
Images that Speak: Generating meaningful images that are speech spectrograms

Hemanth Ram

The University of Texas at Austin

Shruti Sriram

The University of Texas at Austin

Krishanu Saini

The University of Texas at Austin

Abstract

Our work is fundamentally inspired by the principles of multi-modal generative modeling, specifically the combined denoising technique using diffusion models from different modalities, as explored in Images that Sound[1]. This established the possibility of synthesizing a single representation that functions effectively in both the visual and auditory domains. We extend this concept from general audio and images to the specific and structurally constrained domain of human speech spectrograms and images. Our objective is to develop a generative framework at the intersection of vision and speech that produces semantically meaningful images which, when interpreted as spectrograms, yield intelligible human speech.

1 Introduction

Generative modeling has achieved remarkable success in both the visual and auditory domains. Diffusion models like Stable Diffusion can produce high-fidelity images from text descriptions, and analogous latent diffusion approaches have been extended to audio generation. This convergence of vision and audio modalities in generative AI has opened new possibilities for multi-modal content creation. Recent advances show that models adapted from image generation can synthesize realistic sounds, including human speech, directly from text prompts. These developments blur the boundaries between modalities.

A natural bridge between images and audio is the spectrogram, a 2D visual representation of sound. However, spectrograms of speech or music typically appear as abstract patterns that are not recognizable as everyday images, and conversely, ordinary images sound like noise when interpreted as audio. Prior to the advent of modern generative models, attempts to fuse these modalities were largely limited to steganography and art experiments. The challenge is that visual and auditory coherence are usually at odds: optimizing an image to look clear can distort the sound, while ensuring audible clarity can degrade the image. Overcoming this tension requires a method that jointly models both modalities.

Images that Sound[1] provided the first proof-of-concept that such dual-purpose media can be synthesized. By leveraging pre-trained text-to-image and text-to-spectrogram diffusion models within a shared latent diffusion space, their approach generates spectrogram images that look like natural images while sounding like natural audio. In their method, a standard image diffusion model and an audio diffusion model are run in parallel on the same latent representation, iteratively denoising from random noise under the constraints of both models. The outcomes were spectrograms that could be “seen” as images and “heard” as sound, inaugurating a new type of multimodal art. Yet, Images that Sound primarily explored environmental sounds and musical audio; the generated audio, while natural-sounding, was not centered on spoken language or intelligible words.

In this work, we push this vision–audio convergence further into the realm of speech. Our goal is to create images that are not only visually meaningful but also encode intelligible speech in their spectrogram structure. Human speech imposes strict requirements on the spectrogram. The fine temporal structure and frequency patterns must align to phonetic and linguistic content for the speech to be understood. Generating a spectrogram that simultaneously yields a recognizable picture and a clear spoken message is a significantly more challenging task. We explore modifications to the audio diffusion model, Auffusion used in Images that Sound paper to accomplish this task.

2 Related Work

Stable Diffusion[2] and Auffusion[3] The architectural foundation of contemporary multimodal generation is the latent diffusion model (LDM). Stable Diffusion is an open-source deep learning model that utilizes a process called latent diffusion to condition the generation of images from textual descriptions. Building upon this, Auffusion is a text-to-audio (TTA) system that adapts the Stable Diffusion framework to generate text-conditioned sounds, human voices, and music. Critically, Auffusion internally generates mel-spectrograms, which are essentially 2D images. Because these spectrograms are treated as images within the pipeline, Auffusion maintains a high degree of internal similarity and architectural compatibility with Stable Diffusion, facilitating their integration into a single generative process.

Images That Sound A primary inspiration for this work is the Images that Sound framework, which demonstrates the feasibility of composing images and sounds on a single 2D canvas. This method utilizes a combined denoising technique where a single latent representation Z_t is processed simultaneously by a latent image diffusion model and a latent audio diffusion model. By applying a time-dependent weighted average to the noise estimates from both models, the framework generates a unified output that satisfies both visual and auditory constraints

Text-to-Speech (TTS) Diffusion Models and Compatibility While various diffusion-based TTS models like DiTTo-TTS[4] and E3 TTS[5] exist, many are not directly compatible with the combined denoising pipeline required for this project. Standard TTS diffusion models often utilize specialized architectures or latent spaces that differ significantly from the Stable Diffusion framework. This lack of architectural alignment prevents the elegant fusion of denoising steps in a common latent space, which is essential for ensuring that the generated output functions simultaneously as a coherent image and a legible speech spectrogram.

HiFi-GAN[6] (Vocoders) The transition from a generated spectrogram back to an audible waveform is handled by vocoders. HiFi-GAN is a pretrained vocoder frequently used to synthesize audio from mel-spectrograms. However, pretrained vocoders generally expect specific spectrogram distributions and settings to produce high-quality results. In the context of images that "speak", a vocoder must be precisely aligned to the synthesized spectrogram distribution to ensure that the reconstructed speech is clean and intelligible.

Existing Gaps Current TTA models like Auffusion are typically optimized for environmental sounds, generating the sound of a tiger growling when prompted with "tiger", rather than the precise linguistic phonemes required for speech. Furthermore, initial attempts to use zero-shot pipelines involving contour extraction and ControlNet[7] have failed to preserve the fine-grained textual structure of speech within a generated image. There is a critical need for a model that can specifically generate speech spectrograms (e.g., the word "rain" rather than the sound of rain) while maintaining a shared latent space for seamless integration with visual diffusion.

3 Methodology

3.1 Data

Our training data consists of paired textual inputs and corresponding speech spectrograms created via a text-to-speech pipeline. We generated a custom speech spectrogram dataset by converting written sentences into spoken utterances using a Tacotron2[8]-based TTS system. Each input text is first transcribed into a phoneme sequence to capture pronunciation. We use 150 text inputs with 9 words each to capture all phonemes. The phoneme sequence is then synthesized into an audio waveform by first slowing down the video and padding with silence to match 10-second audio waveforms required

for Auffusion, which we convert into a log-scale mel-spectrogram using Librosa. The resulting mel-spectrograms are treated as images of size 256x1024 for training. The data generation pipeline is shown in Figure-1

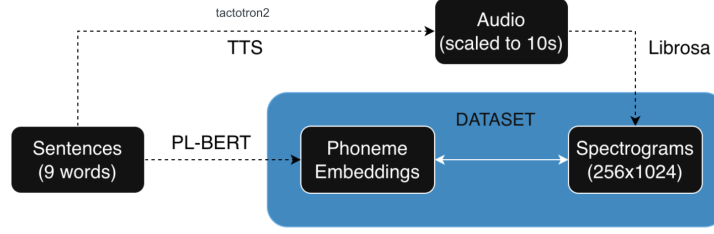


Figure 1: Data Generation Pipeline

For the visual modality, we leverage the existing pre-trained Stable Diffusion model, rather than training on a specific image dataset.

3.2 Model Architecture

Our system comprises two diffusion-based generative models: one for images and one for speech audio. They are integrated via a shared latent representation. In our design, the vision branch is a standard Stable Diffusion v1.5 model (latent diffusion for text-conditioned image synthesis), and the audio branch is a variant of Stable Diffusion fine-tuned to generate mel-spectrograms from text prompts (based on Auffusion). The final latent produced by the diffusion process can thus be decoded as a visual image or interpreted as a speech spectrogram interchangeably.

To better handle the granularity of speech, we introduce a phoneme embedding adapter in the audio branch. Instead of feeding raw text to the audio diffusion model’s text encoder, we first convert each input sentence into a sequence of phonemes that encodes pronunciation. These phoneme tokens are passed through PL-BERT[9] to produce a sequence of conditioning vectors, which then replace or augment the text encoder outputs for the diffusion U-Net. This design ensures that the audio model receives fine-grained phonetic information, which is crucial for generating intelligible speech spectrograms.

Both the image and audio diffusion models use a latent diffusion U-Net architecture with a shared latent space. The latent is a 2D tensor (spatial size 64x64 in our implementation) that represents either an image or a spectrogram in a compressed form. The Stable Diffusion image decoder (a variational autoencoder) transforms this latent into a high-resolution image, while the audio branch incorporates an analogous spectrogram decoder that inverts the latent into a full mel-spectrogram.

To convert the generated mel-spectrogram into an audible waveform, we integrate a neural vocoder. We choose HiFi-GAN as our vocoder due to its state-of-the-art performance in speech synthesis. HiFi-GAN is a generative adversarial network that produces high-fidelity audio from mel-spectrogram inputs by learning an inverse mapping to waveform space. We train a custom HiFi-GAN on our dataset’s spectrograms so that it aligns with the exact spectrogram format and distribution output by our diffusion model. The pre-trained HiFi-GAN has different settings like sampling rate, window/hop size, and mel resolution. This mismatch led to noisy generated outputs. By using a tailored HiFi-GAN, we tried to align the vocoder to our spectrogram distribution, producing cleaner and more intelligible speech.

3.3 Training Procedure

We start from the pre-trained Auffusion model (a text-to-audio latent diffusion model) and fine-tune it on our speech dataset to specialize it for text-to-speech spectrogram generation. The fine-tuning updates the model so that it can produce clear, human-speech mel-spectrograms consistent with input text prompts. To preserve the knowledge in the large pre-trained model while adapting to our specific task, we employ Low-Rank Adaptation (LoRA) for training. In addition to LoRA-based updates, we unfreeze a subset of the original U-Net parameters. This gave better convergence in practice, as some base model weights were adjusted to accommodate the new speech domain.

During fine-tuning, the audio diffusion model is conditioned on the phoneme embeddings for each training example’s text prompt. We use the standard denoising diffusion loss: at each training iteration, a clean spectrogram (the ground-truth mel from our dataset) is corrupted with a random noise level, and the U-Net is trained to predict the noise component. We use the same noise schedule and loss weighting as Stable Diffusion. The model thereby learns to map noised latent inputs to the clean spectrogram latent, given the guidance of the phoneme-text embedding. Training was performed for several 16K steps on our synthetic speech set, which was sufficient for the model to produce intelligible spectrograms matching the prompts. We did not fine-tune the image diffusion model (Stable Diffusion) on any new data.

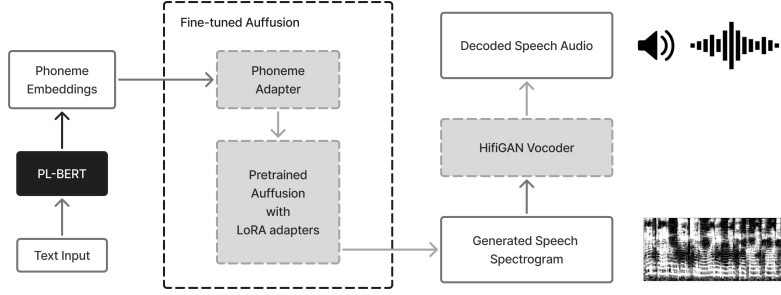


Figure 2: Auffuion and Vocoder fine-tuning pipeline

3.4 Vocoder Finetuning

While the diffusion model generates the visual structure of a spectrogram, a vocoder is required to reconstruct the final audio waveform. We observed that pretrained HiFi-GAN models expect specific spectrogram settings and distributions that may not align with our synthesized outputs. To resolve this, we performed additional training of the HiFi-GAN vocoder directly on our synthetic speech data. This fine-tuning process aligns the vocoder to our specific spectrogram distribution, which is critical for ensuring that the specific speech characteristics learned by the diffusion model, like formants and fricatives are accurately translated into audio.

3.5 Inference

In this pipeline, the pretrained Auffuion model is replaced by our fine-tuned Auffuion model, which has been optimized to generate human speech spectrograms rather than general sounds. The process begins with a random noise latent in a shared space compatible with both the frozen Stable Diffusion image model and our speech-adapted Auffuion. At each step of the iterative process, the latent is processed simultaneously by both models to produce modality-specific noise estimates based on the image and speech prompts. These estimates are merged using time-dependent weights, allowing the model to satisfy visual semantic constraints while maintaining the structural requirements of a speech spectrogram. Once the denoising process is complete, the final latent is decoded into a 2D matrix that functions as both a semantically meaningful image and a log-mel spectrogram. Finally, this spectrogram is converted into an audio waveform using our custom-trained HiFi-GAN vocoder, which is specifically aligned to our speech distribution to ensure the resulting output is intelligible.

4 Experiments

4.1 Setup

To evaluate the performance of our framework, we constructed a test suite consisting of 20 unique image prompts and 20 unique audio prompts. Each audio prompt consists of a sequence of nine words. By pairing every image prompt with every audio prompt, we generated a total of 400 distinct combinations to test the model’s ability to generalize across diverse visual and linguistic inputs.

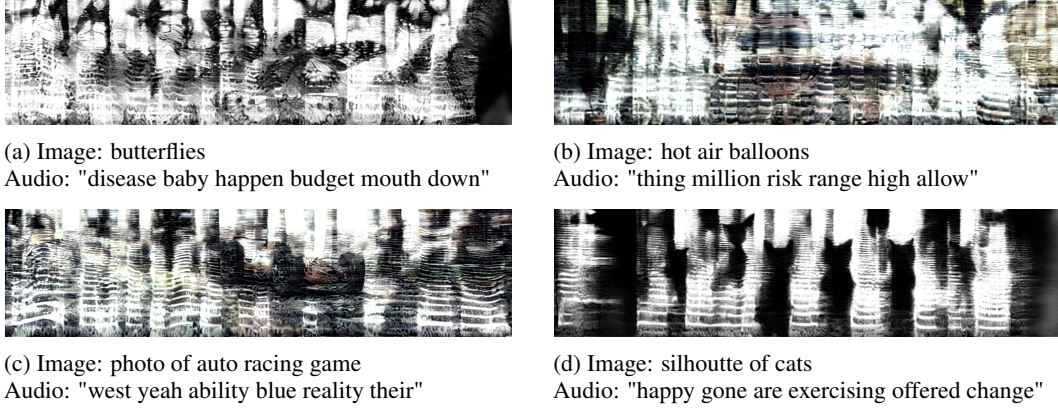


Figure 3: Example images generated by the combined denoising using Stable Diffusion and the fine-tuned Auffusion model. The prompts used for the image diffusion and the audio diffusion models are also shown in the captions. The decoded audio files can be found in the repository.

4.2 Evaluation Metrics

Our model has two objectives: to generate mel spectrograms that align with the image prompt and the decoded audio should speak the target speech. We use two metrics to evaluate the output.

Image-side metric: CLIP[10] similarity between generated spectrogram-image and image prompt .

Audio-side metric: Word Error Rate (WER) using Whisper[11] (automatic speech recognition) on generated audio vs target text (measures intelligibility + correctness).

Results for the metrics as an average across all 400 experimental samples are shown in Table-1.

Metric	Modality	
CLIP score	Vision	22.3%
WER	Speech	93.4%

Table 1: CLIP score and WER over the 400 test samples

4.3 Qualitative Results

We compare how aesthetic the images are and how intelligible the audio is to determine the quality of output generation. Figure-3 shows some example outputs generated by our pipeline.

5 Conclusion

Qualitatively, our results demonstrate that the generated outputs are indeed semantically aligned with the text prompts, and the synthesized images exhibit clear speech spectrogram like structure. When decoded, the resulting audio samples unmistakably resemble human speech. However, the intelligibility of the spoken content remains limited. This is reflected in the high WER values, where many words are unclear or distorted despite the speech-like rhythm and prosody being present.

On the image side, the CLIP similarity scores are reasonably strong, consistent with the qualitative observation that the visual outputs preserve semantic cues from the prompts. The speech-side shortcomings appear to stem from two factors: (1) the fine-tuning of the audio diffusion model, which must learn precise phonetic structure from a relatively small dataset, and (2) the performance of the vocoder, which requires close alignment to the distribution of generated spectrograms. Although our spectrogram generator is already producing patterns close to valid speech, benefiting from phoneme coverage in the curated dataset, the vocoder may require significantly more data to accurately reconstruct subtle phonetic detail necessary for intelligible words.

Overall, our results confirm that joint image–speech generation in a shared latent space is feasible and produces meaningful cross-modal outputs. Improving the fidelity of the audio branch, particularly through expanded vocoder training and larger-scale fine-tuning represents the most promising path toward fully intelligible, high-quality images that "speak".

6 Individual Contributions

Hemanth implemented the initial zero-shot generation pipeline using ControlNet and spectrogram conditioning. He developed the synthetic text–spectrogram dataset along with the PL-BERT embeddings, and fine-tuned the Auffusion model for speech spectrogram generation. He also contributed to writing the Abstract, Related Work, and Conclusion sections, and created the figures included in the final report and presentation.

Shruti generated spectrograms using the fine-tuned Auffusion model and fine-tuned the HiFi-GAN vocoder to reconstruct audio from the generated spectrograms. She also integrated the speech-adapted Auffusion model to enable the combined denoising process and built the full pipeline producing both images and audio from text prompts. She also wrote the Introduction, Methodology, and Qualitative Results sections of the report.

Krishanu evaluated the generated outputs using CLIP similarity and WER computed via Whisper, and analyzed the model’s performance across the full set of experiment prompts. He authored the Experiments section for the report.

References

- [1] Ziyang Chen, Daniel Geng, and Andrew Owens. Images that sound: Composing images and sounds on a single canvas. *Advances in Neural Information Processing Systems*, 37:85045–85073, 2024.
- [2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [3] Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation, 2024.
- [4] Keon Lee, Dong Won Kim, Jaehyeon Kim, Seungjun Chung, and Jaewoong Cho. Ditto-tts: Diffusion transformers for scalable text-to-speech without domain-specific factors, 2025.
- [5] Yuan Gao, Nobuyuki Morioka, Yu Zhang, and Nanxin Chen. E3 tts: Easy end-to-end diffusion-based text to speech, 2023.
- [6] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020.
- [7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.
- [8] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrigiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2018.
- [9] Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions, 2023.
- [10] Sehun Jung and Hyang won Lee. Learning generalizable prompt for clip with class similarity knowledge, 2025.
- [11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.