by audio-to-visual latent alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6430–6440.

[Tang et al., 2023] Tang, Z., Yang, Z., Zhu, C., Zeng, M., and Bansal, M.

(2023). Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*.

[Tumanyan et al., 2023] Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. (2023). Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930.

[van den Oord et al., 2019] van den Oord, A., Li, Y., and Vinyals, O. (2019). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[Wu et al., 2022] Wu, H.-H., Seetharaman, P., Kumar, K., and Bello, J. P. (2022). Wav2CLIP: learning robust audio representations from CLIP. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE.

[Xu et al., 2022] Xu, X., Xie, Z., Wu, M., and Yu, K. (2022). The sjtu system for dcase2022 challenge task 6: Audio captioning with audio-text retrieval pre-training. Technical report, DCASE2022 Challenge.

[Yang et al., 2023] Yang, Y., Zhang, K., Ge, Y., Shao, W., Xue, Z., Qiao, Y., and Luo, P. (2023). Align, adapt and inject: Sound-guided unified image generation.

[Yariv et al., 2023a] Yariv, G., Gat, I., Benaim, S., Wolf, L., Schwartz, I., and Adi, Y. (2023a). Diverse and aligned audio-to-video generation via text-to-video model adaptation.

[Yariv et al., 2023b] Yariv, G., Gat, I., Wolf, L., Adi, Y., and Schwartz, I. (2023b). Adaptation of Text-Conditioned Diffusion Models for Audio-to-Image Generation. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 5446–5450.

[Ye et al., 2023] Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W. (2023). IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*.

[Zhang et al., 2023a] Zhang, L., Rao, A., and Agrawala, M. (2023a). Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847.

[Zhang et al., 2023b] Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., and Qiao, Y. (2023b). LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

[Zhao et al., 2023] Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., and Wong, K.-Y. K. (2023). Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.