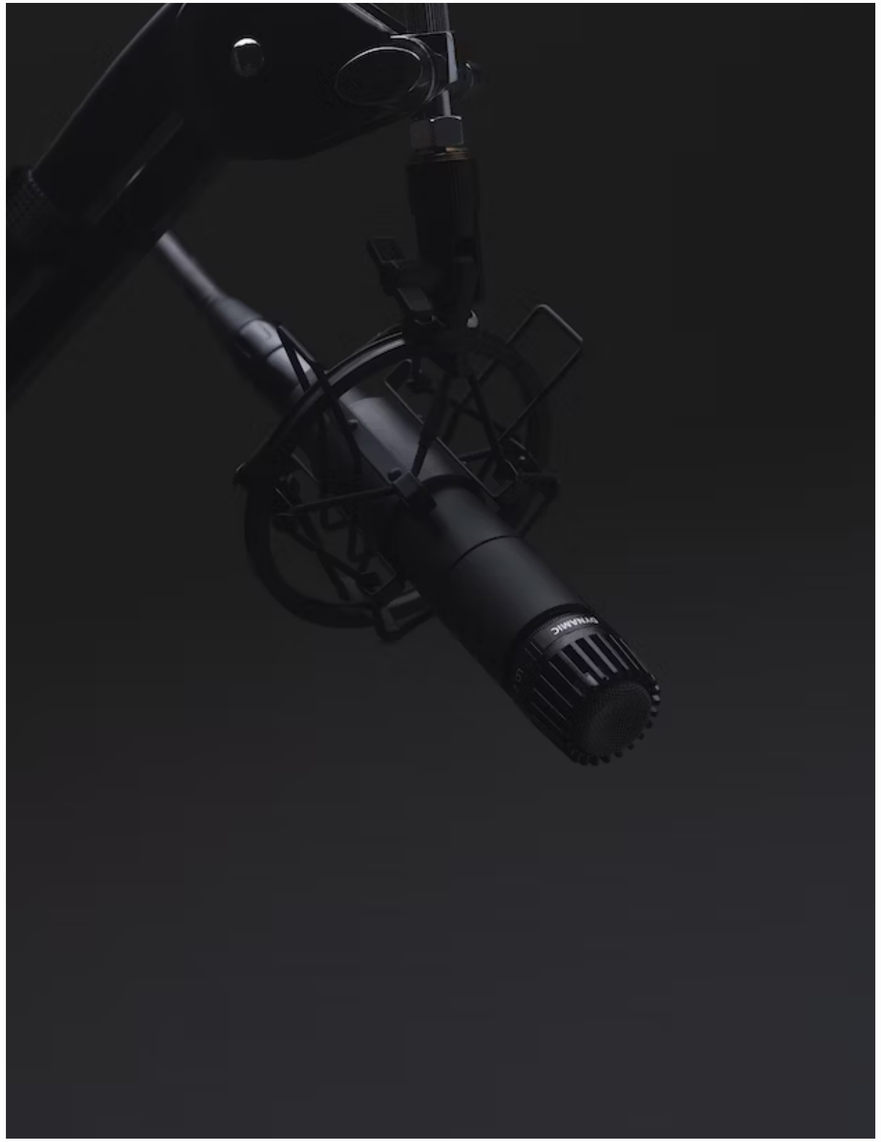
AudioVisio

G1T1

Minh × Bryan Lee × Jian Yi × Luke × Sheffield × Khai Soon

TABLE OF CONTENTS



PROBLEM STATEMENT

Streamlining Audio File Management & Enhancing User Experience

Managing and **organising audio files** can be **challenging**, especially with **large volumes** stored on file-management services.

Without visual cues such as thumbnails, **differentiating** audio files is difficult, and time-consuming to search for specific files.

Music and audio-sharing platforms like SoundCloud and Spotify **do not auto-generate thumbnails** for audio clips, making it harder for users to provide suitable visualisation for their content.

MOTIVATION

Improving Audio File Management and artistic visualization with thumbnail generation

Our project aims to **improve audio file management** by developing an **audio-to-image** conversion **web app** that generates thumbnails for audio files en-masse.

By converting audio to images, users will be able to **visually identify their audio content with ease**, reducing the time and effort required to manage their audio files.

This solution also has the potential to increase user efficiency and enhance the overall **user experience** in **audio and music-related applications**, improving the way users manage and interact with their audio content.

USE CASES & REAL-WORLD APPLICATIONS

AUDIO FILE MANAGEMENT

Persona

Anyone who uses File management systems to store audio files

- Students
- Entertainment industry professionals

Use Case

Bryan is working on a school project that involves **large volumes of audio files**. He needs a tool that can help him quickly **identify** and differentiate between **different tracks**.

Scenario Example

Bryan logs into his Google Drive account to search for a specific track for his project. With the audio-to-image conversion app, he can **generate thumbnails** for all his audio files en-masse and can quickly scan through the thumbnails to **find the file** he needs. This allows him to focus on what he does best – scoring **A+ for projects**.

THUMBNAILS FOR AUDIO SHARING PLATFORMS

Persona

Users of audio-sharing applications like SoundCloud and Spotify.

- Content creators
- Music industry professionals
- Social Media enthusiasts

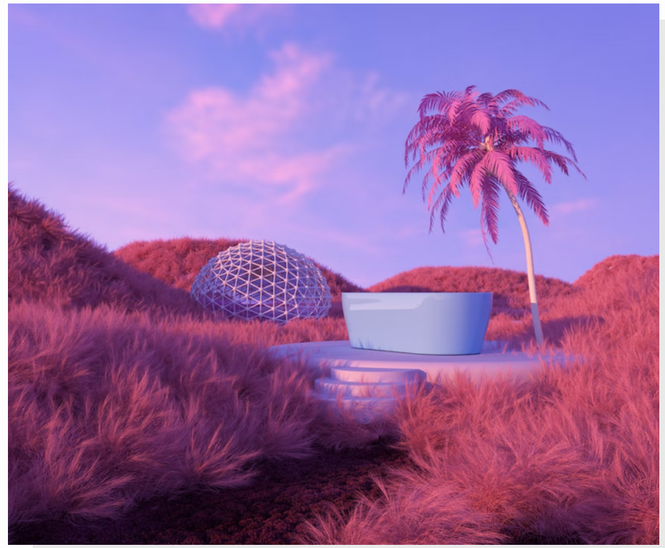
Use Case

Daniel is a **music producer**. His talent is composing music using **sounds in the environment**, but often struggles with finding a suitable thumbnail to express his art. He needs a tool that can help him **generate creative, high quality thumbnails** for his music.

Scenario Example

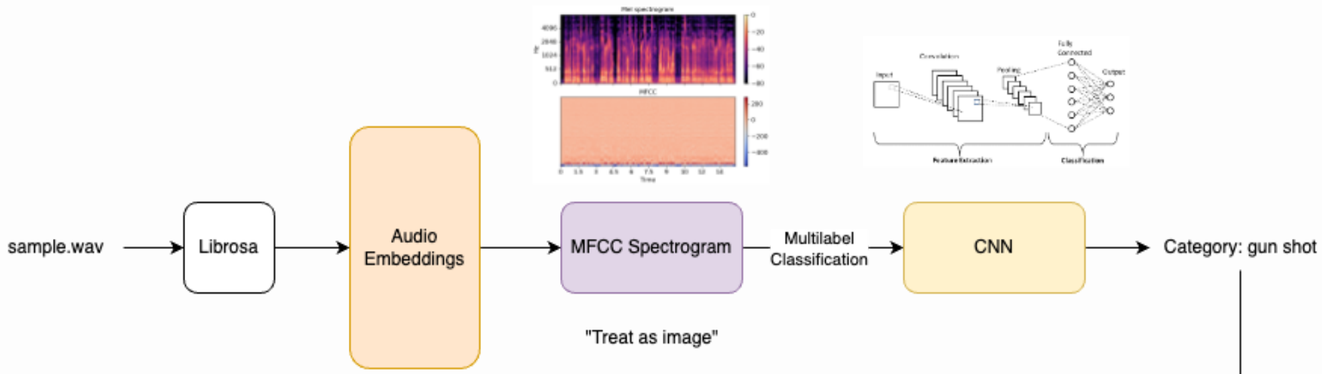
Daniel records a sample for an instrumental that involves **sounds in a neighborhood**. With the audio-to-image conversion app, he can **generate thumbnails** for his latest track by uploading it on the web app. The app generates an **image of a dog barking** as that is the most prominent sound in his sample. With that, he gets **further inspiration** to synthesize a killer beat.

HOW OUR PROJECT WORKS?

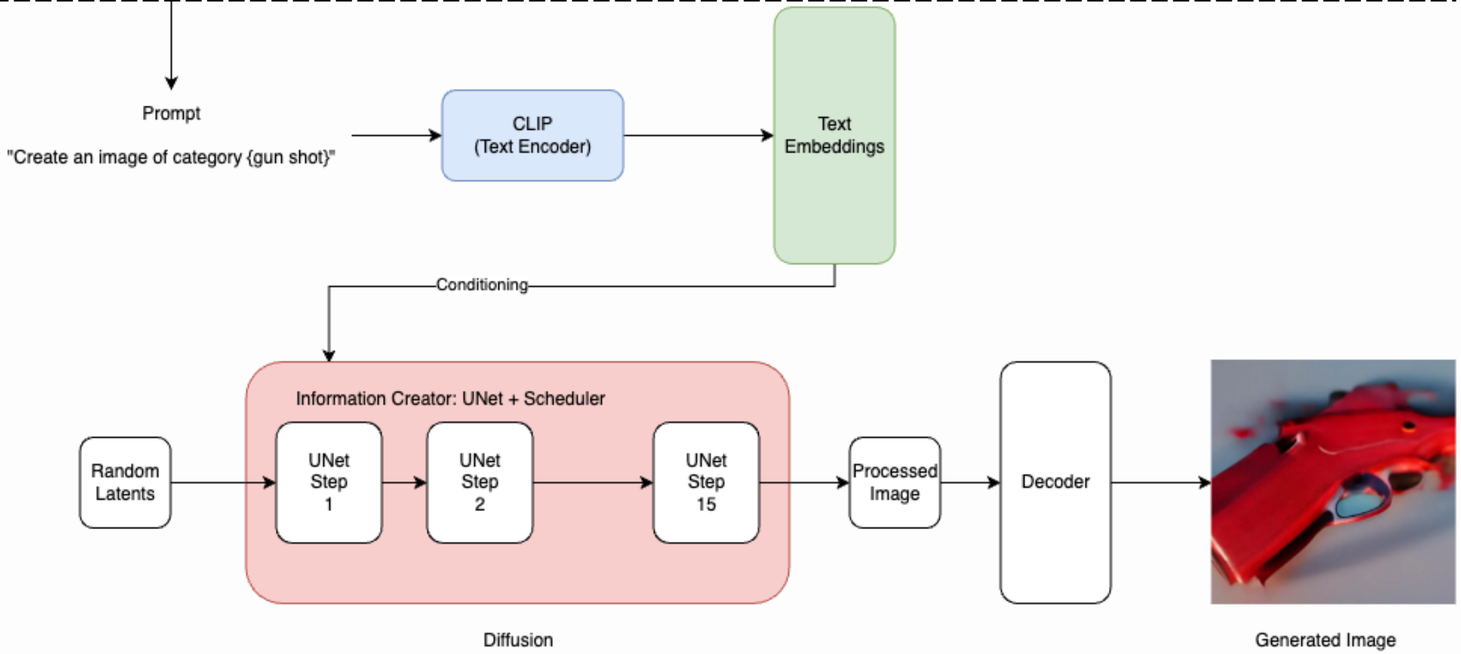


AUDIO2IMAGE PIPELINE

Audio Classification
(MFCC Spectrogram-CNN)



Stable Diffusion
(Prompt2Image)



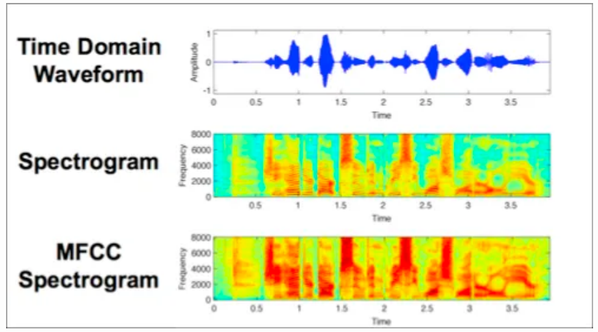
AUDIO2IMAGE PIPELINE

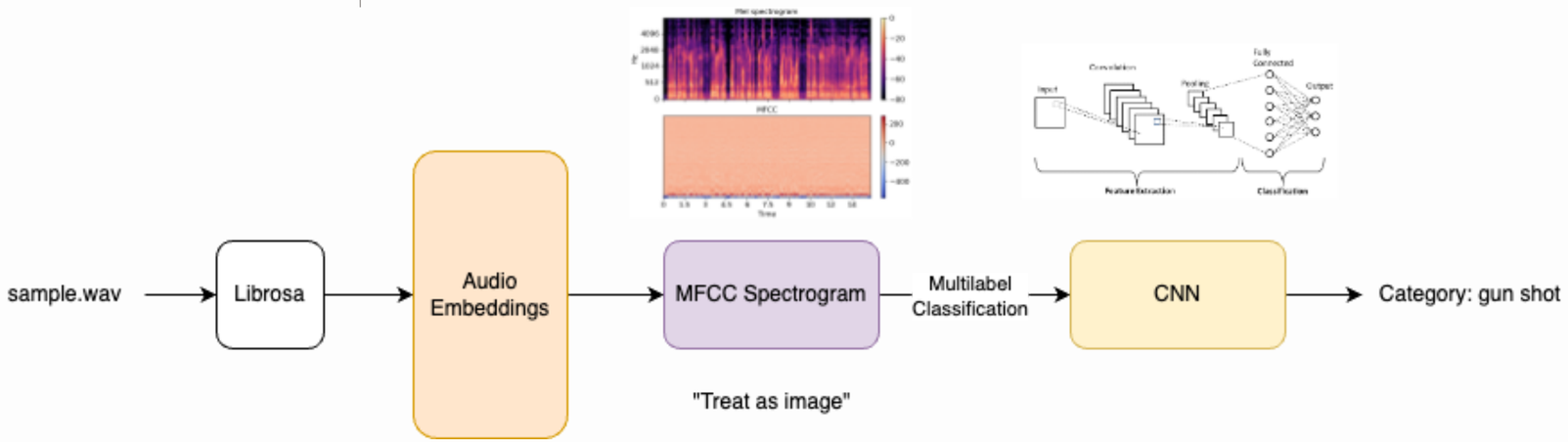
Audio Classification (MFCC Spectrogram-CNN)

Treat the problem as an image classification problem by converting audio into **spectrogram using MFCC technique**, after which it will be fed into a **CNN** to produce predicted category

Datasets

The dataset used for training the sound classification model is **Urbansound8K,** consisting of 8732 .wav files belonging to 10 classes. The test data consists of cleaned, unseen audio clips relevant to the 10 classes.



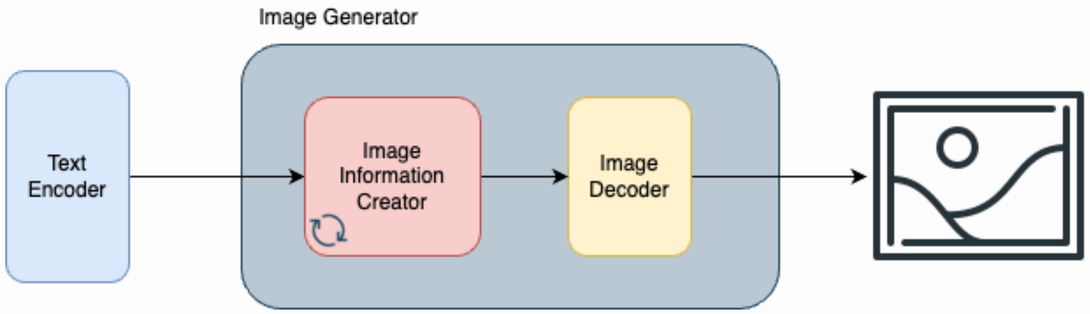


AUDIO2IMAGE PIPELINE

Stable Diffusion (Prompt2Image)

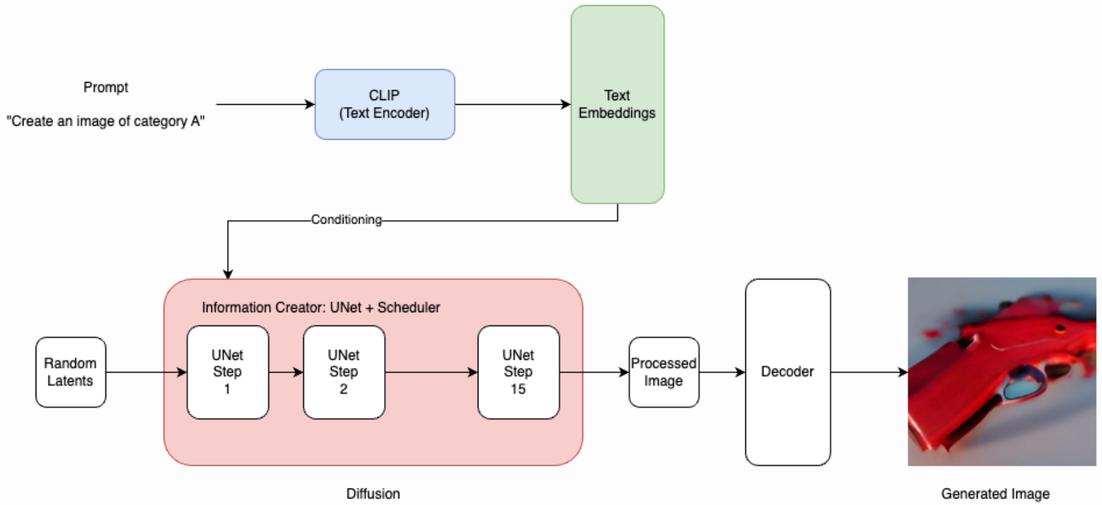
A system of several components that can be broken down to 3 high level components.

- **Text encoder** output numeric vectors, representing each word in the input text
- **Image information creator** works in the latent space that process information that leads to high quality image
- **Image decoder** that produces final pixel image

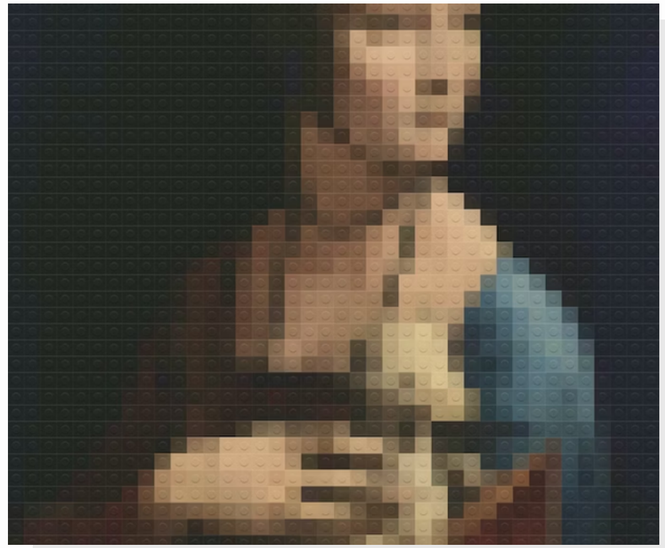


Pretrained models

- **ClipText** for text encoding
- **UNet** + **Scheduler** for gradually diffusing information into latent space



LIMITATIONS



LIMITATIONS

Clean Sound Profiles Required

Noisy sound profiles will result in unclear spectrograms which reduce classification accuracy

Short Clips Only

Data is only trained on four-second long clips

- Time dimension is compressed in longer clips

Difficulty in Differentiating Similar Sounds

Certain sound classes are similar in shape e.g.

1. repetitive sounds
2. sounds with sharp peaks

Image Generation Bias Towards Training Data

Image generation results are highly dependent on training data

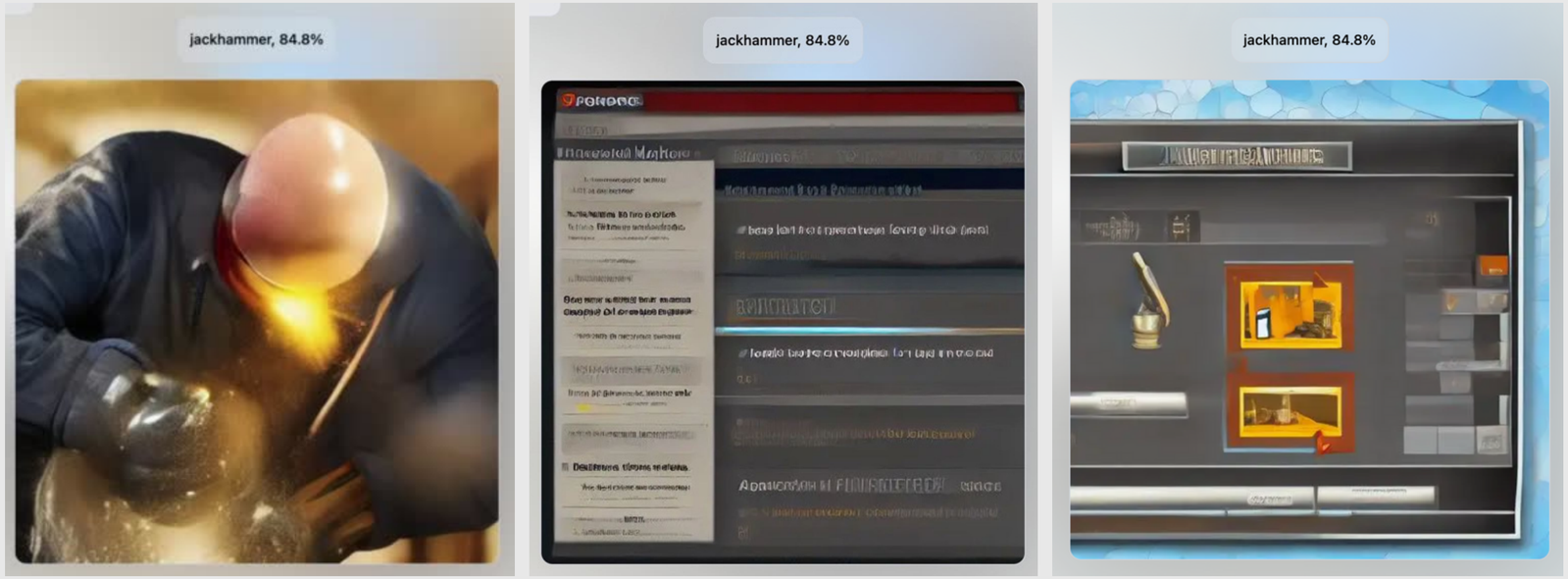
Reliance on Text Prompts

Audio clips still have to be converted into text prompts before image generation

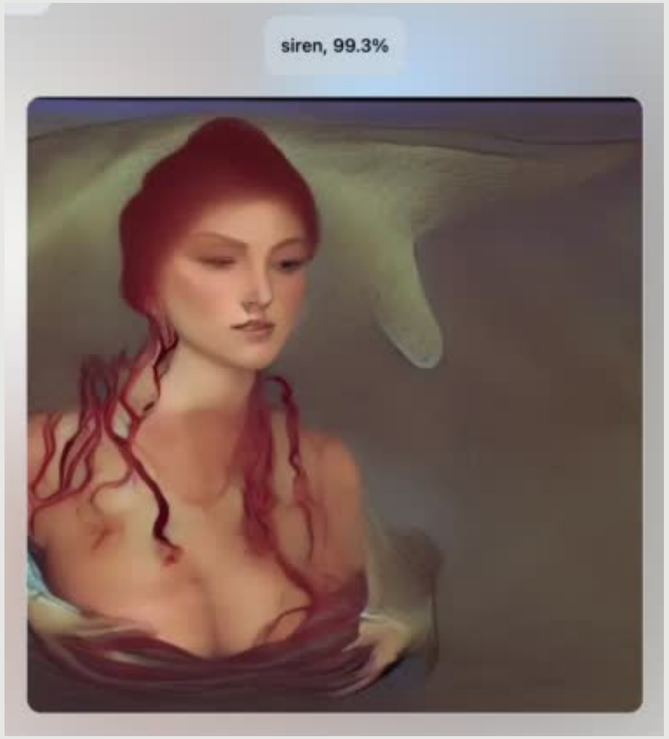
WRONG IMAGES

Bias in stable diffusion training leads to misunderstandings of prompts

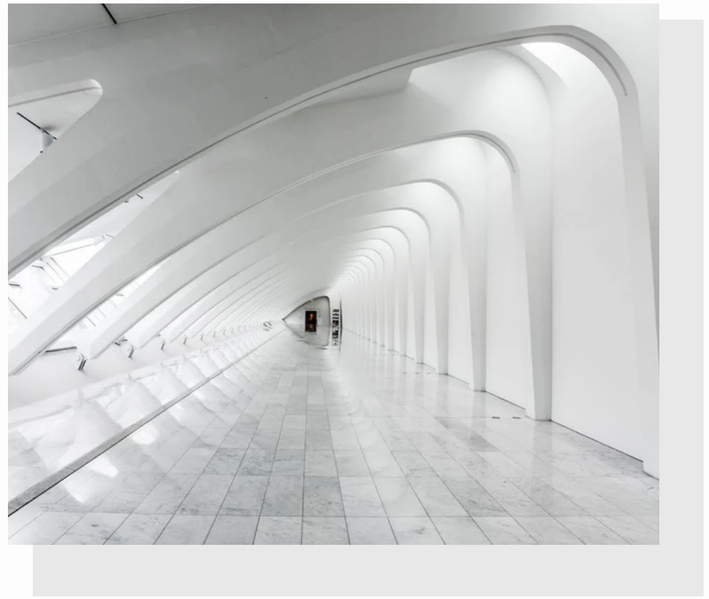
Are these really jackhammers?



Wrong siren...



CONCLUSION & FUTURE WORKS



CONCLUSION

Problem and Use Cases

Our web app:

- addresses the challenges of audio file management
- helps with audio visualization by generating thumbnails for audio files
- increases user efficiency and enhances the overall user experience for anyone who stores or shares audio files.

Our Project

Audio-to-image pipeline:

- Audio classification is done using a CNN that learns spectrograms of audio files to identify the category
- Stable diffusion: uses pretrained components such as CLIPText for text encoding and UNet + Scheduler algo for gradually diffusing information into the latent space to generate high-quality images.

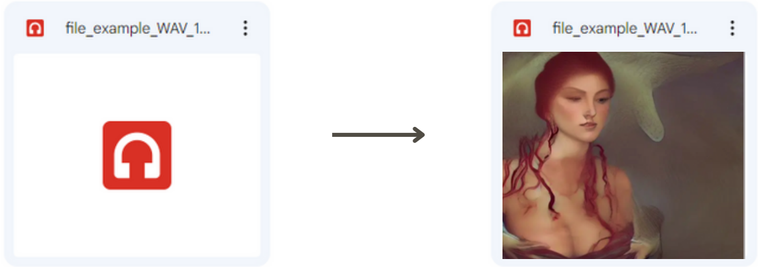
NEXT STEPS

Address Limitations

- Use noise reduction techniques like denoising autoencoders
- Collect and train on longer audio clips
- Incorporate additional features into the classification model
- Augment training data with more diverse sound profiles
- Use speech recognition software and unsupervised learning

Integrate with audio sharing platforms & file-management services

- SoundCloud
  - Auto generate thumbnail
- Dropbox, Google Drive
  - Replace template image for .wav files automatically



FUTURE WORKS & POSSIBLE EXPANSIONS

Audio to Video Generation

Improve audio visualisation to generating a video product:

- Creating music videos
- Enhancing podcast episodes with visuals
- Provide meaningful educational videos

Speech to Slide template Generator

Provide a easier way to create good looking presentations slides that fit the content of the speakers' needs:

- Creates visual aids, such as images and charts that help to illustrate the speaker's message
- Easier and faster to generate the slides format that represent the speakers' message and style.

Systems for safety and security

Provide a visual representation of environmental sounds to help users become more aware of their surroundings:

- Provide assistance to the hearing impaired
- Identify potential hazards or obstacles
- Alert authorities about potential dangers

AudioVisio

G1T1

Thank you!