08th DECEMBER
2025

Images that Speak

Generating meaningful images that are speech spectrograms

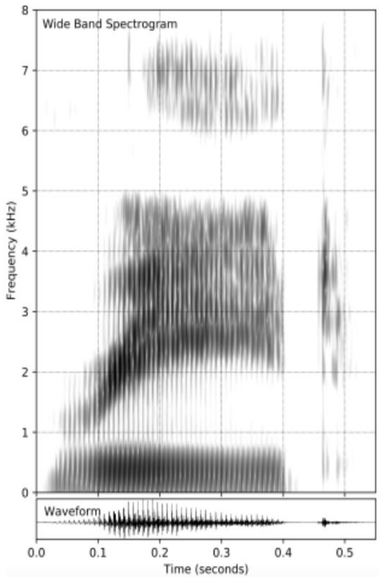
Project presentation for CS395T: Spoken Language Technologies

Hemanth Ram, Shruti Sriram, Krishanu Saini
The University of Texas at Austin

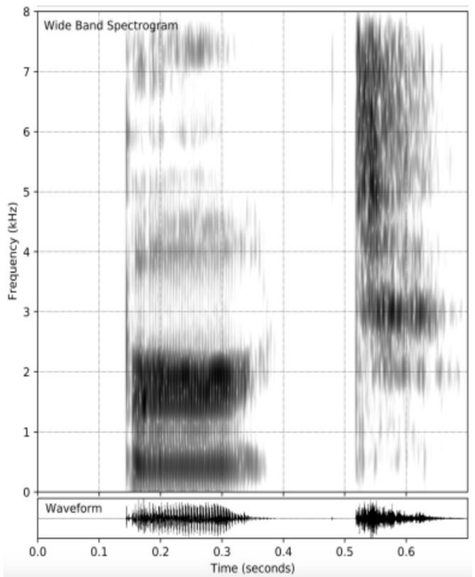
Introduction & Motivation

- Mel Spectrogram share statistical properties with natural images like edges and textures.

- I2S is a steganography method to encode a secret message in spectrograms.



reed
/riʸd/



Burt
/bɜ˞t/

Inspiration

Images that Sound

Composing Images and Sounds on a Single Canvas

Chen et al. (NeurIPS 2024)

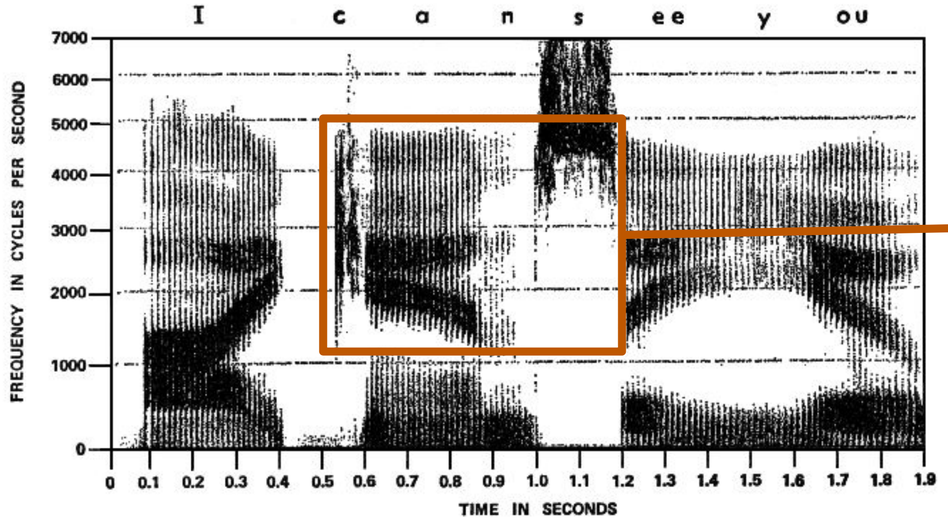
Images that Sound

Image prompt: a colorful photo of tigers

Audio prompt: tiger growling

Objectives

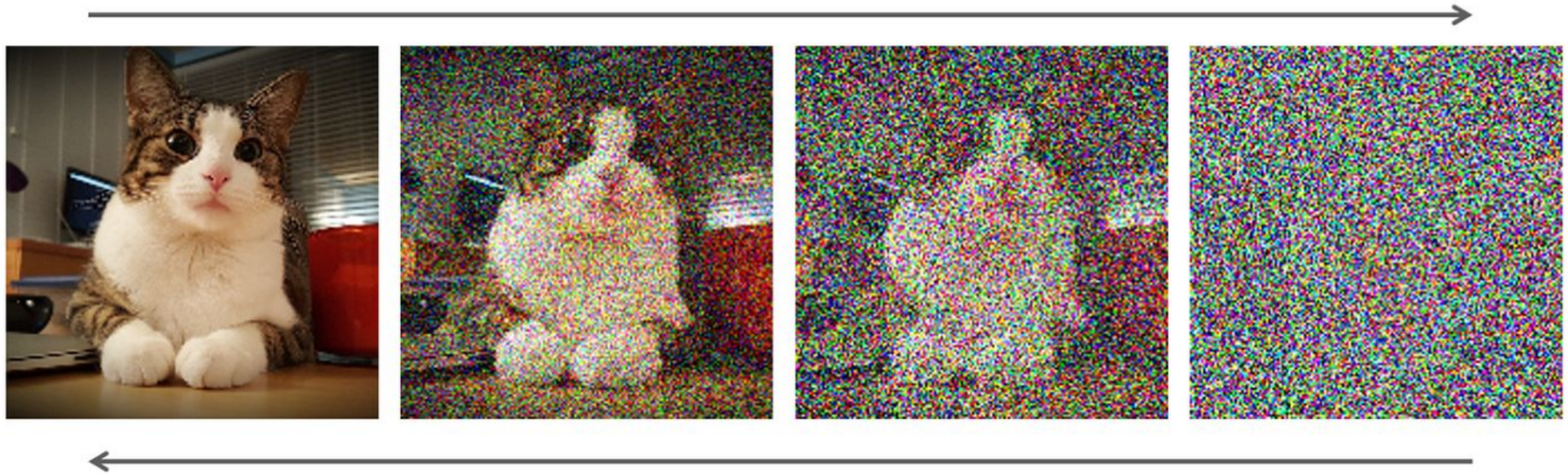
- Generate 2D matrices at the intersection of 2 modalities (vision and Audio - extend to speech)
  - semantically meaningful as images
  - produce meaningful speech when interpreted as spectrograms.



Structural and temporal Patterns
Bursts, Fricatives, Formants

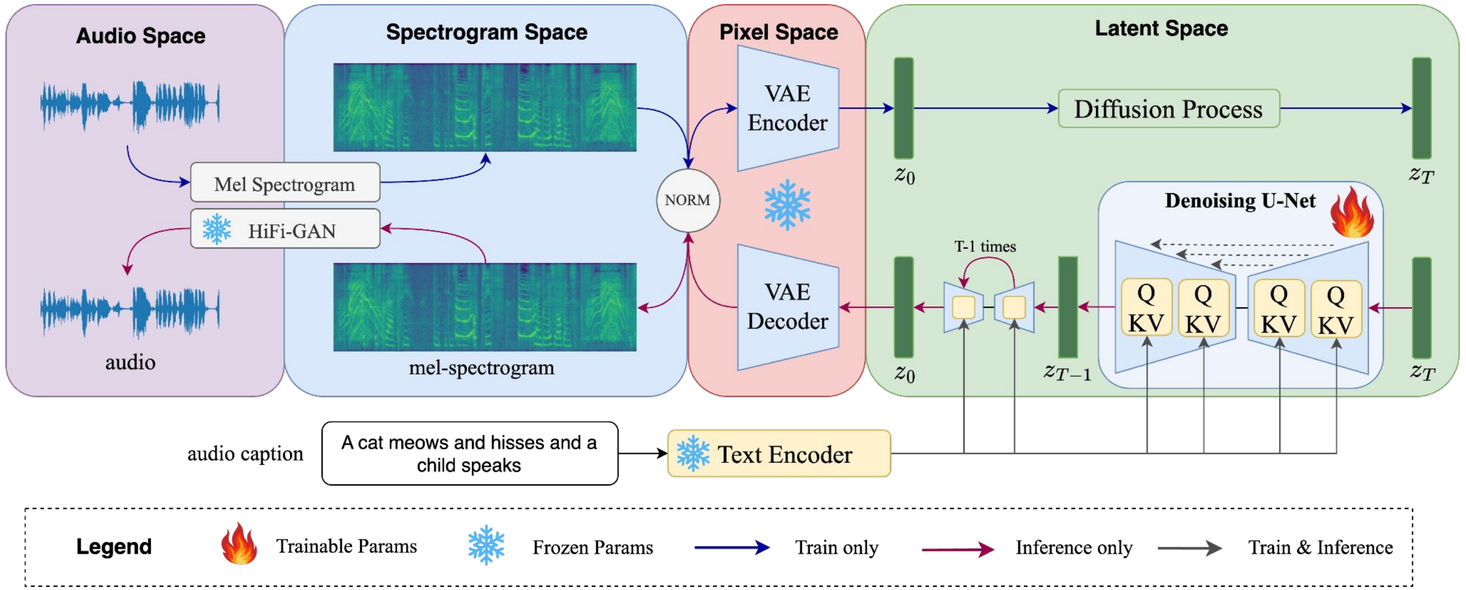
Stable Diffusion

- Diffusion models - reversing a Markov process that gradually adds noise to a data sample.
- Stable Diffusion is an open-source deep learning model conditions generation of an image from textual descriptions using a process called latent diffusion.

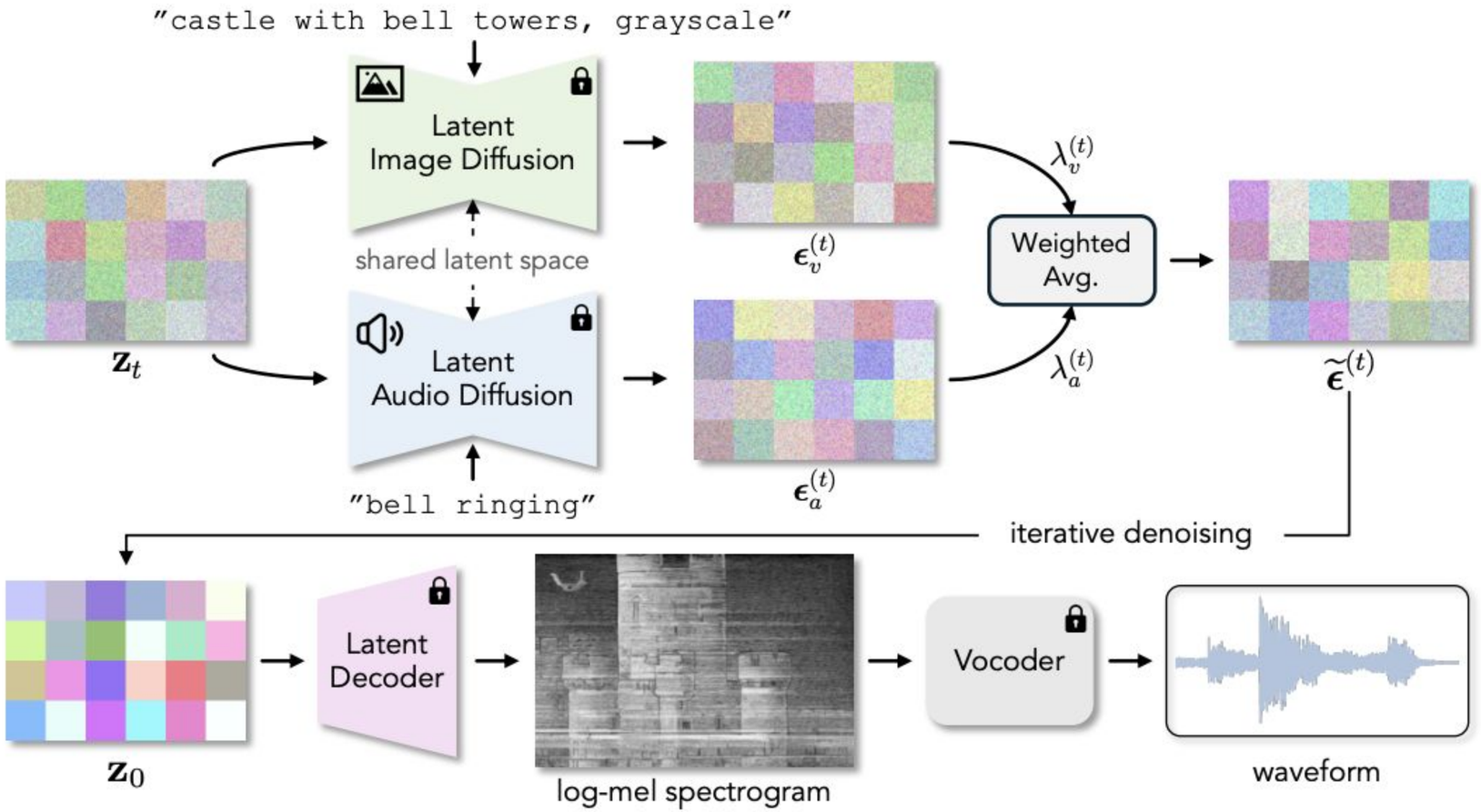


Auffusion

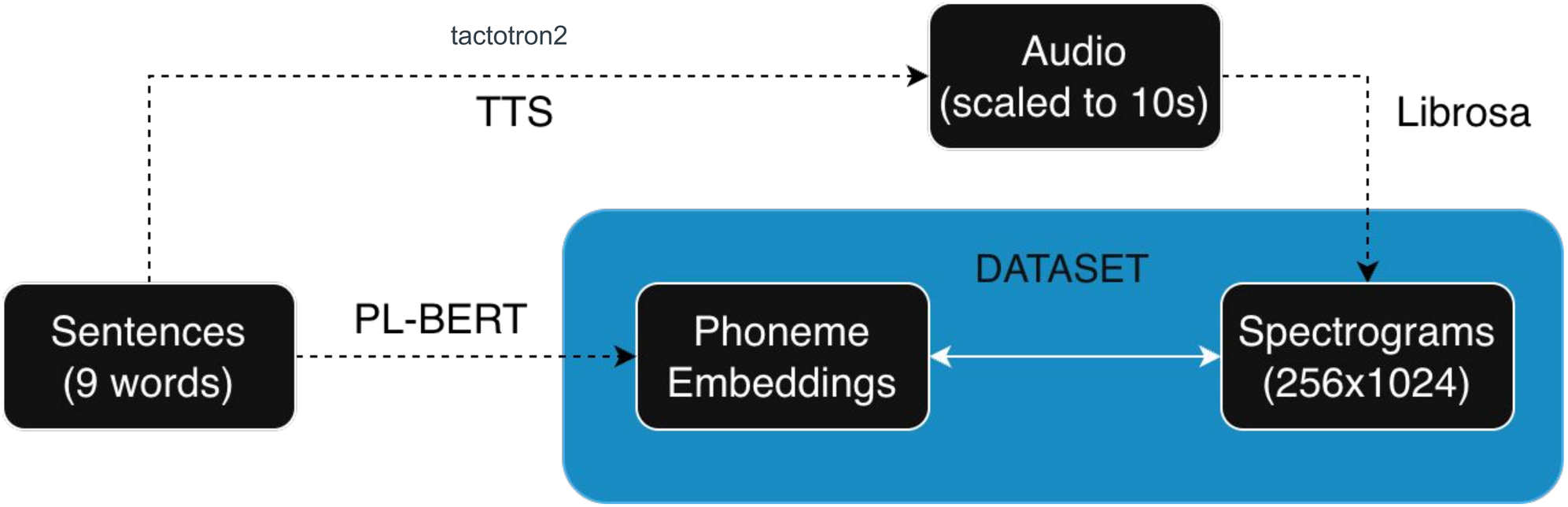
- TTA system adapting Stable Difusion T2I model framework.
- Auffusion generates text-conditioned sounds, human voices, and music.



Combining Image & Audio Diffusion

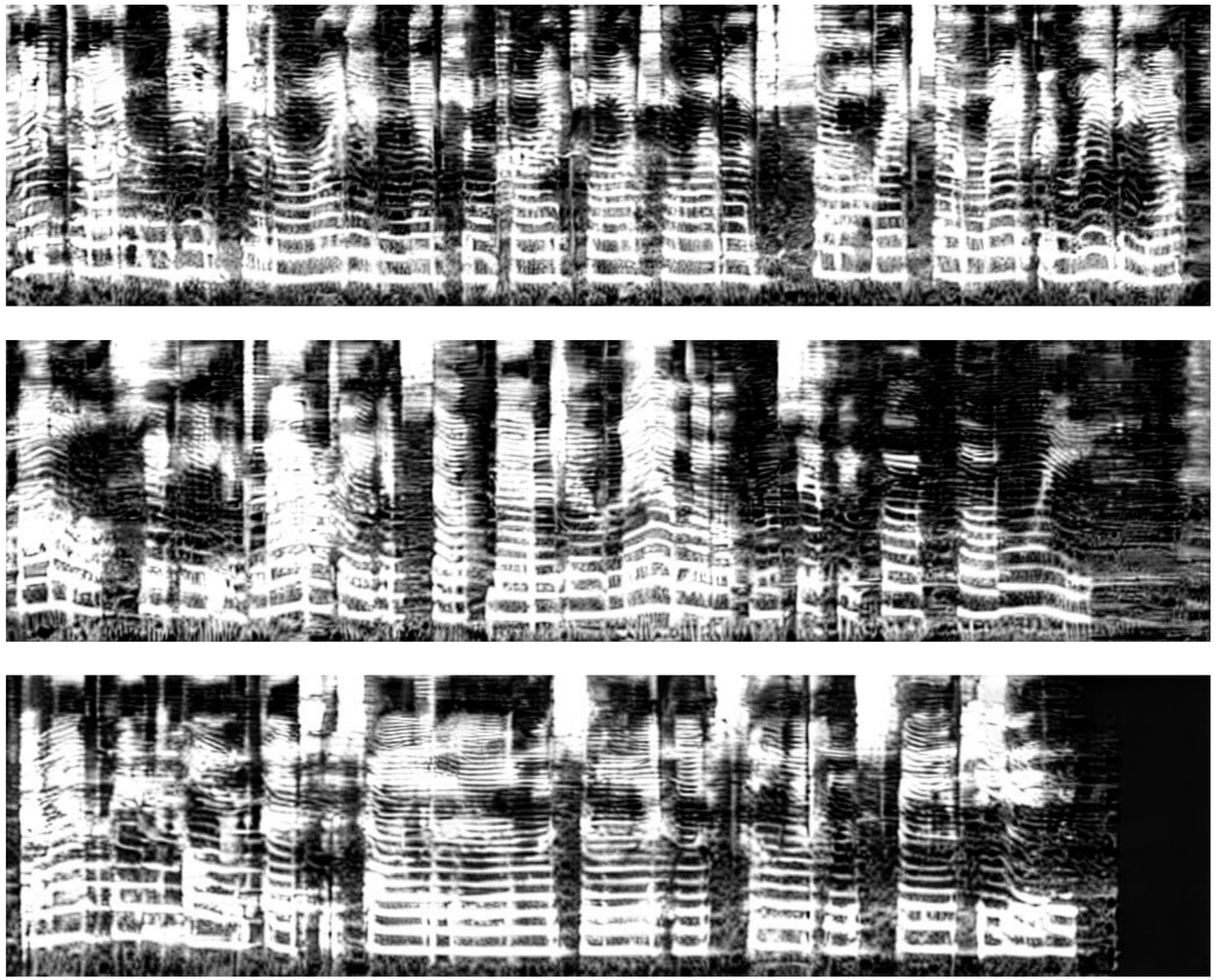


Speech Spectrogram Dataset



Fine Tuning Auffusion

- Added a phoneme embedding adapter before the encoder.
- Used LoRA fine tuning.
- Additionally unfreezed parts of base UNet to improve training.



Decoding Spectrograms

- Pretrained HiFi-GAN expects specific spectrogram settings
- Training HiFi-GAN on our data
- This aligns the vocoder to our spectrogram distribution, producing cleaner and more intelligible speech

Evaluation Metrics

- The objective has 2 parts
- Image-side metric: CLIP similarity
- Audio-side metric: Word Error Rate (WER)

Progress

- Created a human speech dataset
- Fine-tuned Affusion
- Integrated fine-tuned Auffusion with Stable Diffusion
- Trained HiFi-GAN to convert mel-spectrograms to audio
- Model evaluation

Outputs

Image Prompt:

a colorful photo of an auto racing game

Spectrogram Output:

Audio Prompt:

west yeah ability blue reality their

Audio Output:

Outputs

Image Prompt:

butterflies

Spectrogram Output:

Audio Prompt:

disease baby happen budget mouth down

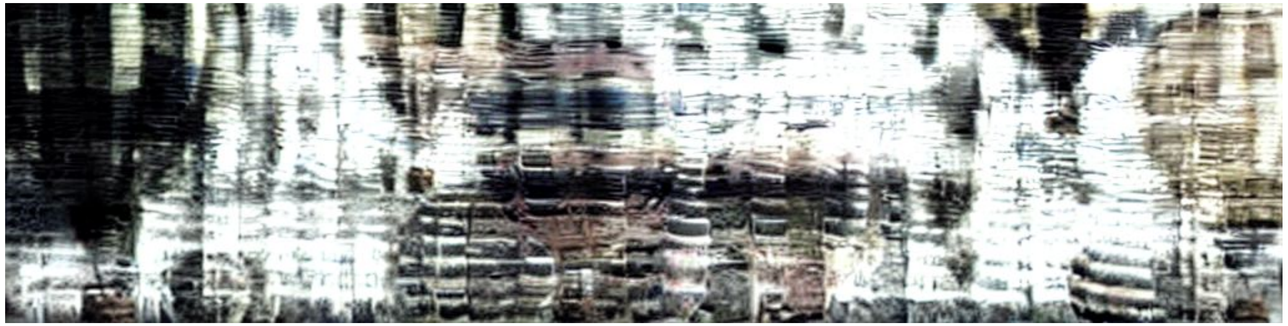
Audio Output:

Outputs

Image Prompt:

hot air balloons

Spectrogram Output:



Audio Prompt:

thing million risk range high allow

Audio Output:

Outputs

Image Prompt:

cats

Spectrogram Output:

Audio Prompt:

happy gone are exercising offered changing

Audio Output:

Code - Github