LAV: Audio-Driven Dynamic Visual Generation with Neural Compression and StyleGAN2

Jongmin Jung[1], **Dasaem Jeong**[2]

[1]Dept. of Artificial Intelligence, [2]Dept. of Art & Technology, Sogang University

Seoul, South Korea

jongmin@sogang.ac.kr, dasaemj@sogang.ac.kr

Abstract

This paper introduces LAV (Latent Audio-Visual), a system that integrates EnCodec's neural audio compression with StyleGAN2's generative capabilities to produce visually dynamic outputs driven by pre-recorded audio. Unlike previous works that rely on explicit feature mappings, LAV uses EnCodec embeddings as latent representations, directly transformed into StyleGAN2's style latent space via randomly initialized linear mapping. This approach preserves semantic richness in the transformation, enabling nuanced and semantically coherent audio-visual translations. The framework demonstrates the potential of using pre-trained audio compression models for artistic and computational applications.

Keywords

Audio-Visual Art, Neural Audio Compression, Generative Adversarial Networks, EnCodec, StyleGAN2

Introduction

Artistic and computational systems that translate audio into visual media have become increasingly sophisticated with the advent of neural networks. However, many current approaches are based on explicit feature extraction, which can lead to a loss of semantic richness in the translation process. This paper presents LAV (Latent Audio-Visual), a system that uses EnCodec's [2] compact and semantically rich latent representations as the foundation for audio-to-visual synthesis. By directly transforming these embeddings into StyleGAN2's [5] style latent space, LAV achieves nuanced outputs while preserving the semantic coherence of audio inputs.[1]

StyleGAN2 [5] uses a style-based generator architecture to enhance control and quality in image synthesis. At the core of its architecture is the style latent space W, an intermediate space derived from a mapping network that transforms input noise or latent vectors from the Gaussian input space Z. The style latent vector w is a specific vector in this space that modulates the weights of convolutional layers in the generator through adaptive instance normalization (AdaIN). This modulation enables precise control over various image features, from coarse attributes like pose to finer details like tex-

---

[1]LAV demo videos:
https://www.youtube.com/playlist?list=
PL3cldJxIjpAzkgpTWlLUxAElsT-5hCHh9

Figure 1: A sample frame produced by the proposed LAV (Latent Audio-Visual) system. ©StyleGAN2 model trained by Michael Friesen [3]
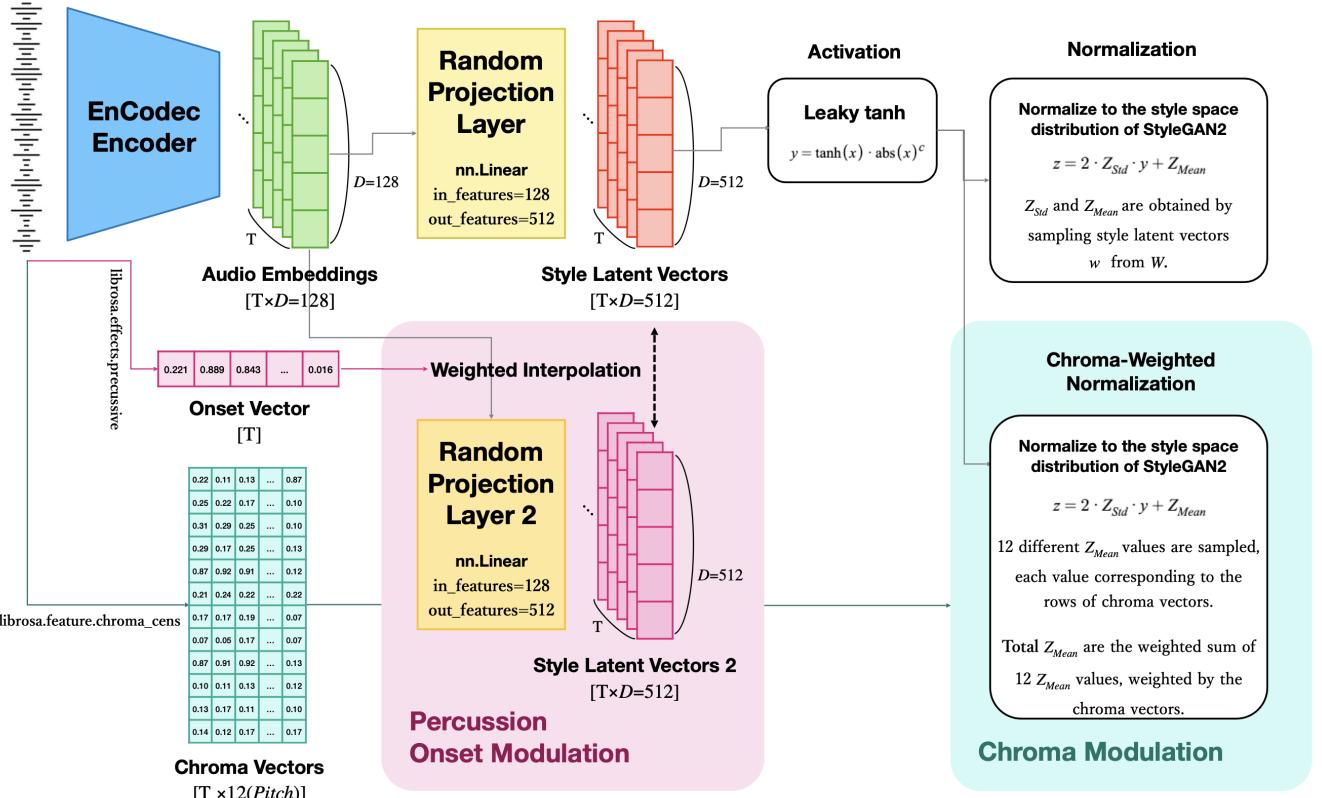
Figure 2: Overview of the proposed LAV (Latent Audio-Visual) pipeline.

ture, making W crucial for generating high-quality, semantically consistent outputs.

Recent advancements in audio-driven visual synthesis have begun addressing this challenge, particularly with the application of StyleGAN [6] for generative image synthesis. Notable contributions include *Audio-reactive Latent Interpolations with StyleGAN* [1], which modulates StyleGAN's latent space using explicitly extracted audio features such as chroma and rhythmic onsets. Although effective in generating visually dynamic output, this approach requires the integration of hand-crafted audio features, which may limit semantic richness.

TräumerAI: Dreaming Music with StyleGAN [4] takes a different approach by training a transformation layer to map audio features directly into StyleGAN2's style latent space. This system pairs music with corresponding visual styles, focusing on creating seamless transitions across musical sections. However, this method relies on task-specific training, which introduces additional complexity and dependency on annotated data for effective performance.

In contrast, LAV (Latent Audio-Visual) eliminates the need for additional training layers by introducing a randomly initialized projection layer. This projection layer transforms audio embeddings directly into StyleGAN2's style latent space via linear mapping. As proposed in *A Foundation Model for Music Informatics* [9], this approach not only simplifies the

pipeline by removing the train process, but also ensures the preservation of high-level semantic information. By leveraging audio embeddings retrieved with pre-trained EnCodec encoder, a neural audio compression model which has the latent embedding sample rate of 50Hz unlike other foundation models, the transformed embeddings inherently retain semantic richness and can fluently and swiftly respond to the immediacy of music, effectively capturing and representing its essence. inherently retain semantic richness. This method overcomes the limitations of previous studies while expanding the possibilities for audio-visual synthesis.

Methodology

EnCodec Embedding Utilization
EnCodec compresses audio into latent embeddings at a sample rate of 50Hz, with a latent dimension of $D = 128$. Like other robust deep-learning based audio encoders, EnCodec embeddings inherently capture semantic audio features, such as timbre and structure, without requiring explicit extraction of chroma or rhythm.

Latent Space Mapping
A randomly initialized single projection layer maps EnCodec embeddings to StyleGAN2's style latent space W of dimensions $D = 512$. As proposed in TräumerAI [4], the transformed embeddings are normalized to the distribution of W
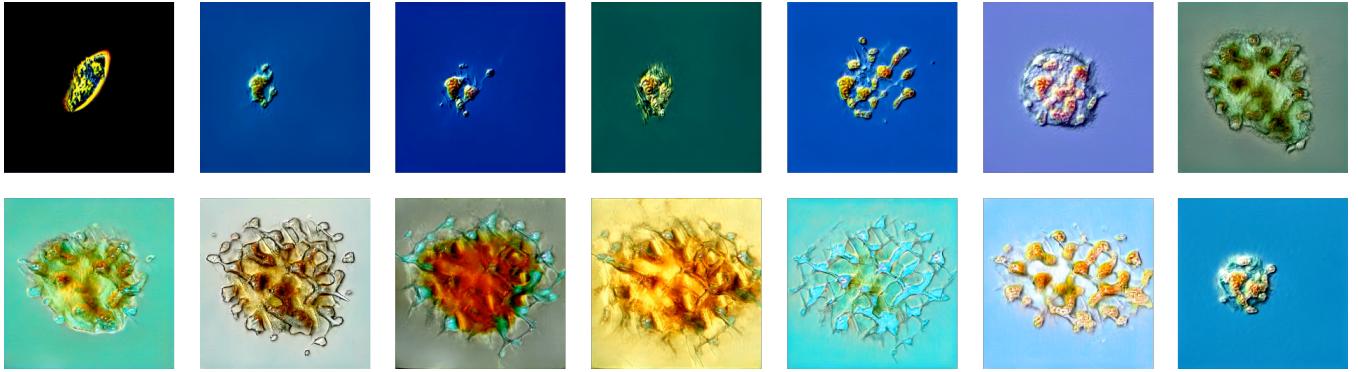
Figure 3: A sequence of frames produced by the proposed LAV (Latent Audio-Visual) system. ©StyleGAN2 model trained by Michael Friesen [3]

using Z_{Std} and Z_{Mean}, which are obtained by sampling style latent vectors w from W, with a controllable coefficient y allowing for further adjustment of the normalization process. This transformation leverages the semantic integrity of En-Codec embeddings, with a leaky tanh activation function with controllable coefficient c that preserves outliers and ensures compatibility with StyleGAN2's distribution.

Modulation by Musical Features

As proposed in *Audio-reactive Latent Interpolations with StyleGAN* [1], modulating with hand-crafted features adds even more dynamic to the output. Musical features such as chroma and percussive onsets are extracted from the input audio using *librosa* [8].

- Onset Modulation: Percussive onset vectors dynamically weight outputs from two projection layers, capturing rhythmic patterns.

- Chroma Modulation: Chroma vectors modulate the style latent space through weighted sums of 12 pitch-specific Z_{Mean} values, enabling harmonic content to influence visual aesthetics.

Hierarchical Style Smoothing

To ensure smooth transitions in the generated images and prevent abrupt changes, an averaging window is applied to the style latent vectors used in the synthesis process. As proposed in *TräumerAI* [4], StyleGAN2's synthesis network g progressively refines image details as convolutional layers are applied, transitioning from coarse to fine-grained features. At each convolutional layer, the AdaIN module applies a corresponding style vector derived from the latent vector w. These convolutional layers are grouped into coarse, middle, and fine hierarchies, and the averaging window size is tailored to each group.

Results and Demonstration

The LAV system successfully transforms audio inputs into visually dynamic and semantically coherent outputs. Using pre-trained EnCodec embeddings, the system demonstrates remarkable capabilities in audio-visual synthesis:

Visual Quality

- High-resolution image generation that maintains visual fidelity while reflecting audio characteristics

- Smooth transitions between frames through hierarchical style smoothing

- Stable and consistent image generation across varied inputs

Audio-Visual Mapping

- Seamless translation of diverse audio inputs into semantically rich visual outputs

- Comprehensive capture of audio semantics through EnCodec embeddings

- Consistent performance across various musical genres and styles

Conclusion

This work introduces LAV (Latent Audio-Visual), a novel framework that bridges neural audio compression and generative visual synthesis. By transforming EnCodec's semantically rich embeddings into StyleGAN2's style latent space through a randomly initialized linear mapping, the system achieves high-quality audio-visual translation without additional training requirements. The framework demonstrates the potential of using pre-trained audio compression models for artistic and computational applications, establishing a foundation for future developments in audio-visual synthesis, particularly in areas where semantic coherence between audio and visual elements is paramount.

Acknowledgement

References

[1] Brouwer, H. 2020. Audio-reactive latent interpolations with stylegan. In *Proceedings of the 4th Workshop on Machine Learning for Creativity and Design at NeurIPS 2020*.

Figure 4: A sample frame produced by the proposed LAV (Latent Audio-Visual) system. ©StyleGAN2 model trained by Krrrl [7]

[2] Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2023. High fidelity neural audio compression. *Transactions on Machine Learning Research*.

[3] Friesen, M. 2020. X.com.

[4] Jeong, D.; Doh, S.; and Kwon, T. 2020. Träumerai: Dreaming music with stylegan. In *Proceedings of the 4th Workshop on Machine Learning for Creativity and Design at NeurIPS 2020*.

[5] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*.

[6] Karras, T.; Laine, S.; and Aila, T. 2018. A style-based generator architecture for generative adversarial networks. *CoRR* abs/1812.04948.

[7] Krrrl. 2020. Runway ml – 3rd "model" (based on long poses).

[8] McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.

[9] Won, M.; Hung, Y.-N.; and Le, D. 2024. A foundation model for music informatics. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1226–1230.

Author Biography

Jongmin Jung is an AI researcher and creative technologist pursuing a master's degree in Artificial Intelligence at Sogang University. Having earned a Bachelor of Art & Science (BAS) in Art & Technology, he works at the intersection of AI and creative expression, with a focus on generative audiovisual art, music generation models, automatic music transcription (AMT), optical music recognition (OMR) and score-image to audio generation task. He actively collaborates with diverse artists, including musicians, tattooists, and visual creators, to integrate cutting-edge AI technology into artistic practices. Through his development of custom audiovisual pipelines and sound generation projects, Jongmin explores novel ways to merge machine learning with artistic innovation.