Audio2Art: Transforming Voice Prompts into Visual Creations Using Transformers

Project Documentation

Introduction

Project Title: Audio2Art

Team Members:

1. Ayush Prajapati
2. Rutu Bhatt
3. Dipen Trivedi

Project Overview

Purpose:

Audio2Art bridges the gap between voice and vision by converting spoken language into stunning AI-generated images using transformer models. It is designed to assist artists, educators, and therapists in transforming creative or expressive audio prompts into rich visual artwork.

Features:

- Voice-to-image conversion using state-of-the-art transformer models
- Wav2Vec2-powered speech-to-text conversion
- Stable Diffusion-powered image generation
- Web interface built using Streamlit
- Colab-based deployment with GPU acceleration
- Public access via Localtunnel

Architecture

Model Workflow:

1. **Input**: User uploads a .wav audio file.
2. **Speech Recognition**: Wav2Vec2 transcribes the audio into a text prompt.
3. **Image Generation**: A Stable Diffusion pipeline interprets the prompt and generates a matching image.
4. **Output**: The user can view and download the image from the web interface.

Component Flow:

- ImageModel.py handles:
  - Audio processing via promptgen()
  - Image generation via text2image()
- app.py handles:
  - User interface with audio upload and model selection
  - Calling backend functions and displaying results

Setup Instructions

Prerequisites:

- Python ≥ 3.8
- Google Colab (GPU Runtime Recommended)
- Installed Libraries:
  - transformers
  - diffusers
  - torch
  - librosa
  - accelerate
  - streamlit
  - localtunnel

Installation Steps (Colab):

1. Clone/prepare the notebook environment.
2. Install required libraries:

   !pip install transformers diffusers torch librosa accelerate streamlit localtunnel
3. Use %writefile to create ImageModel.py and app.py directly in Colab.
4. Run the Streamlit app:

   !streamlit run app.py &
   !npx localtunnel --port 8501

Folder Structure

Since it's Colab-based, there is no local directory. However, virtual structure:

```
/content/
│
├── ImageModel.py    # Model initialization and logic
├── app.py           # Streamlit app UI
├── audio.wav        # Uploaded audio input (example)
```

Model and Function Documentation

promptgen(file)

- Input: .wav audio file
- Output: Transcribed text string
- Libraries used: librosa, transformers (Wav2Vec2)

text2image(prompt, repo_id)

- Input: Transcribed text, model ID from Hugging Face
- Output: PIL Image object
- Libraries used: diffusers, torch, time

Testing

Manual Testing (Colab + Streamlit UI):

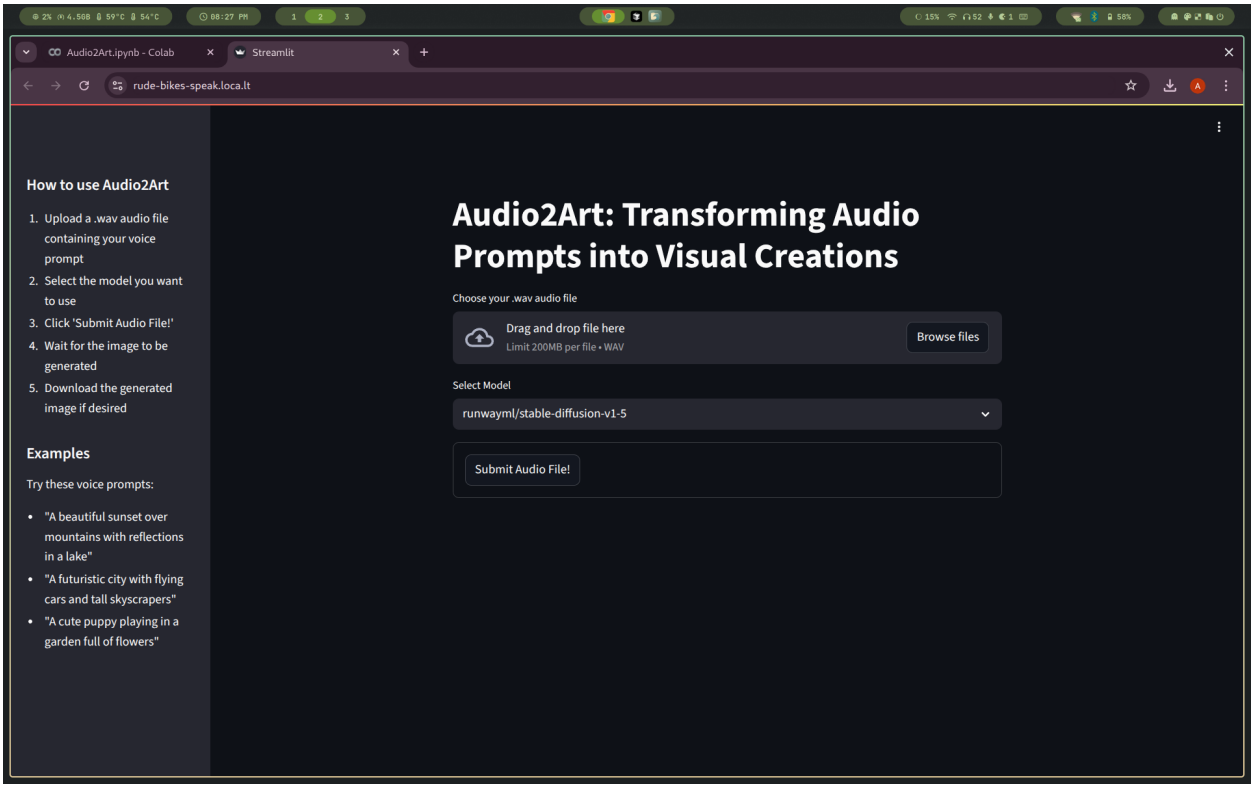
- Audio upload and transcription verified for various accent/pronunciation
- Image generation tested across multiple prompts and models (e.g., runwayml/stable-diffusion-v1-5)

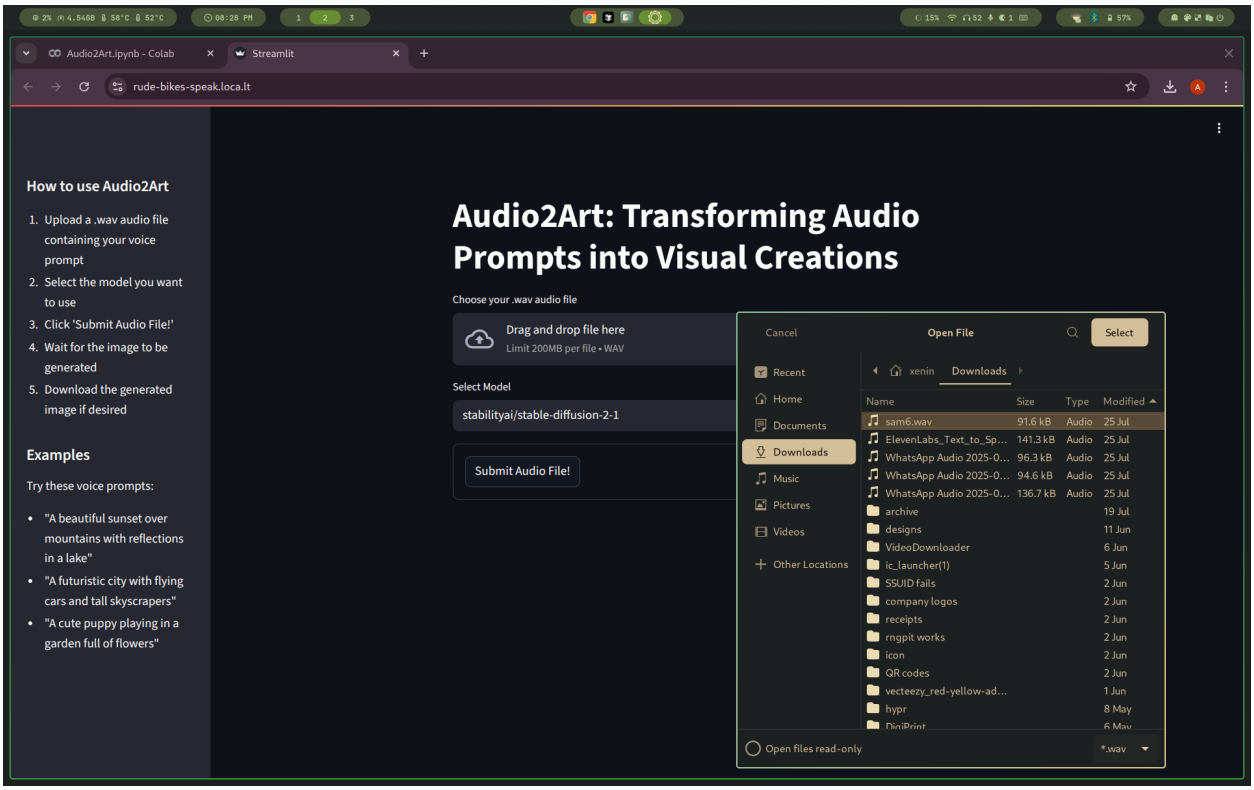
Planned Testing:

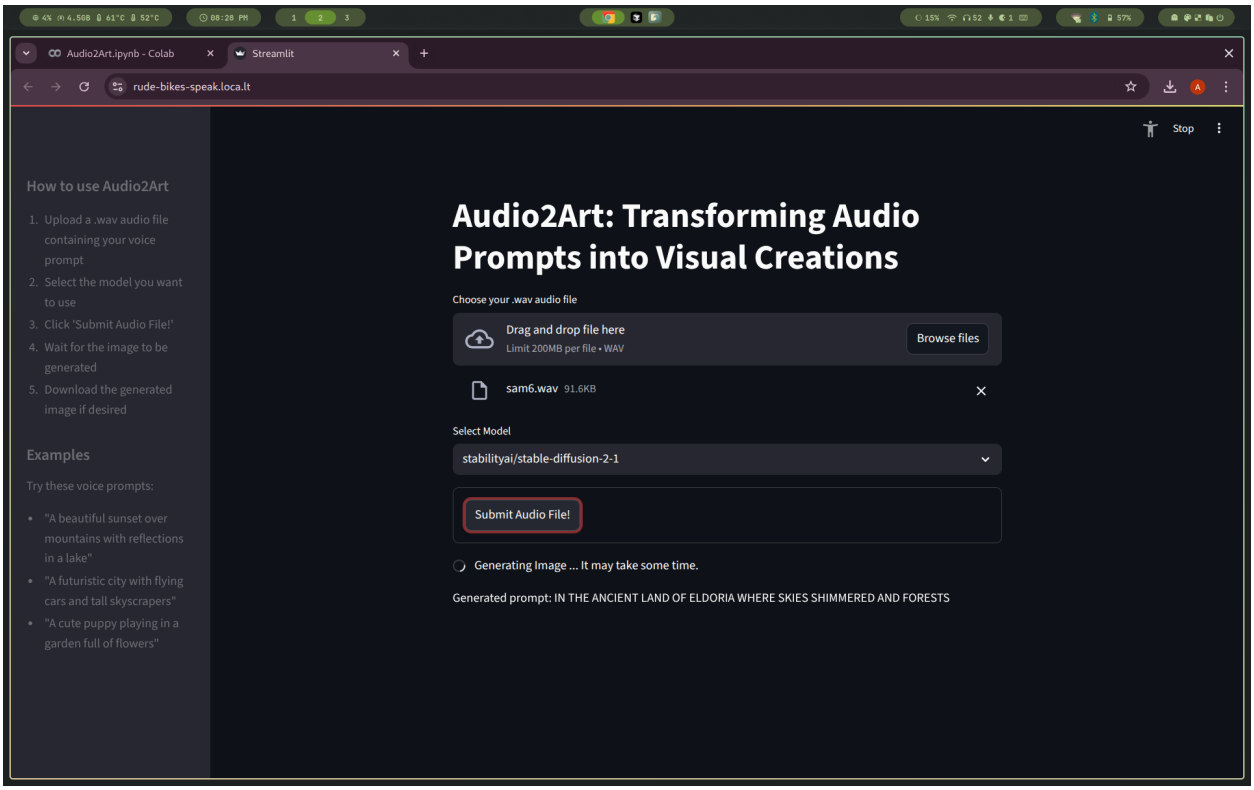
- Unit testing for promptgen and text2image using mock audio/text inputs (future scope)

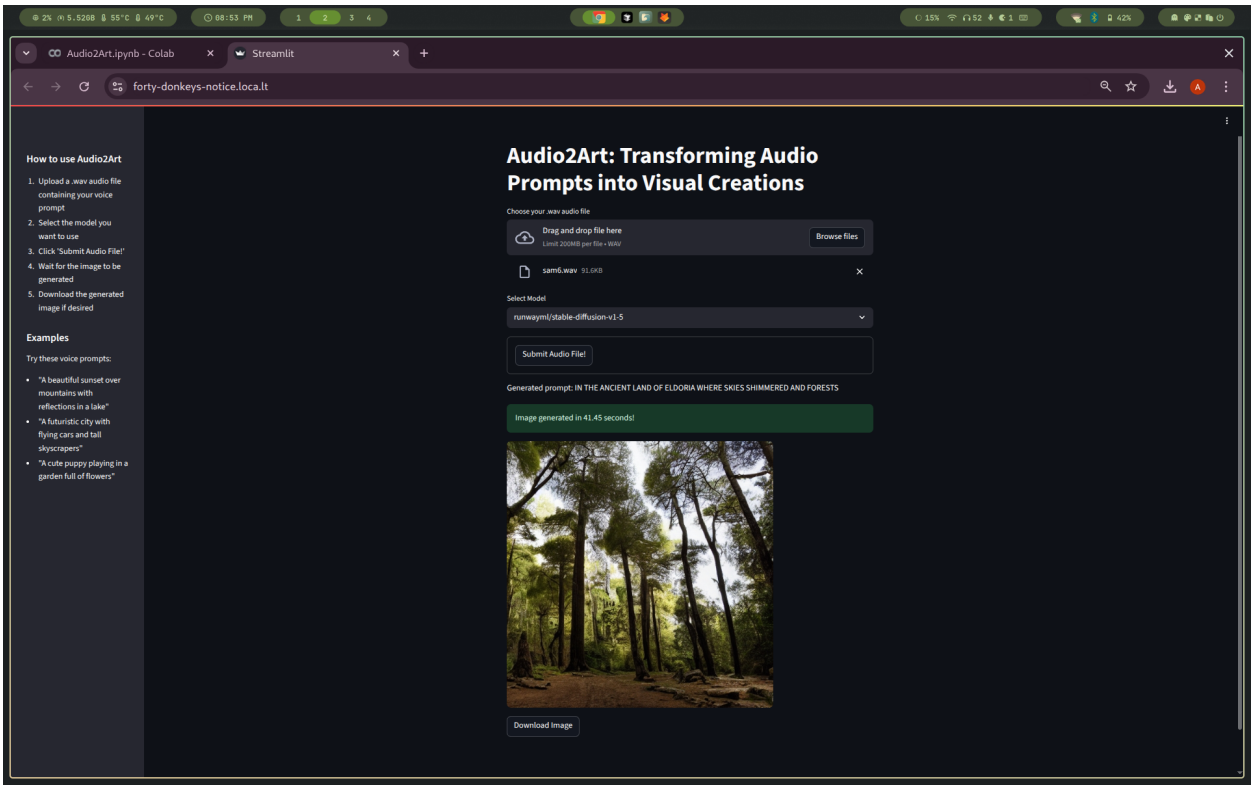
Screenshots or Demo

- UI Screenshot:




Audio2Art.ipynb - Colab
Streamlit
rude-bikes-speak.loca.lt
How to use Audio2Art
1. Upload a .wav audio file containing your voice prompt
2. Select the model you want to use
3. Click 'Submit Audio File!'
4. Wait for the image to be generated
5. Download the generated image if desired
Examples
Try these voice prompts:
• "A beautiful sunset over mountains with reflections in a lake"
• "A futuristic city with flying cars and tall skyscrapers"
• "A cute puppy playing in a garden full of flowers"
Audio2Art: Transforming Audio Prompts into Visual Creations
Choose your .wav audio file
Drag and drop file here
Limit 200MB per file • WAV
Select Model
stabilityai/stable-diffusion-2-1
Submit Audio File!
Cancel
Open File
Select
Recent
Home
Documents
Downloads
Music
Pictures
Videos
Other Locations
xenin
Downloads
Name
Size
Type
Modified
sam6.wav 91.6 kB Audio 25 Jul
ElevenLabs_Text_to_Sp... 141.3 kB Audio 25 Jul
WhatsApp Audio 2025-0... 96.3 kB Audio 25 Jul
WhatsApp Audio 2025-0... 94.6 kB Audio 25 Jul
WhatsApp Audio 2025-0... 136.7 kB Audio 25 Jul
archive 19 Jul
designs 11 Jun
VideoDownloader 6 Jun
ic_launcher(1) 5 Jun
SSUID fails 2 Jun
company logos 2 Jun
receipts 2 Jun
rngpit works 2 Jun
icon 2 Jun
QR codes 2 Jun
vecteezy_red-yellow-ad... 1 Jun
hypr 8 May
Open files read-only
*.wav


Audio2Art.ipynb - Colab
Streamlit
rude-bikes-speak.loca.lt
Stop
How to use Audio2Art
1. Upload a .wav audio file containing your voice prompt
2. Select the model you want to use
3. Click 'Submit Audio File!'
4. Wait for the image to be generated
5. Download the generated image if desired
Examples
Try these voice prompts:
• "A beautiful sunset over mountains with reflections in a lake"
• "A futuristic city with flying cars and tall skyscrapers"
• "A cute puppy playing in a garden full of flowers"
Audio2Art: Transforming Audio Prompts into Visual Creations
Choose your .wav audio file
Drag and drop file here
Limit 200MB per file • WAV
Browse files
sam6.wav 91.6KB
Select Model
stabilityai/stable-diffusion-2-1
Submit Audio File!
Generating Image ... It may take some time.
Generated prompt: IN THE ANCIENT LAND OF ELDORIA WHERE SKIES SHIMMERED AND FORESTS



- Demo URL (Localtunnel/Colab): *[To be generated dynamically during runtime]*

Known Issues

- Latency in image generation on CPU-only environment
- Occasional mis-transcription due to unclear audio
- Limited model customization via UI (single model selection)

Future Enhancements

- Add support for MP3/other audio formats
- Multiple image outputs with style variations
- User authentication & image history
- Enhanced audio noise filtering before transcription
- Mobile-responsive UI via Streamlit customization or React frontend