V2Meow: Meowing to the Visual Beat via Music Generation

Anonymous Author(s)
Affiliation
Address
`email`

1 Introduction

The majority of prior research in video-to-music generation has concentrated on designing complex rhythmic feature extractors to model the physical correspondence between video content and music. For instance, generating music synchronized with dance movements or reconstructing instrumental music based on changes in human poses in silent instrument performance videos (14; 15; 13; 11; 6). Consequently, these approaches are typically tailored to specific visual scenarios, and cannot be generalized to arbitrary video input types, e.g., vlogs or slideshows of images. In contrast, our study delves into the challenge of generating contextually relevant and high-quality background music for a broad spectrum of video input types. Importantly, we achieve this by conditioning the music generation solely on the visual information provided by video frames, without explicitly modeling domain-specific rhythmic or semantic relationships. We further hypothesize that with sufficient data and scale, the generation model is capable of learning the intrinsic video-music correspondence directly from easily accessible music videos and generating relevant background music for unseen video content types through zero-shot transfer.

We propose a video-to-music generation system called V2Meow that can generate high-quality music audio for a diverse range of video input types based on a multi-stage autoregressive model, without the need to explicitly model the rhythmic or semantic video-music correspondence. Compared to previous video to music generation work, the video and text prompts are incorporated as a single stream of embedding inputs and fed into the Transformer with feature-specific adaptors. Trained on O(100K) music audio clips paired with video frames mined from in-the-wild music videos, V2Meow is competitive with previous domain-specific models when evaluated in a zero-shot manner. V2Meow can synthesize high-fidelity music audio waveform solely by conditioning on pre-trained general-purpose visual features extracted from video frames, with optional style control via text prompts. Through both qualitative and quantitative evaluations, we verify that our model outperforms various existing music generation systems in terms of visual-audio correspondence and audio quality.

2 Method

Inspired by MusicLM (2), we take a multi-stage autoregressive language modeling approach (Figure 1) to condition music generation on video frames with optional high-level control over the style of the generated music through text prompts.

Feature Representations. For audio representations, we adopt the SoundStream tokenizer for acoustic tokens modeling and w2v-BERT tokenizer for semantic tokens modeling, both of which are pre-trained on the Free Music Archive (FMA) dataset (3). For all visual features, we use frame rate at 1 fps, following the standard on MV100K (1). For visual feature to music semantic tokens modeling, we use encoder-decoder Transformer and use 10-second random crops of the music video for visual to music semantic tokens modeling and semantic tokens to coarse acoustic tokens modeling. During inference, we take 10-second silent video as input and generate 10-second music clip.

Optional Style Control. To incorporate the control signal, we simply feed the MuLan audio embedding (8) as an additional input with sequence length be one to the Transformer encoder along with the visual features in the first stage. Both Mulan audio embedding and visual features are

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

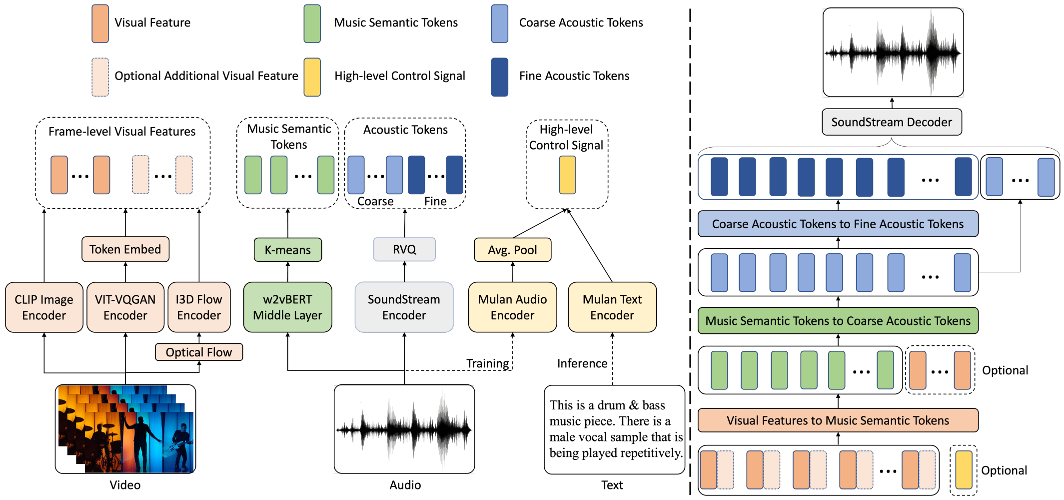

Figure 1: V2Meow architecture overview: (left) Feature extraction pipeline for video, audio and text representations. (right) Overview of multi-stage video to music modeling.

projected to the same feature dimension. At inference, we instead use the MuLan text embedding with the visual features to generate the semantic tokens.

Datasets. Following (12), we filtered a public available video dataset (1) to 110k videos with the label Music Videos and refer to it as MV100K. The training and validation datasets were split into an 80:20 ratio. We trained the Stage 1 model and Stage 2 model on these O(100K) music videos and refer to it as MV100K.

3 Evaluation and Results

We conduct a comprehensive evaluation of the music generation methods across three distinct datasets. To assess the quality of the generated music, we employ a multifaceted evaluation framework that encompasses both quantitative and qualitative metrics. Quantitative measures include evaluations of audio quality, rhythm synchronization, and text alignment. Additionally, we employ subjective metrics like visual relevance and music preference, which are best assessed through human evaluations.

Video conditional music generation. Since there is no open-source video to music in audio waveform, we compare our V2Meow model against the state-of-the-arts video-driven symbolic music representations-based model CMT (4) on the test partition of the MV100K. In terms of visual relevance and music quality, V2Meow significantly outperforms CMT by a large margin. For MV100K, we observe that adding visual input at the acoustic modeling stage significantly improves both audio quality related metrics. We further observe that the combination of CLIP and I3D Flow features yield best music generation quality overall.

Video and text conditional music generation. We compare V2Meow with text-to-music generation models like MUBERT (10) and Riffusion (5) on latest MusicCaps dataset (2), a subset of AudioSet (7) that contains about 5.5k human annotated text captions, music, and video pairs. With video frames as additional control, our approach outperforms Riffusion and MUBERT in visual relevance by 20-30%. It is worth to note that while our V2Meow model only use video-level MuLan embedding and trained on a O(100K) music videos, we still achieve better audio quality and text adherence than pure text-to-music generative model.

Dance to music generation. Evaluation on 20 dance videos in the test split of AIST++ (9) demonstrates that V2Meow can achieve comparable performances to domain-specific dance-to-music generation baselines (14; 15), as measured by beat coverage and beat hit. The evaluation is in zero-shot fashion without any fine-tuning on the AIST++ train split, and only video frames are used for modeling while no motion data is involved.

4 Ethical Implications

Controllable generative models such as V2Meow can serve as the foundation for new tools, technologies, and practices for content creators. While our motivation is to support creators to enrich their creative pursuits, we acknowledge that large generative models learn to imitate patterns and biases inherent in the training sets, and in our case, the model can propagate the potential biases built in the video and music corpora used to train our models.

Such biases can be hard to detect as they manifest in often subtle, unpredictable ways, which are not fully captured by our current evaluation benchmarks. Demeaning or other harmful language may be generated in model outputs, due to learned associations or by chance. A thorough analysis of our training dataset shows that the genre distribution is skewed towards a few genres, and within each genre, gender, age or ethical groups are not represented equally. For example, male is dominant in hip-hop and heavy metal genre. These concerns extend to learned visual-audio associations, which may lead to stereotypical associations between video content (i.e. people, body movements/dance styles, locations, objects) and a narrow set of musical genres; or to demeaning associations between choreography in video content and audio output (i.e. minstrelsy, parody, miming). ML fairness testing is required to understand the likelihood of these patterns in any given model and effectively intervene in them.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and C. Frank. Musiclm: Generating music from text. *ArXiv*, abs/2301.11325, 2023.

[3] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[4] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *Proc. of the ACM International Conference on Multimedia*, 2021.

[5] Seth* Forsgren and Hayk* Martiros. Riffusion - Stable diffusion for real-time music generation. 2022.

[6] Chuang Gan, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2020.

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[8] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. MuLan: A joint embedding of music audio and natural language. In *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[9] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021.

[10] Mubert-Inc. Mubert. https://mubert. com/, https://github.com/mubertai/ mubert-text-to-music. 2022.

[11] Kun Su, Xiulong Liu, and Eli Shlizerman. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems*, 33, 2020.

[12] Dídac Surís, Carl Vondrick, Bryan Russell, and Justin Salamon. It's time for artistic correspondence in music and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10564–10574, 2022.

[13] Jiashuo Yu, Yaohui Wang, Xinyuan Chen, Xiao Sun, and Y. Qiao. Long-term rhythmic video soundtracker. *ArXiv*, abs/2305.01319, 2023.

[14] Ye Zhu, Kyle Olszewski, Yuehua Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and S. Tulyakov. Quantized gan for complex music generation from dance videos. *ArXiv*, abs/2204.00604, 2022.

[15] Ye Zhu, Yuehua Wu, Kyle Olszewski, Jian Ren, S. Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal and conditional generation. *ArXiv*, abs/2206.07771, 2022.