

---

# Synthesizing Visuals from Sound: A Framework for Sound-Driven Image Generation

---

**Shaohong Chen\***  
University of Toronto  
jameschen@cs.toronto.edu

**Eric Zheng\***  
University of Toronto  
eric.zheng@cs.toronto.edu

**Jinyang Zhao\***  
University of Toronto  
jerryzhao@cs.toronto.edu

## Abstract

Conditional image generation has emerged as a prominent area of research, with recent advancements in text-to-image models achieving remarkable success. Inspired by this progress, the design team explores the novel domain of sound-to-image generation, leveraging auditory inputs to synthesize expressive visual outputs. This research investigates the intersection of auditory and visual modalities by translating acoustic features—such as frequency, amplitude, and timbre—into intricate satellite imagery. By focusing on the generation of satellite scenes from background sounds, the study bridges sensory modalities, transforming ephemeral soundscapes into tangible visual representations. Building upon the demonstrated efficacy of transformer and diffusion models in text-to-image generation, the design team examines the potential of these architectures for sound-to-image synthesis. This work contributes to the field of multi-modal generation by blending auditory perception with visual creativity, paving the way for new applications in art, science, and sensory exploration.

## 1 Introduction

Recent advancements in conditional image generation using machine learning architectures such as generative transformers (GPT), diffusion models, and Variational Autoencoders (VAEs) have demonstrated remarkable success. Most of these efforts focus on generating images from textual or visual information, achieving significant breakthroughs across these domains. However, human perception is inherently multimodal; individuals naturally process their surroundings through both auditory and visual senses. For instance, when hearing specific sounds, people instinctively form mental images of corresponding objects or scenes. Despite this, bridging auditory and visual modalities remains underexplored, making it a compelling area for further investigation. Motivated by this gap, the design team aims to explore the intersection of auditory and visual domains, with the goal of developing a machine-learning model that directly generates visual content from auditory input.

Current state-of-the-art methods in audio-to-image generation primarily utilize VAEs and Generative Adversarial Networks (GANs)[1, 2, 3]. While these approaches have demonstrated some success, they come with notable limitations. VAEs, for example, often produce blurry outputs due to the trade-off between reconstruction accuracy and latent space regularization. They also struggle with high-dimensional data, such as complex imagery, and are susceptible to issues like posterior collapse, which can degrade the effectiveness of the latent space. GANs, on the other hand, can generate sharper images but frequently encounter challenges like mode collapse and training instability. These limitations make both approaches difficult to train and less effective for capturing the full complexity of multi-modal data, such as the auditory and visual information humans naturally integrate.

An alternative approach involves converting audio into intermediate representations, such as words or classes, which are then used to generate images[4]. However, this multi-step process diverges from how humans intuitively process sensory information. For example, when hearing the sound of ocean waves, people can directly imagine a seaside scene without the need for explicit classification steps. Inspired by this human capability, the design team desires to simulate this natural process by directly generating images from sound input without relying on intermediate classification.

---

\*Equal Contribution

Given the remarkable success of transformers and diffusion models in applications such as text generation and text-to-image synthesis, this study explores the potential of adapting these architectures for the task of audio-to-image generation. Specifically, the design team focuses on generating visual scenes from background sounds, leveraging these advanced models to push the boundaries of multi-modal learning. Due to the time constraints, the project’s scope is generating relevant scene imagery directly from auditory input.

## 2 Related works

### 2.1 Foundational works for the project

ImageBind introduces an innovative approach to generating embeddings for multiple modalities using a single model[5]. The presenting project leverages the ImageBind model extensively to generate embeddings for audio and image data, which are then used to fine-tune a stable diffusion model. By incorporating ImageBind, the performance of the diffusion model improved significantly, producing highly relevant and accurate images.

Image Transformer introduces the concept of leveraging transformer architectures for image generation[6]. The paper demonstrates success in various image-to-image generation tasks, such as image super-resolution, image completion, and category-conditioned image generation. Drawing inspiration from this work, the design team adapted the transformer architecture to explore audio-to-image generation, extending its application beyond traditional image-to-image tasks.

Anything2Image is an open-source project that inspired the design team to fine-tune a stable diffusion model using embeddings from ImageBind[7]. Anything2Image demonstrates the ability to generate images directly from various multimodal inputs, such as audio and text, without additional training. Intrigued by this approach, the design team adopted and extended this concept by fine-tuning a stable diffusion model to enhance its performance for their specific use case.

Stable Diffusion 2.1 UnCLIP builds upon the diffusion model paradigm, focusing on text-to-image generation by conditioning a diffusion model on textual embeddings from a pretrained CLIP text encoder[8]. The UnCLIP variant enhances performance by incorporating a trained decoder to reconstruct CLIP image embeddings, effectively bridging the gap between textual and visual modalities. The team is fine tuning this model for the audio-to-image synthesis.

### 2.2 Other works related to the project

SoundNet introduced a Convolutional Neural Network that can perform object and scene classification in videos only with their sound data, which showcased the possibility of object recognition with only audio data[9]. The presenting project also uses the intuition of recognizing scenes and objects from audio, but extends that and generates the images of the scenes using state-of-the-art architectures.

Audio to Image Cross-Modal Generation experimented on using shared audio and video feature space for image generation with combined VAE and GAN models[1]. It experimented on generating images of numbers from audio recordings of spoken digits and generated quality results. The presenting project performs a similar task but aiming to generate general quality images rather than only numbers, using a more tractable transformer instead of VAE or GAN.

Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment used ResNet encoders and GAN for an early experiment on the task of image generation from audio data. This model generates scenes from given audio samples[2]. It has the limitation of generating scenes with multiple pieces of sounds from different objects. The presenting project in this study differs from their work since the design team explored this idea using transformers and diffusion models, and this project aims to generate images of scenes when given complex sound samples.

S2I-Bird is another application of GAN on sound-to-image generation, specifically on bird images[3]. It successfully generates quality images of birds of corresponding species when taking in audio data of different bird calls. The presenting study differs from this since the team is using transformers and diffusion models and the design team would like to extend the experiment of sound-to-image generation to a more general group of objects and scenes.

### 3 Formal description of the system

#### 3.1 Problem Domain

The goal of this work is to generate images from audio inputs. Given a sound input (such as a short audio clip), the system is tasked with producing a corresponding image that visually represents the content of the sound. For example, a sound of waves crashing on the shore might produce an image of a beach scene.

The challenge lies in bridging the gap between two very different types of data:

- **Audio Data:** Represented as waveforms or frequency features, which are inherently temporal.
- **Image Data:** Represented as spatial pixel grids (grayscale or color images).

To address this challenge, we developed two systems:

1. A **transformer-based model** that is built and trained from scratch as shown in Figure 1.
2. A **fine-tuned Stable Diffusion 2.1 Unclip model** that leverages pre-trained components for high-quality image generation as shown in Figure 2.

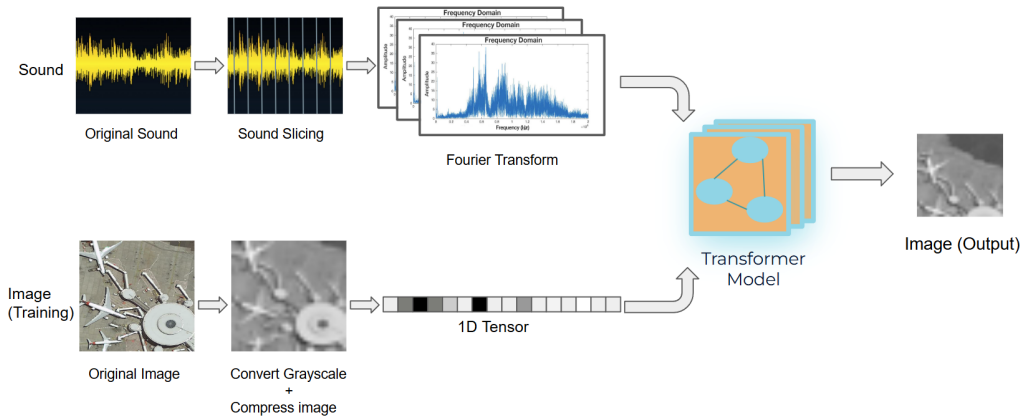


Figure 1: Transformer System Illustration

#### 3.2 Transformer Architecture Illustration

Figures 1 and Figure 4 illustrate the transformer system we built. The sound input is first divided into a specified number of sections, with a Fourier transform applied to each section. The resulting frequency-domain data are then passed to the transformer model, which generates the corresponding image. To reduce computational complexity, the image is converted to grayscale and compressed to a  $32 \times 32$  resolution. This  $32 \times 32$  image is flattened into a 1D tensor for processing by the transformer. Notably, image information is only available during the training phase, where the transformer learns to predict the image output based on the sound input.

We adapted the transformer encoder-decoder architecture introduced in the original *Attention Is All You Need* paper [10]. Unlike text-based tasks, our audio data do not have a fixed vocabulary, so instead of using a standard tokenizer, we applied a linear embedding layer to map the audio tensor to the input dimension required by the transformer model. For the image data, we implemented a tokenizer-like embedding map that treats each pixel value (ranging from 0 to 255) as a class, mapping it to an embedding index. This allows the transformer to process the image as a sequence of pixel classes, facilitating the generation of the image from the audio input.

In order to capture the sequential structure of the audio data and the spatial relationships in the image, we employed sinusoidal positional encoding for both inputs[10, 11]. For the audio, this encoding provides the transformer with relative and absolute positional information, enabling it to model temporal dependencies. For the image, we first reformatted the 2D structure into a 1D array and applied positional encoding to the flattened sequence, ensuring the model retains spatial information about pixel positions during training and generation.

During training, the model’s objective is to predict the next pixel value in the image sequence, framing the task as a classification problem with 256 possible classes corresponding to the range of pixel values. Cross Entropy Loss is used to optimize the model, guiding it to predict the next pixel in the sequence. For the image generation step, the decoder output is passed through a linear layer followed by a softmax function to predict the next pixel value. This process is repeated iteratively for 1024 steps, at which point the output is reshaped into a  $32 \times 32$  grayscale image.

The model architecture incorporates several key features that enhance its performance. Local 1D Self-Attention [11] is used within the multi-head attention mechanism to reduce computational complexity while effectively processing the  $32 \times 32$  image input. The encoder-decoder structure remains largely unchanged from the original design, ensuring that the transformer can effectively learn the mapping between the audio input and the generated image output.

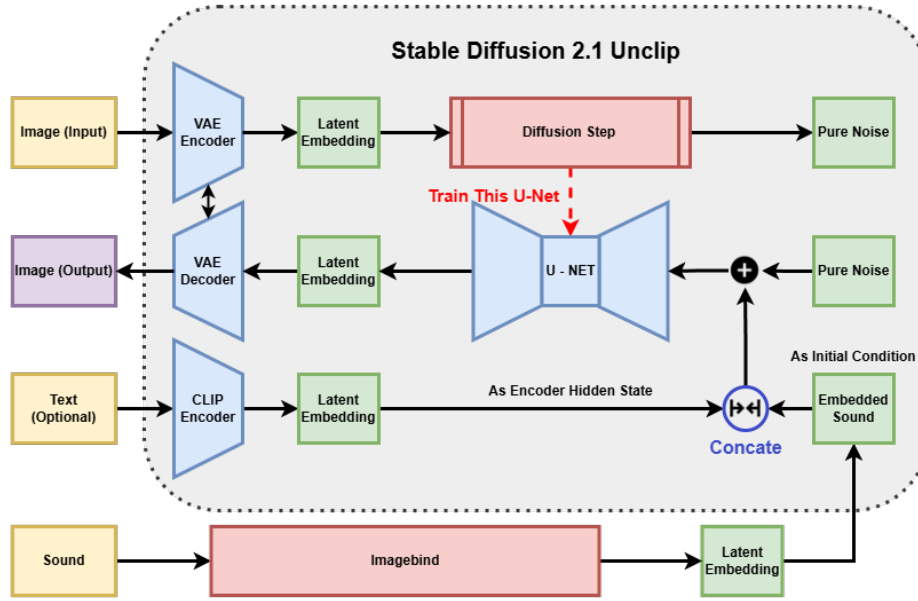


Figure 2: SounDiff System Illustration

### 3.3 Stable Diffusion 2.1 Unclip Fine-tuning System (SounDiff)

Figure 2 illustrates the architecture for fine-tuning the stable diffusion model. The audio inputs are processed by the pre-trained ImageBind model[5], which transforms them into latent space embeddings. These embeddings are used as image embedding inputs to fine-tune the Stable Diffusion 2.1 UnCLIP model[8], allowing it to generate an output image corresponding to the input audio.

The ImageBind model[5] is a multimodal representation system capable of embedding inputs from various modalities, such as images, text, and audio, into a shared latent space. This shared latent space allows embeddings from one modality, such as sound, to act as proxies for embeddings of another modality, such as images. **In our setup, the sound embeddings generated by ImageBind are treated as equivalent to image embeddings and passed directly to the Stable Diffusion 2.1 Unclip model as its conditioning input.**

Our fine-tuning focuses exclusively on the U-Net[12] component of the Stable Diffusion model. The U-Net is responsible for de-noising during the diffusion process, gradually transforming noisy latent representations back into coherent images. During pre-training, the model learns to predict and remove noise added to an input image. **We fine-tune the U-Net following a similar process: the input to the U-Net is an image with added Gaussian noise, and the U-Net is trained to predict the noise itself. The ground truth noise serves as the target during training, and the Mean Squared Error (MSE) loss function is used to minimize the difference between the predicted noise and the ground-truth noise.**

During training for our SoundDiff-S variant, prompts were provided to indicate that the images correspond to satellite maps. In contrast, our SoundDiff-L variant is trained unconditionally, generating images directly from audio embeddings.

The workflow for fine-tuning can be summarized as follows:

- The sound input is embedded into the shared latent space via the ImageBind model.
- This latent embedding, treated as an image embedding, serves as a conditioning input to the Stable Diffusion model.
- Gaussian noise is added to an input image, and the noisy image is passed to the U-Net.
- The U-Net predicts the noise added to the image.
- The loss function **calculates the MSE between the predicted noise and the ground-truth noise**.
- Fine-tuning adjusts the U-Net parameters to minimize this loss, enabling the model to generate coherent images conditioned on the sound embedding.

By leveraging the multimodal nature of ImageBind and the flexibility of Stable Diffusion 2.1 Unclip, our system effectively bridges audio inputs with visual outputs, fine-tuning the U-Net to achieve high-quality sound-driven image synthesis.

## 4 Experiment

We evaluated the transformer and stable diffusion models for the audio-to-visual generation task during preliminary research and experimented with both at different scales. We first provide the common setups in the supplement and discuss the details in subsections.

**Dataset Preprocessing** We used the Audiovisual Aerial Scene Recognition Dataset [13] that contains 5075 satellite images-audio pairs for both models. Each image-audio data pair contains a 512 by 512 resolution satellite image for a labeled scene with a corresponding 10 second audio recording from the same scene. Different preprocessing strategies were applied to the models to fit with resource and time constraints. We resized each image to 32 by 32 pixels grayscale images for the transformer approach to fit each batch into the VRAM of a single GPU. On the other hand, we resized each image to a resolution of 768 by 768 and kept it in RGB color to fit into the stable-diffusion-2.1-unclip pipeline [8]. We performed Fourier transforms to slice each piece of audio into 500 220-d vectors for the transformer experiment, and did not make any changes to audio for our proposed diffusion model approach. For both approaches, we split the dataset into 90% for training and 10% for validation.

### 4.1 Audio to Image Transformer

**Training Setup** We experimented with pre-training a 512-d image transformer with 1-D attention mechanism and with a similar dimension as the original Image Transformer [6] with 6 encoder layers and 6 decoder layers, with 8 attention heads for both encoder and decoder. We trained 300 epochs in 3 GPU hours on a single RTX 4090 GPU with Cross Entropy Loss, Adam optimizer and a scheduled learning rate linearly warmed up to  $10^{-7}$  in 300 batch steps and decays proportionally afterwards for our final model, before we begin our evaluation on the model.

**Evaluation** We experimented with transformer models in different sizes, and trained with subsets of data in different sizes to fully investigate this approach. In the early stage, we performed preliminary research on this approach by pre-training a transformer model that overfits the subset of 20 similar airport image-audio pairs. The transformer model was able to converge in 5000 batch-steps and can generate output images similar to the 20-image subset when being fed with similar audio inputs. However, when we extend our model to the larger dataset, we noticed that the model was not able to converge as effectively as using the subset after the first 50 epochs, and the loss value eventually fluctuates between 4.2 to 5.5 after a certain epoch. With this model, the output images did not show significant patterns. The images discussed in this section are attached in B. Due to time and resource constraints, we decided to move on to the other approach to solve the stated problem. We will discuss the detailed findings in the limitations section.

### 4.2 SoundDiff: Imagebind Embedding to Stable Diffusion UnCLIP

**Training Setup** We fine-tuned the U-Net component of the Stable Diffusion UnCLIP 2.1 (SDU)[8] model in our proposed structure with a scheduled learning rate from  $10^{-6}$  to  $10^{-7}$ . AdamW [14] optimizer was used with momentum betas (0.9,





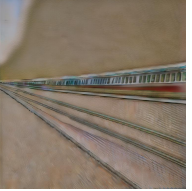

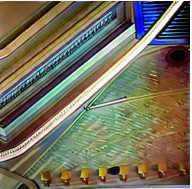
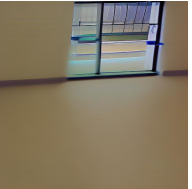
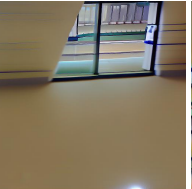
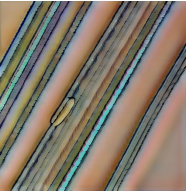
0.999), and with a weight decay rate of 0.01. We used the original noise scheduler from the original stable diffusion model. Two models were trained in the experiment stage, our Soundiff-L model was trained on the entire training dataset without conditioning for 80 epochs and with a batch size of 4, while we also trained a model Soundiff-S on the train station subset of data with minimal prompting to slightly condition the model to generate "satellite image" output in the training stage for 280 epochs and with a batch size of 2. The training losses and learning rates for the main model are shown in D. 1 GPU day was used for training Soundiff-L while 9 GPU hours were used to train Soundiff-S. We fixed the random seeds 42 for reproducibility.

**Conditional and Unconditional SoundDiff** We qualitatively evaluated the performance of both Soundiff-L and Soundiff-S by generating images from sound wav files from the validation set with and without conditioning to direct the model to generate satellite images to match our training dataset. We also generated images from the original ImageBind + SDU model with and without conditioning for satellite images to compare the performance between our proposed model and the baseline solution. Samples of these generations are listed in Table 1.

For highly distinguishable audio from the train station, all generation of images have train elements in them. All of our model generates more structured lines for the trains while the baseline model fails to align the train carriages for both conditional and unconditional generations. The conditioned generation from our Soundiff-S model is also able to draw the train tracks clearly instead of abstract curved lines as in the conditional generation of the baseline model. For another less distinguishable audio from a train station, our Soundiff-L and Soundiff-S models were able to generate the images closest to a train station with a train platform-like structure outside the doors when all generations from the baselines and the unconditional generations are from indoors. In this specific example, the baseline model failed to generate any elements related to the train station but generated an image of a door facing a hallway.

When being conditioned to generate satellite images, our model was able to generate train-track-style parallel lines that are more likely to appear in satellite images in comparison to the curved lines generated by the baseline model with the same condition. Notably, more straight lines and squared structures are used in Soundiff-L generated images than those generated by Soundiff-S given the same random seeds and the same input audios. In the same time, Soundiff-L generated images are slightly more blurry than the images generated by Soundiff-S model.

More generated images from Soundiff models are included in Appendix C.

Audio Description	BaseLine	Baseline + Condition	SounDiff-L	SounDiff-S	SounDiff-S + Condition
Train Rumbling					
Train Station					

**Table 1:** Images generated by our model compared with the Baseline Model.

### 4.3 Analysis and Applications

We evaluated the performance of our pre-trained image transformer and our Soundiff model fine-tuned from ImageBind and Stable Diffusion 2.1 UnCLIP. Since only our fine-tuned stable diffusion model can generate promising images with reasonable context using affordable training time, we will focus on the performance of our Soundiff models.



Other than the observations described in the previous section, we also noticed the images after our fine-tuning process have more details on the main elements in each image, and have much fewer details on other places, than our baseline model. Specifically, in the train rumbling sound example, although all our generations have similar layouts, SounDiff model generated images have more abstract backgrounds but depict the main element of our input audio (the train) in detail, suggesting the audio embeddings in our SounDiff pipeline has directed the Stable Diffusion model to focus more on the main elements in the audio.

Our generated images also use straight lines more frequently than the baseline model. This may be a result from our training dataset of satellite images, as streets and buildings are often shown as squares and straight lines in them. The SDU model may have used these features of satellite images as artistic styles in its image generation process.

Overall, it is evident that our fine-tuned model has learned features from the image-audio training dataset we provided during the training process. With this observation, we can reasonably suggest that if we have larger datasets with more variant audio-image pairs, it is plausible to expect the model to converge better and generate more realistic images describing the input audio data. Although our baseline SDU pipeline can already generate images directly by piping in most of the audio embeddings, some fine-tuning is still needed to align these embeddings to realistic images.

This research is useful as an inspiration for further developments in multi-modal image generations, as it explores the possibility of aligning latent embeddings of two modalities from two different models and use them for image generation by simple fine-tuning processes.

## **5 Limitations**

### **5.1 Transformer Approach Limitations**

The current Transformer model successfully overfits small datasets, producing good images. However, it suffers from a severe under-fitting problem when the training dataset exceeds 20 images, with outputting images close to pure noise.

#### **5.1.1 Limitations Reasoning**

The design team suggests that the Transformer requires significantly more parameters to learn the features of input audio. However, the current model is already at the maximum size that can be trained on an A6000 GPU with 48GB of VRAM, the largest available to the team. Additionally, training such a Transformer from scratch typically requires millions of samples, far exceeding the few thousand samples the team currently possesses. Given the team’s disk space limitations on CSLab machines, it is impractical to train the model to achieve optimal results.

#### **5.1.2 Potential Methods to Address Limitations**

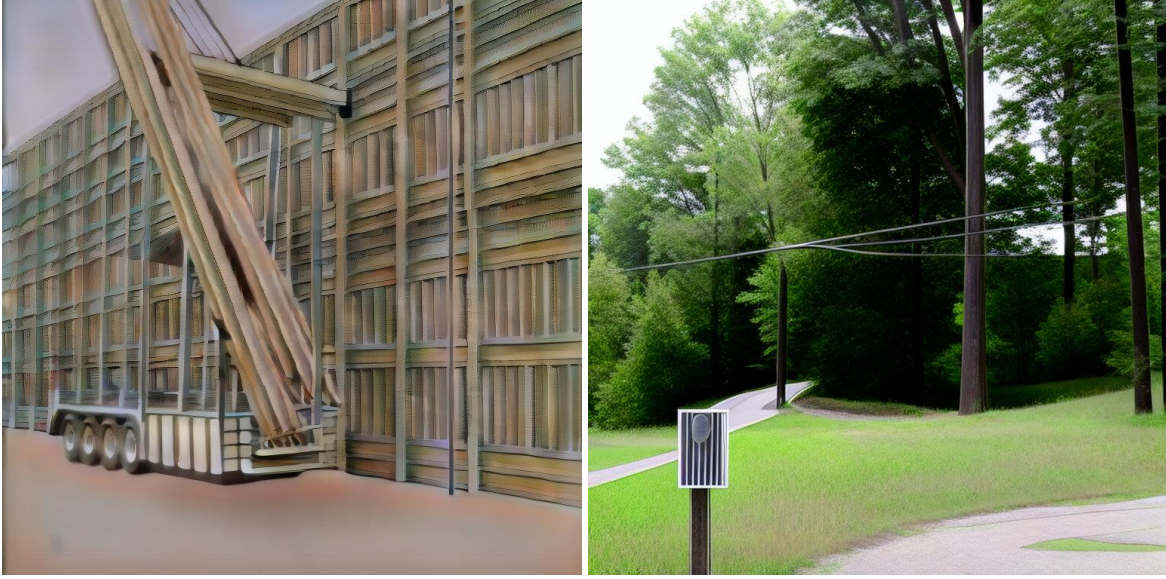
To address the aforementioned issues in the future, one potential approach is to train a much larger model using a network of GPUs. In addition, employing a larger dataset is critical. The design team has identified a suitable SoundNet data set, which contains more than 300GB of data; however, it is currently infeasible for the team to use due to resource constraints [9]. Another promising strategy is fine-tuning a pre-trained Transformer, as this approach is more efficient in terms of data and computational requirements. Unfortunately, there is no suitable pre-trained Transformer currently available for the task of image generation.

### **5.2 Diffusion Model Fine tuning Limitations**

The fine-tuned diffusion model generally produces high-quality and relevant images from the sound input. However, it occasionally generates images that are inconsistent with the input audio. For instance, Figure 3 illustrates two failure cases using train station sounds: one audio results in a workplace image, while another generates a forest scene.

#### **5.2.1 Limitations Reasoning**

For the failure cases mentioned above, the design team revisited the original sound input files and identified several issues with the audio quality. For example, the image in Figure 3a was generated from an extremely noisy audio file. Even human listeners would struggle to discern the background context from this audio. In this case, the output image is somewhat reasonable, as the input audio resembles the ambient sounds of a workplace.



(a) Fail case: Workplace generated from train station sound      (b) Fail case: Forest generated from train station sound

**Figure 3:** Fail cases generated from train station sounds.

Another potential cause of failure is the inherent bias introduced by ImageBind during the embedding process. For example, in Figure 3b, the input sound clearly contains a train siren. However, the output image is incorrect, possibly due to the bias in ImageBind’s audio embeddings, which may influence the model to generate irrelevant or inaccurate images.

### 5.2.2 Potential Methods to Address Limitations

Given the presence of poor-quality examples in the dataset, performing a thorough data examination and evaluation is crucial. Gathering additional high-quality data and filtering out low-quality examples can significantly improve the dataset, making it more suitable for fine-tuning.

Regarding the potential inherent bias in ImageBind, fine-tuning the model to eliminate the bias is a viable solution. However, due to the extremely large size of ImageBind, fine-tuning it is infeasible for this project. Alternative approaches include implementing a post-processing step to verify whether the generated image aligns with the expected context or experimenting with other models for generating embeddings.

## 6 Conclusions

This study explores the novel task of sound-to-image generation, leveraging transformer architectures and fine-tuned diffusion models to bridge auditory and visual modalities. While the transformer approach demonstrated potential in over fitting small datasets, it faced significant challenges with scalability and convergence on larger datasets. In contrast, the fine-tuned diffusion model, based on Stable Diffusion 2.1 UnCLIP and ImageBind embeddings, effectively generated high-quality, contextually relevant images from sound inputs. The results highlight the importance of leveraging advanced architectures and fine-tuning processes to align multi-modal embeddings for practical applications.

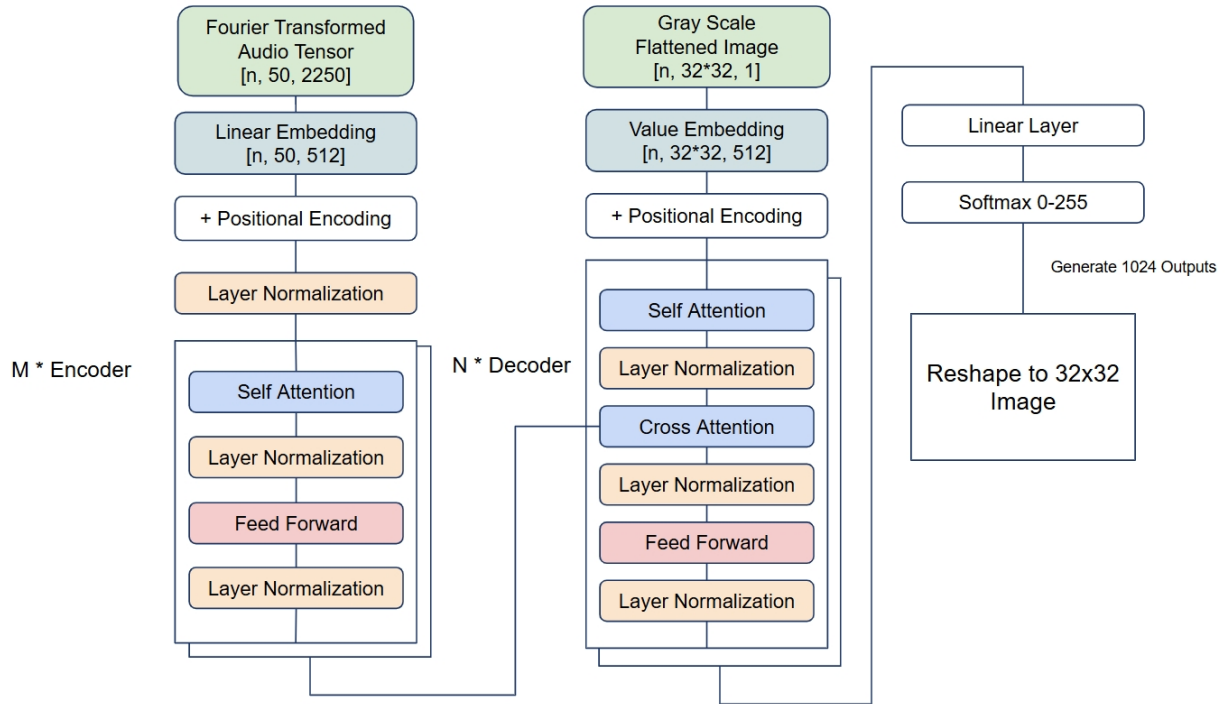
Despite its success, the model exhibits limitations, such as occasional failures to generate images consistent with the audio input due to dataset quality and potential biases in embedding spaces. These limitations underscore the need for higher-quality datasets, improved embedding alignment, and post-processing methods to enhance robustness. This research contributes to the growing field of multi-modal learning by demonstrating the feasibility of directly synthesizing visuals from sound, paving the way for future exploration in artistic, scientific, and sensory applications.



## References

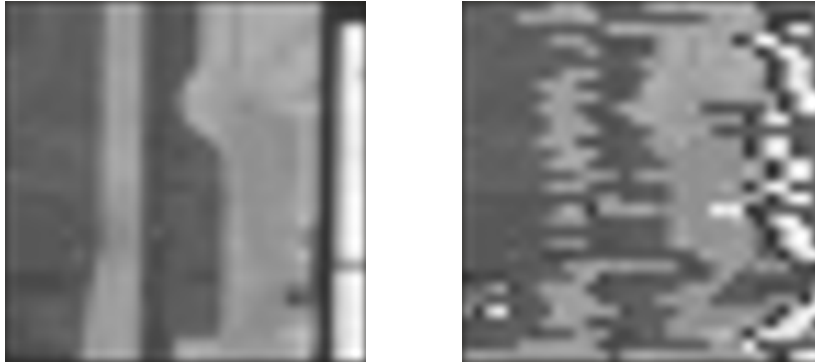
- [1] Maciej Zelaszczyk and Jacek Mandziuk. Audio-to-image cross-modal generation. *CoRR*, abs/2109.13354, 2021.
- [2] Joo Yong Shim, Joongheon Kim, and Jong-Kook Kim. S2i-bird: Sound-to-image generation of bird species using generative adversarial networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2226–2232, 2021.
- [3] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2023.
- [4] Anonymous. Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation. *arXiv preprint arXiv:2305.13050*, 2023.
- [5] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, Antoine Miech, Mathilde Caron, and Piotr Bojanowski. Imagebind: One embedding space to bind them all. *arXiv preprint arXiv:2305.05665*, 2023.
- [6] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, and Aidan N Gomez Ku. Image transformer. In *International Conference on Machine Learning (ICML)*, pages 4055–4064. PMLR, 2018.
- [7] Chris Chan. Anything2image: Generate images from text, audio, and other modalities using imagebind and stable diffusion. <https://github.com/chrischan0204/Anything2Image>, 2023. GitHub repository.
- [8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [9] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.
- [10] Noam Shazeer Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, 2023.
- [11] Ashish Vaswani Niki Parmar, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv*, 2018.
- [12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.
- [13] Di Hu, Xuhong Li, Lichao Mou, Pu Jin, Dong Chen, Liping Jing, Xiaoxiang Zhu, and Dejing Dou. Cross-task transfer for multimodal aerial scene recognition. *CoRR*, abs/2005.08449, 2020.
- [14] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

## A Image Transformer Architecture



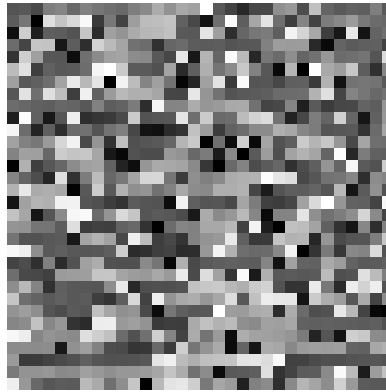
**Figure 4:** Our Transformer Model

## B Image Transformer Outputs



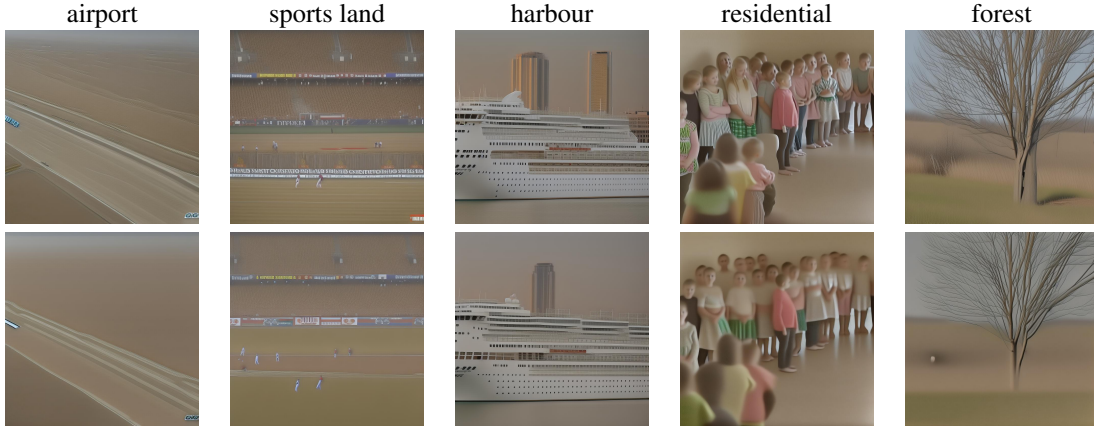
(a) Original image for an audio from airport      (b) Generated image from the same audio

**Figure 5:** Example generated image using image transformer overfitted to a subset of similar image-audio data pairs.



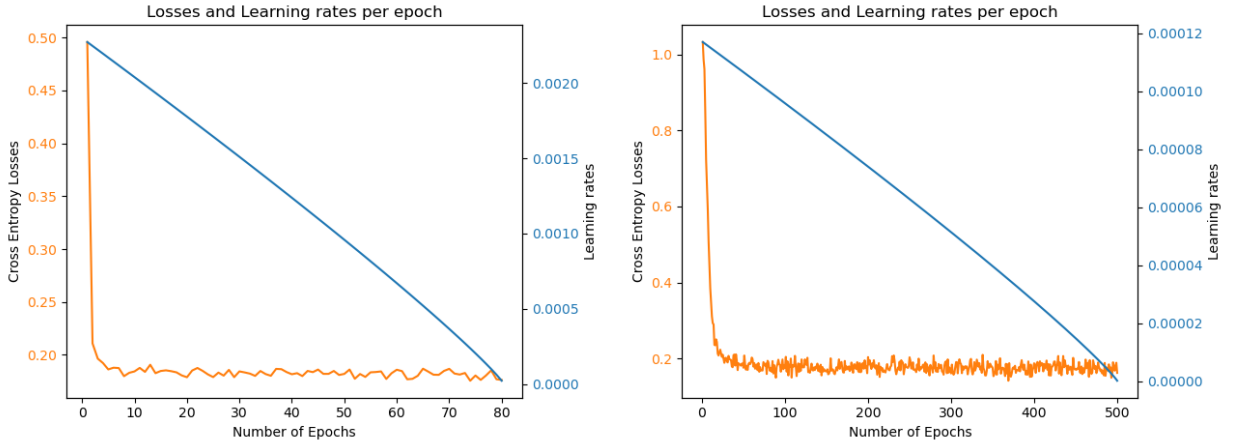
**Figure 6:** Output of Image transformer trained with full dataset after 1500 epochs

## C More SoundDiff Image Generated Examples



**Table 2: More Images Sampled By SoundDiff Models.** The image in the first row are generated using SoundDiff-S model, and images in the second row are generated using SoundDiff-L model. The category of the input audio is recorded on the top. Most images generated by SoundDiff-L model have more physically accurate main elements, including more accurate human faces and object structures. These images may also have less details in other places than the images generated by SoundDiff-S models.

## D Training Losses and Learning Rates for Soundiff Models



(a) Training Losses and Learning Rates for **SoundDiff-S** model    (b) Training Losses and Learning Rates for **SoundDiff-L** model

**Figure 7: Training Losses and Learning Rate changes during SoundDiff Model trainings**

## E Project Code

The source code for this project can be found in this github repo:

<https://github.com/JamesChenSH/Audio2Image>