CIS 6200 Spring 2024, Project Report

AudioVisGen: Video Generation using Audio

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

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[3] 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 [9], 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 [3], 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. This approach aims to eliminate the need for text as an intermediary, thereby streamlining the translation process and potentially increasing the accuracy and relevance of the generated videos.

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

Instead of utilizing the BEATs feature extractor as referenced in [1], our approach employs a CAVP encoder sourced from Diff-Foley [3]. This substitution is based on the assumption that the CAVP embedding encapsulates more comprehensive audio information, thereby providing the video generation model with a richer understanding of the desired video output.

Following the replacement of the embedding model, we proceed to train the model to evaluate potential improvements compared to our baseline. This step is crucial as merely altering the embedding method without subsequent training would disrupt the learned audio-to-text transformation layers.

Subsequently, we construct a model integrating CAVP with a video generation model. We hypothesize that bypassing the text-embedding stage may enhance the flow of information to the generation model, thereby potentially improving its performance.

We propose three designs for our video diffusion model.

1. **Reverse Diff-Foley**: We would basically swap the way the diffusion model accepts the inputs and the denoising objective would be for the generation of video frames instead of a spectrogram. This is a very novel idea and we would have to start training such a diffusion model from scratch.

2. **Finetuning Power of Sound Diffusion Model[2]**: The Power of Sound incorporates stable diffusion to generate audio reactive videos and it would be beneficial to finetune it to generate videos from CAVP embeddings instead of training a diffusion model from scratch.

3. **Finetuning Seeing and Hearing Model[8]**: Similar to the previous proposed method, this is another model we can look to finetune and compare as this is a very recent paper that came out in February 2024.

Table 1: Result for the baseline

Model	FVD(↓)	CLIPSIM(↑)	IS(↑)	AV-Align(↑)
	VGGSound			
ModelScope Text2Vid	**801**	**0.69**	**15.55**	0.27
ModelScope Random	1023	0.47	6.32	0.26
Tempotoken	923	0.47	6.32	**0.35**

5 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)**[6]: 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**[7]: 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 the audio descriptions. Higher scores indicate better content relevance.
3. **Inception Score (IS)**[4]: Utilizing a pre-trained C3D model[5], the Inception Score assesses the clarity and diversity of the generated video frames. Higher IS values suggest that the video frames are both clear (high probability of predicting the correct class) and varied (diverse across the video).
4. **Audio-Video Alignment**[9]: Specifically proposed in Tempotoken, 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 and IS focus on the visual quality and diversity, CLIP Similarity targets the relevance of the visual content to the audio input, and Audio-Video Alignment assesses the temporal coherence between the audio and video. Together, these metrics provide a robust framework for evaluating our model's performance across different dimensions of quality and alignment.

6 Project Plan

Week 12: 04/15

- Construct the CAVP + video generator pipeline.
- Start training the pipeline.
- Optional: Train one more model.

Week 13: 04/22

- Evaluation on the pipeline.
- Write the report for the pipeline.

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] Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7822–7832, 2023.

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

[4] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[5] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[6] 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.

[7] 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.

[8] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. *arXiv preprint arXiv:2402.17723*, 2024.

[9] 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.