Audio-to-Image Generation: A Pipeline for Sound Analysis and Visual Synthesis

Vijayalakshmi M., Aayush Anshul and Kishu Raj Tyagi
Department of Computing Technology, SRMIST, Kattankulathur, Chengalpattu, Tamil Nadu, India

Keywords: Audio-to-Image Translation, YAMNet, Whisper Model, Image Synthesis, Cross-Modal Learning.

Abstract: This paper presents a novel pipeline for transforming audio into high-resolution images, leveraging advanced neural networks and modern generative models. The proposed approach integrates robust audio feature extraction, cross-modal mapping, and state-of-the-art image synthesis techniques to enhance the fidelity and generalizability of audio-to-image generation. Utilizing deep learning architectures such as YAMNet for sound classification, Whisper for speech recognition, and Stable Diffusion for high-quality image synthesis, the system ensures scalability and realism. Additionally, real-time processing and an interactive user feedback mechanism enable iterative refinement, optimizing the relevance and precision of generated visuals. The proposed methodology holds significant potential across various domains, including multimedia content creation, environmental monitoring, and educational applications, offering a transformative step toward seamless audio-visual synthesis.

1 INTRODUCTION

With advancements in artificial intelligence, creating reasonable images from audio data has become a promising area of research. This paper introduces an Audio-to-Image pipeline that leverages deep learning networks and signal processing techniques to convert audio input into semantically correct images. By combining various modules such as environmental sound classification, speech recognition, and generative AI, the system can effectively convert audio signals and produce corresponding visual outputs. The aim is to bridge the gap between hearing and visual imagination and provide a tool that can be applied across a broad spectrum of applications, ranging from media creation and accessibility enhancement to automated content creation.

The motivation for this system comes from the power of converting sound events in real life into entertaining graphics so that people can imagine what they listen to. It is particularly useful in cases where sound cues visualization can be applied to enhance narrations, accessibility for the hard of hearing, and automatic scene generation. Utilizing state-of-the-art AI models such as YAMNet, Whisper, and Stable Diffusion, the project promises high-quality image generation based on the context of the audio that is extracted. The proposed pipeline presents a new paradigm for multimodal AI applications by combining the field of sound analysis with visual generation, opening up the artificial creativity horizon. The other key advantage of this project is that it can generate high-fidelity visual results that record not just environmental audio but also speech material. The multimodal nature of the system is such that it can be applied to any industry, from entertainment and virtual reality experiences to security and surveillance. The continued developments in AI-image creation and sound recognition technology improve continuously the precision and legitimacy of the produced content, which is an exciting field of AI research and application. This work is an advancement in the field of AI-image generated multimedia, creating new possibilities of auditory-visual experience interaction. These advances have the potential to transform industries by enabling AI to naturally interpret and visualize sound, creating immersive experiences that integrate auditory and visual modalities.

The ability of the Audio-to-Image pipeline to close the gap between sound and visual representation has far-reaching implications for numerous industries. From an accessibility perspective, the technology can potentially empower the hearing impaired by providing a visual representation of their sound world. In the world of entertainment, game developers and

M., V., Anshul, A. and Tyagi, K. R.
Audio-to-Image Generation: A Pipeline for Sound Analysis and Visual Synthesis.
DOI: 10.5220/0013907500004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 4*, pages 24-32
ISBN: 978-989-758-777-1

producers may employ this system to generate scene concepts from ambient soundscapes. The surveillance and security market may employ the technology to develop visual alarms for sound events, such as alarm ringing or breaking glass. Such general applications make it crucial to establish a robust and scalable audio-to-image transformation system.

2 RELATED WORKS

Study There are a number of studies that have investigated the conjunction of audio and image processing that have helped establish multimodal AI systems. Wu et al. 2016 proposed a cross-modal ranking analysis approach to fill the gap between music and image modalities and show how it is possible to enhance alignment by using structured relations between audio and visual data. Zheng et al. developed the concept further based on adversarial metric learning for audio-visual matching, advancing the capability of deep learning-based models to connect sound with their corresponding images.

Hsu et al., 2018 introduced a local wavelet acoustic pattern descriptor for birdsong classification, underlining the role of time-frequency analysis in audio classification. In the same way, Gemmeke et al., 2017 presented the Audio Set dataset, a large human-labeled audio event classification dataset, that forms the basis of most contemporary sound classification algorithms.

Yang et al., 2024 proposed an audio-to-image generation model with semantic and feature consistency, tackling major issues in audio-driven visual synthesis. Widyagustin et al. employed the YAMNet model for classifying lung sounds and proved its usefulness in detecting disease-related audio patterns. Valliappan et al. 2024 implemented YAMNet for gun detection, highlighting its versatility in security use cases. Rochadiani et al., 2023 compared transfer learning approaches to environmental sound classification, further establishing the efficacy of pretrained models.

Haz et al. tested the Whisper model's capabilities for speech-to-text translation, affirming its high transcription accuracy. Abdulsalam et al., 2003 investigated Stable Diffusion for AI image synthesis, shedding light on how to optimize diffusion models for high-quality visual output.

Kamachi et al., 2014 investigated identity matching between audio and visual modalities, focusing on the psychological nature of cross-modal perception. Goodfellow et al. 2002 proposed Generative Adversarial Networks (GANs), which are the basis for most contemporary generative AI methods, such as audio-to-image synthesis. Tzanetakis and Cook 2008 created a musical genre classification system, adding to early work on computational audio analysis.

Kliegr et al., 2002 examined the use of image captions and visual analysis for enhanced image concept classification. Dupont and Luettin 2011 introduced an audio-visual speech modeling framework, paving the way for multimodal speech recognition. Alameda-Pineda et al. 2019 investigated the detection of audio-visual events in social settings, affirming the need for contextual multimodal understanding.

Wan et al. 2016 introduced a GAN-based technique for generating scene images from audio, which is very much in line with the goals of the proposed system. Reed et al. 2016 and Chen et al. 2017 extended generative adversarial methods for text-to-image and audio-visual synthesis, respectively. These works establish a solid theoretical and practical ground for building stable audio-to-image pipelines.

3 PROPOSED SYSTEM

The system that has been proposed aims to process audio files and create images that clearly depict the material of the audio as shown in Table 1. It comprises several parts which collaborate to pick out meaningful content from sound and apply it in creating a visual representation. The pipeline starts by loading audio files from Google Drive, environmental sound classification, and speech recognition. These features are then utilized to craft prompts, which instruct the image generation process via a diffusion model. This multi-step process guarantees generated images with high contextual relevance and coherence.

Table 1: Comparison between existing and proposed system.

Features	Academic Precedents	This System
Environmental Handling	Single-label classification	YAMNet's 521-class taxonomy
Speech Integration	Text-only transcription	Whisper + validation gates
Error Handling	Basic thresholding	Hybrib fallback system

One of the major innovations of the system being presented is its capacity to combine environmental and verbal components to produce images with a higher degree of contextual richness. Unlike traditional image-generation systems that only use textual prompts, the system incorporates multimodal data to generate images that are actually descriptive of the actual audio environment. By using a structured prompt generation technique, the system guarantees that the resulting output closely matches the input audio features. This makes the resulting images more real-like by incorporating several layers of contextual information.

For accuracy and continuity, the system combines aggressive preprocessing methods like resampling, noise removal, and classification thresholding. The biggest strength of the system lies in analyzing both environmental and voice content, making it a wide-range tool for many fields, such as security, entertainment, and accessibility. Through the incorporation of optimizations like FP16 acceleration and GPU-based inference, the pipeline is optimized for efficiency and scalability. This guarantees that the model runs within a sensible computational budget with high output quality. The modular design of the system also enables simple future extension and flexibility across various applications.

The proposed system's success hinges on its capacity to process sound effectively and produce a meaningful visual output. Toward this end, the system is structured with a focus on modularity, efficiency, and scalability. Utilizing cloud-based resources allows the system to engage in big-scale processing without compromising performance. The capacity for further integration with other AI models and frameworks also makes this pipeline an attractive basis for future multimodal AI studies and applications.

4 METHODOLOGY

The methodology of the proposed system is to convert audio to images. The initial step is to load the audio file, which is then processed through the YAMNet model for classifying sounds in the environment. Concurrently, the Whisper model is employed to transcribe speech, deriving textual information from the audio input. The second step entails prompt engineering, in which a well-defined template integrates accurate classifications and speech transcription into an accurate descriptive prompt to be used in image creation. This formalized process makes the end result uphold both visual and auditory consistency.

The efficacy of the system is increased through the use of complex processing mechanisms like confidence thresholding and temporal averaging to preprocess and sharpen the derived features. By taking advantage of the strengths of machine learning models, the system only utilizes data that is meaningful and relevant while generating prompts. This reduces noise and meaningless information, hence enhancing the coherence of the images generated. Moreover, unnecessary labels like "background noise" are removed in order to keep the focus on significant audio factors.

Stable Diffusion XL is utilized for image generation, using inference parameters optimized like controlled guidance scale, widescreen resolution, and pre-defined negative prompts. Saving the output images and showing results in a Colab environment is the last step. Error handling techniques like CUDA memory fallbacks and NSFW content filtering make it robust across different deployment contexts. The integration of state-of-the-art AI models and structured processing guarantees that the system runs smoothly, delivering high-quality results every time as shown in Figure 1. The approach is scalable, and future enhancements in the efficiency and accuracy of the pipeline are possible.

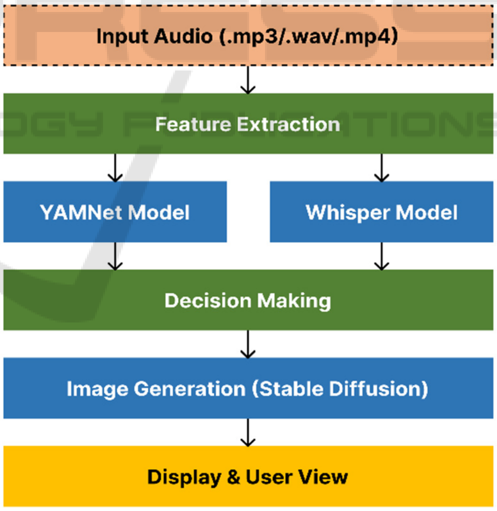

Figure 1: Architecture diagram.

The modularity of the approach provides flexibility in the extension of the system to accommodate new features. With the integration of progress in AI-based audio analysis and image generation, the pipeline can be developed to facilitate real-time audio-to-image translation. Additional optimizations can be made to increase the precision of sound classification, enrich the contextual depth of the generated prompts, and allow the system to accept a wider variety of audio inputs. As the research continues, the incorporation of more machine learning methodologies, including reinforcement learning and transformer-based models, could continue to enhance the system's performance.

5 MODULES USED

The Audio-to-Image pipeline consists of various expertise modules, each of which serves an important purpose in mapping raw audio information to meaningful visual outputs. Each module is designed to ensure the system has high accuracy and efficiency in processing, allowing seamless flow from input receipt to final image creation. The following is a description of each module, its purpose, and how it helps the overall pipeline.

5.1 Audio File Loader

Audio File Loader is used to preprocess input audio files in a way that they are in the correct format for subsequent analysis. The module uses librosa, a well-known Python package for audio processing, and the Google Drive API to download and preprocess audio files on cloud storage.
There are a number of preprocessing steps involved before the audio is handed over to the following modules. They are:

- Resampling: The audio is resampled at a sample rate of 16 kHz, as required by models like YAMNet, for compatibility with varying inputs.
- Mono Channel Conversion: Stereo audio input, if available, is converted to mono for ease of processing.
- Normalization: Audio levels are normalized to provide uniformity and prevent distortion during subsequent processing.
- Trimming & Padding: Silent or unwanted segments of the audio are trimmed, and brief audio samples can be padded to have an even input length.

By carrying out these vital preprocessing steps, the Audio File Loader cleanses the audio input to be structured and ready for the next analysis models.

5.2 YAMNet Audio Analyzer

The YAMNet Audio Analyzer uses TensorFlow Hub's YAMNet model for audio classification. YAMNet is a deep learning-based model that was trained on an enormous dataset of environmental sounds and can classify more than 521 distinct sound events refer to Figure 2. This module is important for determining the environmental context of the audio, which helps in producing meaningful visual outputs.
Some of the main functionalities of the YAMNet Audio Analyzer are:

- Feature Extraction: The model identifies useful sound features, determining events like "dog barking," "ocean waves," or "background chatter."
- Temporal Averaging: As sounds fluctuate with the passage of time, the module averages out the classifications throughout the length of the audio file for consistency.
- Threshold Filtering: A confidence threshold (usually 0.15) is used to remove uncertain or irrelevant classifications so that only salient sounds are used to generate the final image.

Through the analysis of the environmental sound, this module offers contextual information that improves the quality of image generation so that the system can represent real-world scenes more precisely.

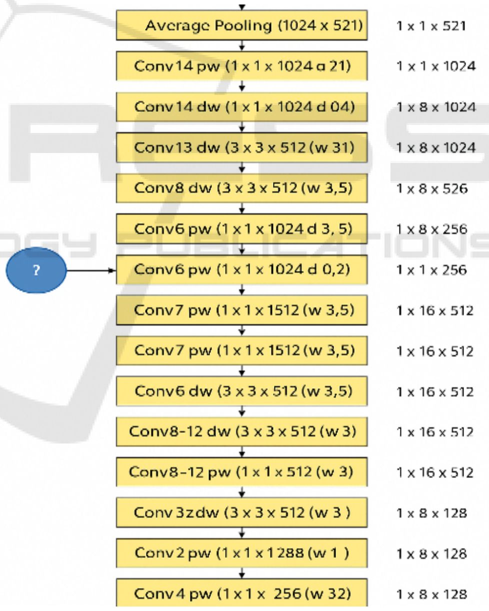
Average Pooling (1024 x 521)
1 x 1 x 521
Conv14 pw (1 x 1 x 1024 a 21)
1 x 1 x 1024
Conv14 dw (1 x 1 x 1024 d 04)
1 x 8 x 1024
Conv13 dw (3 x 3 x 512 (w 31)
1 x 8 x 1024
Conv8 dw (3 x 3 x 512 (w 3,5)
1 x 8 x 526
Conv6 pw (1 x 1 x 1024 d 3, 5)
1 x 8 x 256
?
Conv6 pw (1 x 1 x 1024 d 0,2)
1 x 1 x 256
Conv7 pw (1 x 1 x 1512 (w 3,5)
1 x 16 x 512
Conv7 pw (1 x 1 x 1512 (w 3,5)
1 x 16 x 512
Conv6 dw (3 x 3 x 512 (w 3,5)
1 x 16 x 512
Conv8-12 dw (3 x 3 x 512 (w 3)
1 x 16 x 512
Conv8-12 pw (1 x 1 x 512 (w 3)
1 x 16 x 512
Conv 3zdw (3 x 3 x 512 (w 3)
1 x 8 x 128
Conv2 pw (1 x 1 x 1288 (w 1)
1 x 8 x 128
Conv4 pw (1 x 1 x 256 (w 32)
1 x 8 x 128

Figure 2: YAMNet body architecture.

5.3 Whisper Speech Recognizer

The Whisper Speech Recognizer module employs OpenAI’s Whisper model, a powerful speech-to-text (STT) transcription system. This module extracts spoken words from audio files, ensuring that the generated images reflect any verbal context present in the input in Figure 3.

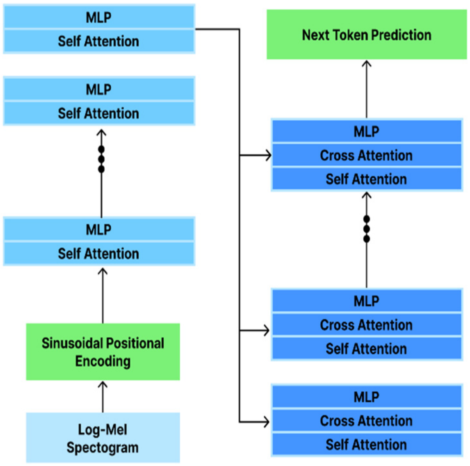

Figure 3: Whisper model architecture overview.

Key features of the Whisper Speech Recognizer include:

- High-Accuracy Transcription: The model supports multiple languages and delivers state-of-the-art transcription accuracy, even in noisy environments.
- Confidence-Based Validation: The module strips speculative transcriptions from confidence scores to prevent deceptive textual information.
- FP16 Acceleration: On running on a GPU, the model operates in float16 (FP16) mode, which is the fastest mode while maintaining quality.
- Content Sanitization: The system applies minimal text sanitization, stripping away unwanted words and semantically significant phrases that help generate images.

This module plays a critical role in integrating words spoken into the image generation process, thus adding to the overall accuracy and realism of the output.

5.4 Contextual Prompt Generator

Contextual Prompt Generator is a prominent part that transforms the extracted speech transcriptions and sound events into formal prompts to be used in image generation. The module maintains high relevance of textual prompts being delivered to the Stable Diffusion model, ensuring proper alignment of the audio output and image.

The logic sequence for this module entails:

- Merging Environmental and Speech Data: It blends labeled environmental sounds (e.g., "rainfall," "birds chirping") with transcribed speech (e.g., "It's a beautiful day").
- Filtering Non-Essential Labels: Background noise and irrelevant sounds (e.g., "mic static") are removed.
- Structuring the Prompt: The system uses a structured format, for example:
- "A realistic scene with: [Environmental sounds]. [Speech context]."
- Enhancements & Style Descriptors: It adds style elements to the prompt, such as:

Cinematic lighting
High detail textures

Through the creation of significant prompts, this module fills the gap between sound analysis and image generation so that the resulting images reflect the input sounds accurately.

5.5 Stable Diffusion XL

Stable Diffusion XL module is used to produce images from well-structured prompts. It uses the diffusers library, an API for executing state-of-the-art text-to-image models in Figure 4. This module is fine-tuned for maximum performance and accuracy, generating images that are visually engaging.

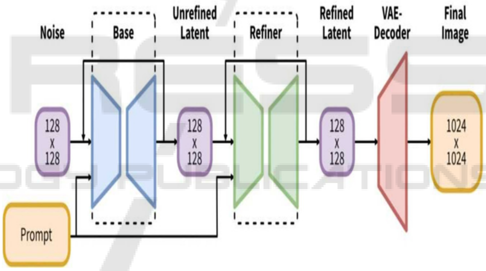

Figure 4: Stable Diffusion XL Pipelining Model.

Important parameters utilized in this module are:

- Resolution: Made optimum for widescreen format.
- Inference Steps: (25 steps) – controlling quality and computational overhead.
- Guidance Scale: (8.0) – providing high adherence to prompts.
- Negative Prompts: Eliminates unwanted characteristics (e.g., "disfigured, blurry, low-quality").
- Optimization: Takes advantage of T4 GPU acceleration, torch. float16, and xformers in order to speed up image generation.

By taking advantage of state-of-the-art AI-based image synthesis, this module guarantees that the output is visually precise, aesthetically pleasing, and contextually relevant to the input audio.

6 FORMULAE

- Mel-frequency cepstral coefficients (MFCCs) used in audio processing: It is extensively employed for processing speech and audio signals in areas like speech recognition, speaker identification, and emotion recognition. MFCCs extract the fundamental characteristics of an audio signal by encoding its spectral attributes in a manner that replicates human auditory perception. Shown below in (1)

$$MFCC = DCT(log(Mel(FFT(signal)\ |^{2}))) \qquad (1)$$

- Classifier Confidence Calculation: Classifier confidence refers to a degree of how sure a machine learning algorithm is in its prediction. It is widely applied in classification problems to determine the confidence of an output label. Shown below in (2)

$$H(p,q) = -\sum xp(x)logq(x) \qquad (2)$$

- Cross-entropy loss used in speech recognition: Cross-entropy loss is a popular loss function in classification tasks, such as speech recognition, where one wishes to reduce the difference between probability distributions that have been predicted and actual ones. Shown in below in (3)

$$confidence = (\sum i = 1.ci)/n \qquad (3)$$

7 EVALUATION METRICS

In order to measure the efficiency of the Audio-to-Image pipeline, we use CLIP Score as the most basic evaluation metric. CLIP (Contrastive Language-Image Pretraining) is an OpenAI multimodal model that measures how well an image matches a specific textual description. Given that our pipeline produces images based on structured prompts based on the audio input, CLIP Score can be used as a consistent metric to measure the semantic similarity between the produced images and the source audio context.

The CLIP score is used to assess the Audio-to-Image pipeline, which captures the semantic consistency between the created images and the respective audio-generated prompts. Comparative evaluation of the speech and environment audio samples allows for understanding of the quality with which the system understands and depicts various forms of auditory inputs.

For speech samples in Figure 5, the CLIP scores range depending on the level of complexity in the verbal content. The best score (31.47) was seen for the most basic speech input—\"Don't be rude\". This indicates that short and to-the-point speech inputs produce improved semantic alignment in image creation. On the other hand, longer and more contextually complex inputs like the United Nations Speech (19.69) and Inaugural Presidential Speech (23.02) had relatively lower CLIP scores. This shows that as speech complexity is higher, creating a highly aligned visual representation becomes more difficult. The confidence in transcription was stable at 0.70 for speech samples in Table.2, reflecting the robustness of the speech recognition model. The sound classification confidence also varied, with a lower mean for structured speech, reflecting the model's difficulty in classifying spoken words in wider environmental contexts.

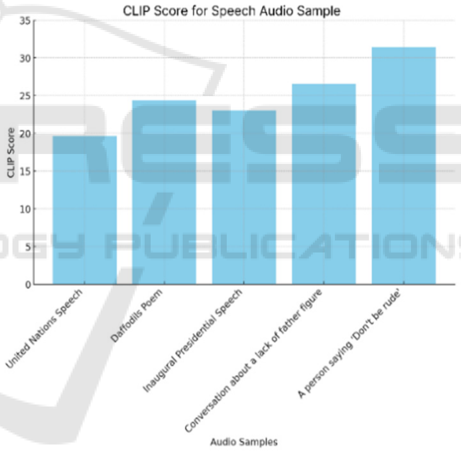

Figure 5: CLIP score for speech audio sample.

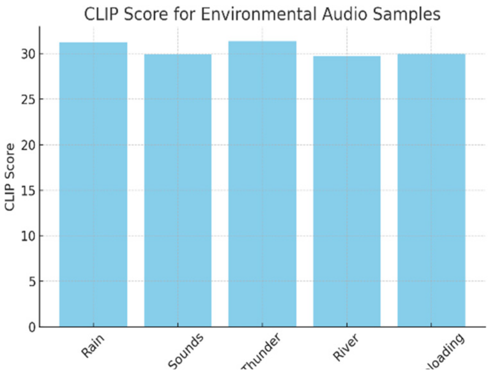

Figure 6: CLIP score for environmental audio sample.

Table 2: Metric overview for speech audio sample.

Sample	Image Generated Time (s)	Avg. Sound Confidence	Transcription Confidence	Transcription Valid	CLIP Score	Success
United Nation Speech	41.32	0.63	0.7	TRUE	19.69	TRUE
Daffodils Poem	45.66	0.88	0.7	TRUE	24.39	TRUE
Inaugural Presidential Speech	44.79	0.47	0.7	TRUE	23.02	TRUE
Conversation about lack of father figure	44.91	0.43	0.7	TRUE	26.52	TRUE
"Don't be rude"	45.72	0	0.7	TRUE	31.47	TRUE

For environmental audio samples in Figure 6, the CLIP scores had a relatively narrower range, ranging from 29.75 (River) to 31.39 (Thunder). This indicates natural sounds tend to produce high-quality alignment with audio descriptions and synthesized images. In contrast to speech samples, the transcription model usually considered environmental audio inputs invalid because they contained no identifiable words. The sound classification confidence in Table. 3 tended to be lower, with the lowest confidence from Gun Reloading (0.16) and Rain (0.29), which have subtle and changing acoustic properties. But the system was still able to produce high CLIP scores, suggesting that environmental sounds naturally offer more distinct contextual signals for image creation than spoken language.

Table 3: Metric overview for environmental audio sample.

Sample	Image Generated Time (s)	Avg. Sound Confidence	Transcription Confidence	Transcription Valid	CLIP Score	Success
Rain	39.66	0.29	0.7	FALSE	31.22	TRUE
Cat Sound	38.41	0.7	0.7	FALSE	29.94	TRUE
Thunder	42.27	0.37	0.8	FALSE	31.39	TRUE
River	43.44	0.22	0.7	FALSE	29.75	TRUE
Gun Reloading	43.22	0.16	0.8	FALSE	29.97	TRUE

In comparison between the two types, environmental audio samples always achieved greater CLIP scores than speech samples. This implies that natural soundscapes present more distinctive auditory cues that convert well into visual forms this can be seen in Figure 7. Speech inputs, particularly those containing abstract or multifaceted meanings, seem harder for the system to accurately understand. This gap identifies areas where things can be made better, such as enhancing prompt engineering methods for spoken materials and multimodal alignment methods to better enhance visual coherence.

In total, the assessment illustrates that the Audio-to-Image pipeline successfully transforms both speech and sounds from the environment to semantic images. But additional enhancements in the structuring of the prompt, the accuracy of speech recognition, and the confidence in sound classification can further amplify the overall semantic congruence, particularly for speech inputs. These findings open up doors for future enhancement in AI-based multimodal applications, enhancing the cross-modal connection between auditory and visual modalities.

In total, the assessment illustrates that the Audio-to-Image pipeline successfully transforms both speech and sounds from the environment to semantic images. But additional enhancements in the structuring of the prompt, the accuracy of speech recognition, and the confidence in sound classification can further amplify the overall semantic congruence, particularly for speech inputs. These findings open up doors for future enhancement in AI-based multimodal applications,

enhancing the cross-modal connection between auditory and visual modalities.

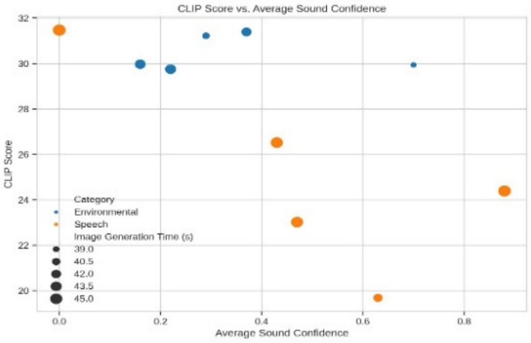

Figure 7: CLIP score vs average sound confidence.

8 MODULE OUTPUT

The model output is obtained by training a model and passing an audio file which Refer Figure 8 and Figure 9 for audio to image generation via both environmental and figure 10 and figure 11 audio to image generation via both speech audio.

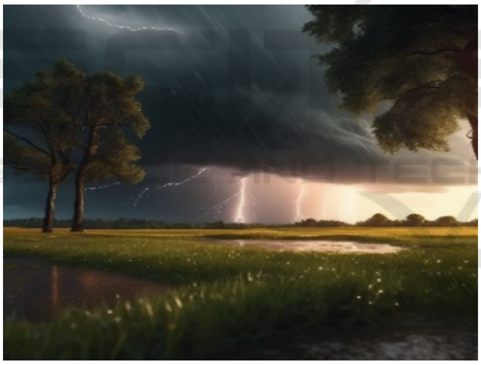

Figure 8: Environmental audio rain.

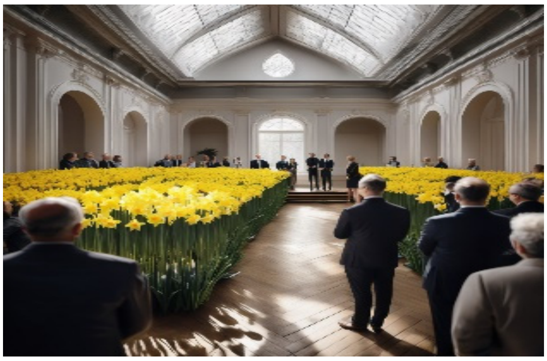

Figure 9: Environmental audio river.

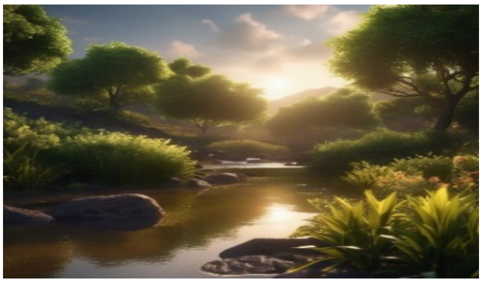

Figure 10: Speech audio of Daffodils poem.

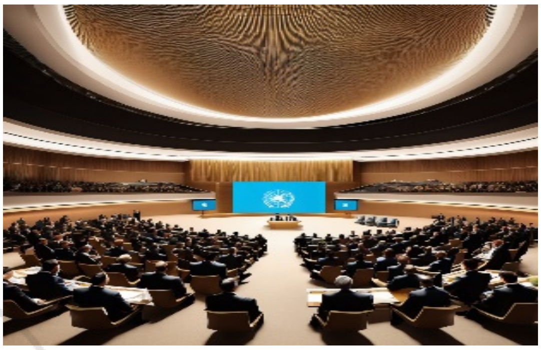

Figure 11: Speech at U

9 CONCLUSIONS

The proposed audio-to-image generation pipeline represents the next big leap in multimodal AI, providing a novel means of translating sound into high-resolution visual representations. Integrating the latest in-state-of-the-art technologies like Whisper, YAMNet, and Stable Diffusion, this system overcomes traditional methodologies' weaknesses by ensuring scalability, realism, and flexibility across diverse audio inputs. The extraction of meaningful audio features and accurate mapping to visual elements enhance the quality and precision of generated images, making this approach highly applicable across various domains.

Together with the real-time processing and iterative refinement mechanisms, the system further enhances its features and promotes dynamic user interaction and customization. In contrast to static models, the framework enables continuous improvement through user feedback, ensuring that the produced visuals are relevant and contextually accurate for the intended objects. This kind of mechanism provides exceptional benefits for auditory-to-visual translation applications when precise translation is critical; this includes assistive technologies, environmental analysis, and creative multimedia production.

This technology transcends artistic and entertainment landscapes. In scientific research, environmental monitoring, and education, transforming sound into meaningful visual representations truly opens up possibilities for analysis and interpretation. Bridging the gap between audio perception and visual synthesis, this work belongs to the class of evolving cross-modal AI with new avenues opened up for interdisciplinary innovation.

The future work shall involve diversification of the datasets, optimizing the computational efficiency of the algorithm, and further fine-tuning to be able to interpret more complex soundscapes. The continuous progress in the areas of AI and deep learning shall help redesign how auditory information can be represented and shall play a very versatile tool in diverse applications.

REFERENCES

Aihua Zheng, Menglan Hu, Bo Jiang, Yan Huang, Yan Yan, Bin Luo, “Adversarial-Metric Learning for Audio-Visual Cross-Modal Matching” [IEEE Transactions on Multimedia (Volume: 24)].

Amma Liesvarastranta Haz, Evianita Dewi Fajrianti, Nobuo Funabiki, Sritrusta Sukaridhoto, “A Study of Audio-to-Text Conversion Software Using Whispers Model” [2023 Sixth International Conference on Vocational Education and Electrical Engineering (ICVEE)].

C.H. Wan, S.P. Chuang, and H.Y. Lee, “Towards audio to scene image synthesis using generative adversarial network,” in Proceedings of the [IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 496–500.]

G.Tzanetakis and P.Cook,“ Musical genre classification of audio signals,” [IEEE Trans. Speech Audio Process., vol. 10, no. 5, pp. 293–302, Jul. 2002.]

I. J. Goodfellow et al., “Generative adversarial networks,” in Proc. Adv. Neural Inf. Process. [Syst., 2014, pp. 2672–2680.]

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, “Audio Set: An ontology and human-labeled dataset for audio events” [2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)].

L. Chen, S. Srivastava, Z. Duan, and C. Xu, “Deep cross-modal audio-visual generation,” [in Proceedings of the on Thematic Workshops of ACM Multimedia, 2017, pp. 349–357.]

M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, “Putting the face to the voice’: [Matching identity across modality,” Biol., vol. 13, no. 19, pp. 1709–1714, 2003.]

Mahmoud Abdulsalam, Nabil Aouf, “Using Stable Diffusion with Python: Leverage Python to control and automate high-quality AI image generation using Stable Diffusion.”

N. Harihara Valliappan, Sagar Dhanraj Pande, Surendra Reddy Vinta, “Enhancing Gun Detection with Transfer Learning and YAMNet Audio Classification” [IEEE Access (Volume: 12)].

Pei-Tse Yang, Feng-Guang Su, Yu-Chiang Frank Wang, “Diverse Audio-to-Image Generation via Semantics and Feature Consistency” [2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)].Hazimah Widyagustin, Hendra Kusuma, Tri Arief Sardjono, “Deep Learning-Based Classification of Lung Sound Using YAMNet for Identifying Lung Diseases” [2024 2nd International Symposium on Information Technology and Digital Innovation (ISITDI)].

S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” [IEEE Trans. Multimedia, vol. 2, no. 3, pp. 141–151, Sep. 2002.]

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” [in Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 1060–1069.]

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” [in Proceedings of the 33rd International Conference on Machine Learning, 2016]

Sheng-Bin Hsu, Chang-Hsing Lee, Pei-Chun Chang, Chin-Chuan Han, Kuo-Chin Fan, “Local Wavelet Acoustic Pattern: A Novel Time–Frequency Descriptor for Birdsong Recognition” [IEEE Transactions on Multimedia (Volume: 20, Issue: 12, December 2018)].

T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo, “Combining image captions and visual analysis for image concept classification,” [in Proc. Int. Workshop Multimedia Data Mining: Held Conjunction ACM SIGKDD, 2008, pp. 8–17.]

Theresia Herlina Rochadiani, Yulyani Arifin, Derwin Suhartono, Widodo Budiharto, “Exploring Transfer Learning Approach for Environmental Sound Classification: A Comparative Analysis” [2024 International Conference on Smart Computing, IoT and Machine Learning (SIML)].

X. Alameda-Pineda, V. Khalidov, R. Horaud, and F. Forbes, “Finding audio-visual events in informal social gatherings,” [in Proc. Int. Conf. Multimodal Interfaces, 2011, pp. 247–254.]

Xixuan Wu, Yu Qiao, Xiaogang Wang, Xiaoou Tang, “Bridging Music and Image via Cross-Modal Ranking Analysis” [IEEE Transactions on Multimedia (Volume: 18, Issue: 7, July 2016)].