CIS 6200 Spring 2024, Project Report

AudioVisGen: Video Generation using Audio Github Repository

Team Members:

- Erica Wu; Email: `wuerica@seas.upenn.edu`
- Kaiwen Hu; Email: `kaiwenh4@seas.upenn.edu`
- Nidhi Bali; Email: `nbali@seas.upenn.edu`
- Vyaas Valsaraj; Email: `vyaas@seas.upenn.edu`

Abstract

This project explores the development of an audio-to-video generation model that integrates Contrastive Audio-Visual Pretraining (CAVP) with existing frameworks to enhance video generation from direct audio inputs. Motivated by the need to improve semantic coherence and temporal synchronization in generated videos, our approach leverages this integration to address the challenges of translating complex audio cues into visually coherent outputs. While the model showed improvements in aligning audio with video content, significant discrepancies were observed between quantitative metrics and qualitative assessments. These discrepancies highlight not only the limitations of current evaluation metrics but also underscore the inherently subjective nature of art and visual content, which can vary widely in individual perception. This project points to the necessity for developing more comprehensive evaluation metrics that better align with human perception and acknowledge the subjective experience of art. The findings lay the groundwork for future research in enhancing audio-to-video synthesis models, emphasizing the need for larger datasets, expanded computational resources, and new evaluation methods to truly gauge the effectiveness and appeal of generated video content.

1 Motivation

Humans have a remarkable ability to visualize scenes from auditory cues, an intuitive process that remains challenging to replicate in machines. Our project builds on the foundation laid by TempoTokens, which transforms audio tokens into visual sequences via an intermediary text representation. By integrating the Contrastive Audio-Visual Pretraining (CAVP) strategy from the Diff-foley[2] project, we aim to enhance the fidelity and contextual alignment between the audio inputs and the generated videos.

This improvement could revolutionize the way we experience multimedia, making it more immersive and accessible. For example, it could enable the dynamic generation of visual content from live music in real-time, or help produce educational tools that offer visual interpretations of complex audio. Moreover, enhancing audio-to-video synthesis could significantly improve accessibility for individuals with hearing impairments, providing them with new ways to experience audio content visually.

Our enhancements are designed to address the limitations of current models by improving the synchronization and semantic coherence of the audio-visual output, paving the way for more expressive and context-aware multimedia experiences.

2 Related Work

Audio-to-video generation is a rapidly advancing field that seeks to bridge the sensory gap between hearing and seeing. One notable project in this domain is TempoToken [6], which innovatively transforms audio tokens into text tokens before utilizing a pre-trained text-to-video generation model to create visual content.

While TempoToken demonstrates promising results in generating coherent video sequences, its dependency on text as an intermediary format might limit the direct translation of audio nuances into visual elements.

Another significant contribution is made by Diff-foley [2], which employs contrastive learning to map audio samples and corresponding videos to a shared embedding space, enhancing the model's ability to synchronize and align audio-visual content. This method shows considerable improvements in the semantic alignment of the generated videos with the audio input. However, the approach requires high-quality, pre-aligned audio-video pairs for training, which can be a limitation in scenarios where such data is scarce or of inconsistent quality.

Our work aims to synthesize the strengths of these approaches while addressing their limitations. By integrating the Contrastive Audio-Visual Pretraining (CAVP) strategy from Diff-foley with the audio-to-text-to-video framework of TempoToken, we propose a hybrid model that benefits from the robust feature extraction of contrastive learning and the expansive capabilities of text-to-video synthesis. This integration seeks to enhance the direct audio-to-visual translation process, potentially leading to more accurate and dynamically generated video content from diverse audio inputs.

3 Problem Formulation

The primary challenge in audio-to-video generation lies in accurately translating complex audio cues into corresponding visual scenes without direct human intervention. Traditional methods often rely on intermediate representations, such as text, which can introduce discrepancies due to the abstraction and interpretation differences between audio and visual modalities. This can result in videos that lack contextual alignment with the original audio input or fail to capture subtle audio nuances.

Objective: Our project aims to develop a model that directly translates audio inputs into visually coherent and contextually relevant video sequences. This involves overcoming the limitations of dependency on textual intermediaries and improving the direct mapping from audio features to visual outputs.

Specific Challenges:

1. **Semantic Coherence:** Ensuring that the generated video not only matches the general theme of the audio but also aligns closely with specific cues and subtleties present in the sound.

2. **Temporal Synchronization:** Maintaining accurate alignment between the timing of audio events and the corresponding visual changes. This is crucial for the authenticity and immersion of the generated video content.

3. **Data Diversity:** Handling a wide range of audio types, from simple spoken words to complex musical compositions, each requiring different approaches for effective visual translation.

Proposed Solution: We propose a hybrid model that integrates the Contrastive Audio-Visual Pretraining (CAVP) strategy from Diff-foley with the structured transformation process of TempoToken. By leveraging contrastive learning, our model aims to refine the feature extraction process, enabling a more nuanced understanding of the audio content that can be directly translated into visual elements.

By addressing these challenges, our model seeks to push the boundaries of current audio-to-video synthesis technologies, paving the way for more dynamic and expressive multimedia applications.

4 Methods

In this section, we describe the problem for the existing dataset and propose a new customer dataset to solve the problem. Then, we introduce the end-to-end audio-to-video model structure.

4.1 Dataset

1. VGG Dataset: VGG-Sound is an audio-visual correspondent dataset consisting of short clips of audio sounds, extracted from videos uploaded to YouTube. It consists of 310+ classes and 2,00,000+ videos

with 550+ hours of data. We chose a subset of 900 videos where ”playing volleyball” was mentioned in the captions of the VGG-Sound Dataset. Upon experimentation, we found that the video quality of this subset was very poor due to a lot of background noise with respect to audio and visual content. Hence, we decided to create our own dataset to work with.

2. Customized Dataset: We chose appropriate youtube videos and trimmed them into smaller clips to train our model with better and more reliable data. First, we trimmed a youtube video of ”Rain and Thunder” class into ten second clips. This is a continuous class and does not require a hgh audio-visual alignment. As our next step, we decided to go for a more discrete class with respect to both the visual and the audio elements, so, we chose another youtube video of dogs barking for the ”Dogs” class and trimmed them into two second clips. We obtained 400 training and 100 testing clips for each class.

4.2 Architecture

The overarching model pipeline is depicted in Figure 1. In contrast to the BEATs feature extractor utilized in prior work [1], our approach integrates a CAVP encoder sourced from Diff-Foley [2]. This substitution is predicated on the assumption that the CAVP effectively projects audio into a latent space shared by both audio and video, facilitating the transformation of requisite information for video generation. Consequently, we anticipate that this modification will lead to improved convergence and generalization during the training process.

To maintain continuity in the generated videos, we implement a moving average technique with varying window sizes (1, 2, 4, 8). This approach enables the construction of more comprehensive global information across frames, thereby ensuring consistency in the overall video theme. Additionally, we incorporate a self-attention layer after the MLP layers to further intermix frame information before it reaches the diffusion model. This self-attention mechanism enables the extraction of crucial ’class’ information, enhancing the model’s ability to generate coherent and contextually relevant video sequences.

We leverage a pre-trained text-to-video model [3], which enables us to input optional text prompts during both the training and inference stages. Incorporating text prompts serves the purpose of providing additional guidance to the model, aiding it in identifying the appropriate video distribution for the corresponding class. This approach is hypothesized to alleviate the lack of context issue encountered during generation conditioned solely on audio inputs.

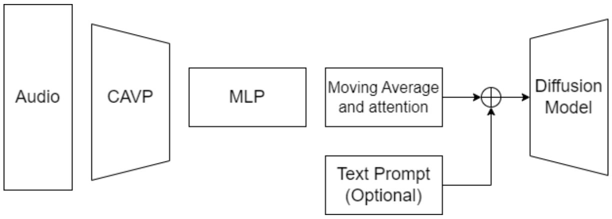


Figure 1: Model Structure

5 Experimentation

We are using the pre-trained CAVP embedding model[1] and the pre-trained text-to-video model[2]. Throughout the training process, we maintained the CAVP embedding model’s parameters frozen while employing the AdamW optimizer. The MLP layers were initialized using the default PyTorch initialization scheme. Initially, we attempted to train the diffusion model concurrently with the MLP layers, but the resultant model exhibited poorer performance compared to the model trained solely with MLP layers. Consequently, for the remainder of the experiment, we opted to freeze the diffusion model during training. We are using the hyperparameters in the table 1 during the training and inference stage.

[1] https://huggingface.co/SimianLuo/Diff-Foley

[2] https://huggingface.co/cerspense/zeroscope$_v2_576w$

Table 1: Model Parameters

Parameter	Value
learning rate	1e-4, 1e-5, 1e-6
Adam Weight Decay	0.01
Train Batch Size	1
Seed	64
epoch	5-20
Inference Steps	25
Guidance Scale	25

5.1 Audio Input

1. **Single Class: Volleyball**
   We utilized the volleyball subset outlined in 4.1 to train our pipeline over 10 epochs. Subsequently, our model was evaluated on volleyball video clips not included in the training set, yielding results akin to those depicted in Figure 2a. Disappointingly, the results fell far short of our expectations, with amalgamated sports scenes and unprofessional player portrayals. Upon closer examination of the dataset, it became evident that the poor quality of the video clips hindered our ability to discern volleyball gameplay through audio alone. Consequently, the model struggled to effectively learn the nuances of authentic volleyball scenes within this dataset, prompting us to curate a higher-quality dataset.

2. **Single Class: Thunder sounds** We utilized the rain dataset outlined in 4.1 and trained our model over 15 epochs. The resulting output is illustrated in Figure 2b. Interestingly, while the generated videos accurately depict rainfall, they unexpectedly resemble waterfalls, a phenomenon not present in the training dataset. We hypothesize that the similarity in sound between waterfalls and rain posed a challenge for the model's identification process. Consequently, we aim to incorporate distinct and informative audio cues from the video to improve model performance.

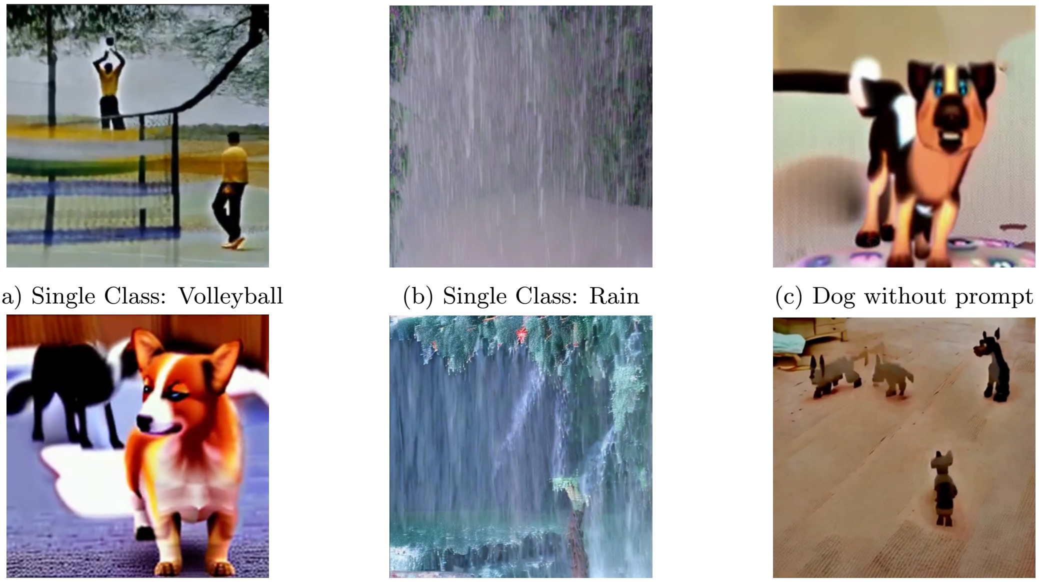
3. **Single Class: Dog** We trained our model using the dog dataset specified in 4.1 for 20 epochs. The resulting output is presented in Figure 2c. Notably, there are no instances of fusion observed this time, indicating that the model found it easier to learn from informative audio cues. This underscores the critical role of data quality in model training.

5.2 Audio + Text Input

1. **Single Class: Dog** While the dog dataset produces satisfactory results post-training, we noticed that all generated dogs appear blurred and fail to conform to the physical laws within the video. To tackle this challenge, we propose a solution in the subsequent section, utilizing text prompts as guiding cues during training. We hypothesize that integrating text prompts will aid the model in learning the appropriate distribution of dogs and reduce hallucinations involving other class objects. The outcomes of this approach are demonstrated in Figure 2d. Comparative analysis indicates that the model trained with text prompts produces more realistic dog representations compared to the model without such guidance, validating our assumption regarding the effectiveness of text prompts. However, it is worth noting that the model still generates scenes that defy physical laws, underscoring the difficulty in achieving real-world simulation.

2. **Multi Class: Dog and rain** After achieving success with single-class object generation, we ventured into multi-class training. In this experiment, we merged the dog dataset with the rain (with thunder) dataset, employing the same configuration utilized for the dog with a prompt model. The results for testing rain are illustrated in Figure 2e, while the corresponding dog representations are depicted in Figure 2f. As anticipated, the quality of the generated videos has diminished compared to the single-model approach. This outcome aligns with our expectations, as multi-class video generation demands

a larger dataset and increased computational resources to achieve results comparable to those of the single-class model.



(a) Single Class: Volleyball

(b) Single Class: Rain

(c) Dog without prompt

(d) Dog with prompt

(e) Multi-class: Rain

(f) Multi-class: Dog

Figure 2: Generated videos

6 Evaluation

Evaluating the quality and effectiveness of our audio-to-video generation model requires a comprehensive set of metrics that assess both the visual quality of the generated videos and their alignment with the audio cues. We have selected the following metrics for their relevance and robustness in measuring different aspects of video generation:

1. **Frechet Video Distance (FVD)**[4]: This metric computes the distance between the feature vectors of the generated video and a target video derived from a pre-trained model. A lower FVD indicates that the generated videos are closer to the target in terms of visual content and dynamics, suggesting higher quality and realism.

2. **CLIP Similarity Metric**[5]: By averaging the CLIP similarity scores of each frame with the corresponding text input, this metric evaluates how well the visual content matches the expected content derived from text labels of the original videos. Higher scores indicate better content relevance.

3. **Audio-Video Alignment**[6]: Specifically proposed by TempoTokens, this metric measures the synchronization between audio changes and corresponding visual changes in the video. Effective alignment is crucial for ensuring that the video not only looks realistic but also accurately reflects the audio cues.

Each metric addresses a specific aspect of the video generation challenge: FVD focuses on video realism and quality of the generation, CLIP Similarity targets the relevance of the visual content to the audio input (through the associated text label), and Audio-Video Alignment assesses the temporal coherence between the audio and video. Together, these metrics provide a framework for quantitative evaluation of our model's performance. Metrics for our experiments are reported below.

Table 2: Experimental Results

Experiment	FVD(↓)	CLIPSIM(↑)	AV-Align(↑)
	Audio Input Only		
Volleyball	2955	0.23	0.40
Thunder	3312	0.21	0.10
Dog Barking	3105	0.24	0.33
	Audio + Text Input		
Dog Barking	2844	0.24	0.35
	Audio Input, Multiclass Training		
Thunder	867	0.23	0.10
Dog Barking	-1.7e9	0.23	0.32
TempoTokens Baseline	923	0.47	0.35

7 Results and Discussion

We find that conditioning on text input in addition to audio input improves all experimental metrics, and qualitatively produces better-looking videos, as shown in Figure 2d as compared to Figure 2c. This significant effect from conditioning on some other information in addition to the input audio shows a resolution of ambiguity that can arise with only audio inputs. Intriguingly, we also find that multi-class training improves metrics for results from the thunder experiment compared to single class training. This shows that information learned from training on a certain class can also be leveraged in generating outputs from a different class.

Generally, experimental results from our new model show poorer performance across these metrics than the baseline, save for the AV-Align metric, which is improved over the baseline for the experiment using the Volleyball dataset. However, because the AV-Align metric does not account for semantics of what occurs in the video and only compares changes in the video, videos can be nonsensical to viewers but still achieve high AV-Align scores. We note the importance of considering a suite of different strategies, including qualitative evaluation, for determining performance of models for this task, among others.

We posit that results from our experiments arise from several factors, including constrained data quality (as discussed in 4.1), and constraints on resources available for training, including training time. With more resources available, more of the model could be trained, on a larger dataset, and embedding spaces could be better aligned between inputs to the diffusion model and CAVP embeddings. This would lead to higher quality outputs that are more similar to the original videos associated with the audio inputs.

The mixed success across different experiments provides valuable insights into the challenges of audio-to-video synthesis. It emphasizes the need for models that can more accurately interpret complex audio cues and convert them into visually coherent outputs. Future work should focus on enhancing model architecture to better handle diverse and multi-class datasets, improving alignment techniques, and developing more nuanced evaluation metrics that can better capture the subjective quality of generated videos.

8 Conclusion

This project aimed to develop a robust model capable of generating video content directly from audio inputs by integrating Contrastive Audio-Visual Pretraining (CAVP). Our findings indicate improvements in temporal synchronization for certain experiments but also highlight significant discrepancies between quantitative metrics and qualitative assessments.

8.1 Summary of Findings

The integration of CAVP improved audio-visual alignment, as evidenced by quantitative metrics. However, the observation that some videos perceived qualitatively as poor had decent scores suggests that these metrics may not fully capture user-perceived video quality. Generally, integration of CAVP worsened quantitative performance, which we attribute to constraints on training compute and data.

8.2 Qualitative vs. Quantitative Evaluation

The divergence between qualitative observations and quantitative results underscores the need for developing more comprehensive evaluation metrics that better align with human perception and the subjective experience of video content.

8.3 Lessons Learned

- **Metric Development:** There is a clear need for new or refined metrics that more accurately reflect the qualitative aspects of generated videos, potentially incorporating user studies or perceptual tests.
- **Importance of Data Quality:** Our results reinforced the necessity for high-quality, well-curated data to train models effectively.
- **Impact of Compute Resources:** Computational constraints highlighted the need for adequate resources to fully leverage model capabilities and explore new methodologies.

8.4 Future Work

- Investigating the scalability of our model with larger and more varied datasets could provide insights into its generalizability and robustness.
- Expanding computational resources to allow for architectural innovations and more comprehensive training regimes could further enhance model performance.
- Exploring real-time processing capabilities could open new avenues for live audio-to-video applications, enhancing user interaction and accessibility.

Overall, while the project achieved significant advancements in audio-to-video synthesis, the findings also highlight crucial areas for improvement and future research opportunities, particularly in developing more nuanced evaluation strategies and enhancing model adaptability.

References

[1] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers, 2022.

[2] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models, 2023.

[3] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[4] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018.

[5] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. GODIVA: generating open-domain videos from natural descriptions. *CoRR*, abs/2104.14806, 2021.

[6] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation, 2023.