Making Music for Videos

Pei-Huan Tsai
Columbia University
pt2630@columbia.edu

Yenchu Chen
Columbia University
yc4360@columbia.edu

Abstract

Nowadays, people love to incorporate music into their video contents. However, integrating appropriate music into video content poses significant challenges, often involving extensive searches through public music libraries or incurring substantial costs for creating original compositions.

In this regard, we developed a framework that combines state-of-the-art models to automatically generate a variety of high-quality, diverse, and engaging music through seamless integration. We also incorporates methods to repurpose these compositions and meld them with additional public music data, thereby enriching the framework's versatility.

We present several examples to illustrate our framework's capability to produce suitable music for diverse video types. Our experimental results include comparisons of music generated under various configurations, evaluated through both subjective metrics designed for this study and objective assessments.

1. Introduction

Music has become an indispensable part of people's lives in recent years. For instance, when people want to share snippets of life on social media, they often incorporate music into their video posts to enrich the atmosphere and emotions of their videos, making them more appealing to their audience.

Nevertheless, finding suitable music for videos is a labor-intensive process. Content creators frequently filter through existing music libraries, dedicating considerable time to assessing audio suitability for their videos. Often, these libraries may not have appropriate tracks, forcing creators to either compromise on the music quality or invest significant time in creating custom compositions. This situation underscores the necessity for a tool that can automatically generate appropriate audio content tailored to various video materials, thereby eliminating the need for manual selection from existing libraries.

Several works have been done in the music generation field. Di et al. [2] developed a transformer-based music generation method that utilizes three rhythmic relations between video and music to enhance control over the generated output. To facilitate the music generation process, Zhuo et al. [10] proposed a video-music generation framework with progressive decoupling control to further provide more control to the transformer. Additionally, Kang et al. [6] created a framework that uses video features as conditioning inputs to generate matching music via a Transformer architecture. Despite these advancements, these methodologies primarily focus on generating simple melodies that match the video's characteristics. These methodologies often lack the flexibility to adapt to the video's plot and scenes, resulting in music that does not fully complement the video content. For example, in the demo of the (https://amaai-lab.github.io/Video2Music/), although the music tempo adjusts to the video's speed, the compositions remain monotonous and fail to capture the semantic aspect of the video.

To address this challenge, we propose a framework that automatically generates audio corresponding to the content of non-audio videos by analyzing scenery. By leveraging state-of-the-art machine learning models, we can analyze visual elements and narrative flow, synthesizing background music that enhances the mood and context of the video. Additionally, we have developed a music retrieval flow, enabling the reuse of our generated music library with an improved semantic understanding of the compositions.

In summary, our main contributions are:

1. We develop a framework that can generate diverse and engaging music based on the context of videos.

2. We show that the proposed framework can generate different music themes based on the different scenes in the given video.

3. We propose a method to reuse the music generated by our flow given the generated music we have.

2. Method

Our project has two primary objectives: 1) Explore a viable approach for executing a comprehensive task of music

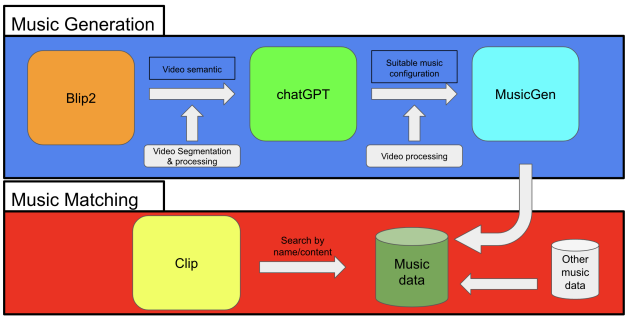


Figure 1. The music generation flow and music matching flow.

generation based on video content. 2) Develop a strategy for effectively reusing music after its initial generation.

In this regard, we developed a framework designed to generate music content that is well-suited to various types of videos. As illustrated in Fig. 1, the framework consists of two main components. The first is the music generation flow, which takes videos as input to produce music that corresponds to the video content. The second component involves selecting appropriate music for the video from our generated music database. The following sections detail this process.

2.1. Music Generation

1. In the first step, input data is processed by the Blip2 model [7]. This model captures essential descriptive information from the video and utilizes its Visual Question Answering (VQA) capabilities to extract specific details, including emotional content from each video clip. This process extracts the desired features for the user, serving as inputs for the subsequent music generation phase.

2. The extracted information is then propagated to the ChatGPT model (https://openai.com/blog/chatgpt), which enhances the video description by providing more concrete and detailed directives for music generation models. For instance, with information indicating a city by the sea, ChatGPT generates the following sentence:

   A serene piano melody intertwines with gentle waves, gradually building as the city's lights twinkle in the dusk, blending urban rhythms with the calming sounds of the sea.

   The resulting music encompasses not only the melodic piano tones suitable for urban scenes but also accurately simulates the sounds of the sea, making it highly compatible with the video content.

3. The enriched textual data is then forwarded into the musicgen-large [1] model, synthesizing music corresponding to the provided text. Our findings suggest that the specificity and clarity of musical instructions correlate with the model's ability to produce fitting and satisfying music compositions, as demonstrated in our experimental results.

2.2. Music Generation in Reality

A single video often comprises multiple scenes, requiring the selection of appropriate music for each scene transition. To support this feature, we utilized a Python video segmentation package PySceneDetect (https://www.scenedetect.com/) that segments the video into scenes. Following the segmentation process, we collect information such as the frame number and timestamp of each scene transition. This information is then utilized as input for both the MusicGen and Blip2 models, generating text descriptions for each theme and determine the corresponding music length for them. Finally, we concatenate these video segments using MoviePy (https://pypi.org/project/moviepy/) and reconstruct the video with synchronized audio.

2.3. Music Matching

For the music matching flow, we use the CLIP model [8] to match the corresponding text description for the music generation. Also, we downloaded several open music from the music library online to perform the music name matching to further enhance the diversity of our music library.

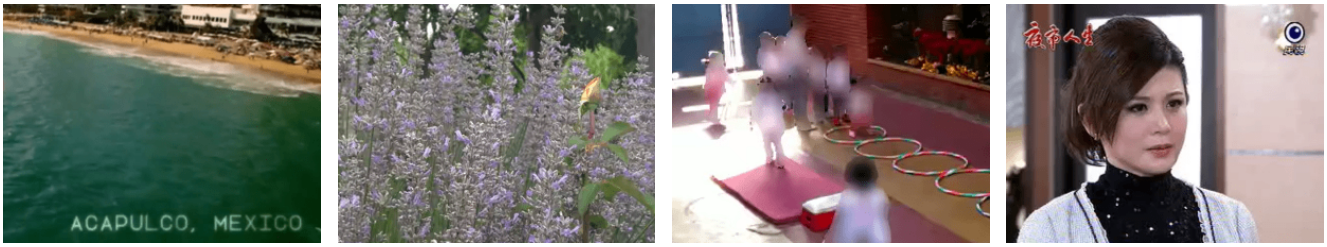
2.4. Generated Examples

The generated results demonstrate the significant promise of our framework. The first example (see Fig. 2a) demonstrates the difference between our approach and others. This is the first video from https://amaai-lab.github.io/Video2Music/. Compared to existing methodologies, our framework exhibits enhanced capabilities in producing a wider scope of music styles that better align with the content of the provided video. Notably, our generated content even captures the ambient sound of the sea. Subsequent examples, depicted in Fig. 2b and Fig. 2c, further show our framework's adeptness in crafting music suitable for various contexts, including natural scenery and human activities. Notably, an intriguing outcome emerges when applying our methodology to a dramatic context (see Fig. 2d).

3. Experimental Result

3.1. Dataset

We utilize the video dataset from the VCSL [4], employed in the 2023 Video Similarity Challenge Codebase organized by Meta. Although the VCSL was initially designed for segment-level video copy detection, our experiments leverage its extensive coverage of diverse video topics, demonstrating the versatility of our framework in generating diverse music for various kinds of videos.



(a) https://youtu.be/SkvyQqMoURc (b) https://youtu.be/x6vyCrWmyD0 (c) https://youtu.be/3wrogd_Nfqk (d) https://youtu.be/rPNGSi7lbGE

Figure 2. Some examples of our generated results.

For the reference video-audio dataset, we utilize data from SymMV [10]. SymMV comprises a video and symbolic music data collection, enriched with various musical annotations. The primary content of these videos includes piano covers and their corresponding music videos, offering a specialized dataset for our analysis.

3.2. Subjective Evaluation and Results

Subjective evaluations are commonly employed in previous studies on generated audio [5, 10]. We also believe that a user study is the most effective method for evaluating our work. To this end, we conducted a user study utilizing Google Forms (https://forms.gle/NSgNfj9Xo5Y6ZJAWA). The questionnaire takes about 10 minutes to complete. The participant pool for our study consisted of 15 individuals.

In our study, the questionnaire is divided into three sections designed to evaluate various aspects of our framework. The first section, "Emotion of Audio and Video", focuses on the framework's ability to identify and reproduce emotions accurately.This section contains three tasks. The initial two tasks involve evaluating emotions in silent videos and audio-only clips; the result is present as the precision of emotion matching (P_{SEmo}). The emotions evaluated include happiness, sadness, disgust, anger, fear, surprise, and neutrality. The final task requires participants to rate the alignment of emotions between the video and its accompanying audio (R_{Cong}) on a five-point scale.

The second section, "Quality of the Video with Generated Audio", evaluates the quality of the generated music and its harmonious integration with the video. The assessment is conducted across several specified dimensions on a five-point scale. The evaluation criteria include: The aspects are as follows: 1) **Richness**: Evaluates whether the generated music exhibits sufficient diversity. 2) **Fitness**: Assesses whether the audio complements the visual content effectively. 3) **Humanness**: Determines the expressive quality of the generated audio. 4) **Fondness**: Seeks to capture the overall enjoyment or appeal of the video.

The final section, entitled "Best Video in Different Configurations", is focused on identifying which configuration yields audio that most accurately corresponds with the associated video. This evaluation first compares the balance of the video generated by configurations: Simple, Brief@30, Long@50, and Audio-matching-desc, then assesses the generative capabilities of the Audio-matching-desc and Audio-matching-title models. The distinction among Simple, Brief@30, and Long@50 relates to the length of the instructions generated by ChatGPT. Audio-matching-desc and Audio-matching-title denote different targets for matching music. Participants are asked to rank the audio track they believe best complements the video.

Config	R_{Ric}	R_{Fit}	R_{Hum}	R_{Fond}
Simple	2.92	2.71	2.75	2.56
Brief@30	3	**3.06**	**2.98**	2.69
Long@50	**3.31**	3.02	2.88	**2.88**

Table 1. Subjective evaluation result - Section 2: Quality of the Video. (1 is the lowest, and 5 is the highest.)

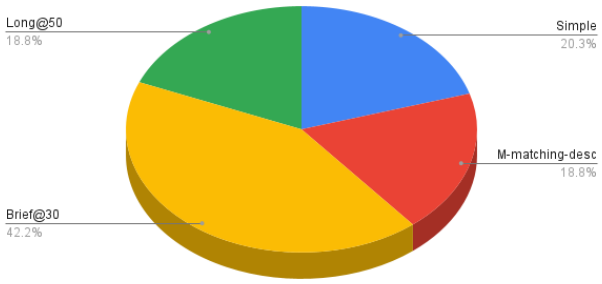


Figure 3. Subjective evaluation result - Section 3: Best Video in Different Configurations.

Results and Analysis For emotion recognition in Drama video, a low precision score (P_{SEmo}=0.10) and a moderate congruence score (R_{Cong}=3.06) on a 5-point scale suggest that while individual emotional recognition via video and

audio is imprecise, with 17% of audio rated as neutral, the overall matching is reasonably effective. The discrepancy may arise because videos typically convey a more concise range of emotions, whereas a single audio track can interpret a broader spectrum of emotions.

Regarding video-audio alignment, as illustrated in Fig. 3, shows that the Brief@30 method achieves superior matching between video and audio, followed by the Simple method. Complex motions like animal migration are more significantly aligned with the Brief@30 method. However, simple motions, such as a bird's-eye view of track and field events, show no significant differences between the Simple and Brief@30 methods.

In terms of audio quality, as shown in Tab. 1, the Brief@30 model scores highest in fitness and humanness, whereas the Long@50 model excels in richness and fondness. This suggests that while medium-sized prompts like Brief@30 provide sufficient alignment with less complex musical compositions, longer prompts like Long@50, despite their richer layers, may cause the audio to diverge from the video content.

3.3. Objective Evaluation and Results

The subjective nature of how individuals perceive sound and vision complicates the objective evaluation of the compatibility between audio and video. However, we propose three methods to comprehensively evaluate video-audio integration: video-description accuracy, audio-motion accuracy and music quality assessment.

We developed a Video-Music CLIP precision model inspired by [10] for video-description accuracy. Our method involves extracting three frames from each video scene, using the CLIP model to compare against a pre-generated list of textual descriptions. We rank the results by cosine similarity and calculate precision (P_{desc}) based on the accuracy of the top-ranking textual descriptor for each video.

For audio-emotion accuracy, we categorized each scene of a Drama video into one of four emotions: happiness, sadness, anger, or neutrality. Using an emotion recognition model from SpeechBrain [9], we predict the emotion for each scene's audio. Precision is calculated for all scenes (P_{emo}) and scenes with explicit facial expressions ($P_{emoFace}$), assessing model accuracy by comparing predicted emotions against the annotated labels. Higher precision reflects more accurate model performance.

Also, we selected objective music metrics using MusPy, developed by [3], to evaluate music quality. We implemented three pitch-related metrics and one rhythm-related metric: 1) **Scale Consistency (SC)**: the maximum rate of pitches aligning with any major or minor scale. 2) **Pitch Entropy (PE)**: the Shannon entropy of the normalized note pitch histogram. 3) **Pitch Class Entropy (PCE)**: the Shannon entropy of the normalized note pitch class histogram. 4) **Empty Beat Rate (EBR)**: the proportion of beats without any note played to the total number of beats.

These metrics assess how closely the generated audio resembles natural audio. For this purpose, we employ the SymMV [10] test set as a benchmark for natural audio.

Results and Analysis As illustrated in Tab. 2, the Brief@30 model excels in video-music correspondence, consistent with subjective evaluations. This indicates that expanding the output of the ChatGPT model might result in the generation of imprecise language, thereby introducing redundant information for music generation models. On the other hand, it can also cause the performance degradation during the music matching process.

For emotion congruence within the Drama method, a precision of 0.5 is recorded when a face is present in the video ($P_{emoFace}$). A potential reason for the model not achieving higher precision is the limitation of the SpeechBrain model. Based on the result it provided, 36% of the results were labeled as neutral, suggesting that SpeechBrain may struggle to identify the emotions conveyed by the generated music accurately. However, it still reflects the trend we observed: the model exhibits a higher capacity for detecting emotions when human faces are present, consequently having a higher probability of generating music that better matches the detected emotion.

For the music quality evaluation, the Brief@30 configuration excels in terms of music quality compared to other configurations, which is consistent with subjective evaluations. The possible reason for V-MusProd to have higher music quality in other metrics is that the music in SymMV and V-MusProd primarily comprises piano and classical pieces. In contrast, our model employs a more diverse instrumental structure, incorporating sounds such as ocean waves, drums, and techno music. This feature enhances video-music integration but potentially lower the performance value during objective evaluation.

4. How to Reproduce the Work

For our code and more generated results, please refer to https://drive.google.com/drive/folders/1VeMgawUTRC5NY1UwCI5q_YSY8jyqveRm?usp=sharing

We also make our code available on Github: https://github.com/janisme/Making-Music-for-Videos-Evaluation.git.

5. Conclusion

In this project, we propose a generalized framework to create diverse and engaging music for non domain-specific videos. Based on state-of-the-art models, our framework generates music that fits well with video content while making music not limited to simple melodies or specific instruments. We also provide various evaluation metrics and insight into our work.

Methods	Video-Description Correspondence			Music Quality			
	P_{desc}	P_{emo}	$P_{emoFace}$	SC	PE	PCE	EBR
Real (SymMV)	-	-	-	0.986	4.197	2.633	0.023
V-MusProd [10]	-	-	-	**0.983**	**3.940**	**2.607**	0.004
Simple	0.877	-	-	0.967	3.05	1.95	0.06
Brief@30	**0.911**	-	-	0.990	3.62	2.13	**0.01**
Long@50	0.863	-	-	0.977	3.28	2.04	0.06
Drama	-	0.455	0.5	0.964	2.938	1.857	0

Table 2. Objective evaluation result.

6. Contribution

Equal contribution.

Pei-Huan is in charge of the methodology design and implementation.

YenChu is in charge of the evaluation part of the project and also helping the project development.

References

[1] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2024. 2

[2] 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 *Proceedings of the 29th ACM International Conference on Multimedia*, page 2037–2045, 2021. 1

[3] Hao-Wen Dong, Ke Chen, Julian McAuley, and Taylor Berg-Kirkpatrick. Muspy: A toolkit for symbolic music generation, 2020. 4

[4] Sifeng He, Xudong Yang, and et al. Jiang, Chen. A large-scale comprehensive dataset and copy-overlap aware evaluation protocol for segment-level video copy detection. *arXiv preprint arXiv:2203.02654*, 2022. 2

[5] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs, 2021. 3

[6] Jaeyong Kang, Soujanya Poria, and Dorien Herremans. Video2music: Suitable music generation from videos using an affective multimodal transformer model. *Expert Systems with Applications*, 2024. 1

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[9] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624. 4

[10] L. Zhuo, Z. Wang, B. Wang, Y. Liao, C. Bao, S. Peng, S. Han, A. Zhang, F. Fang, and S. Liu. Video background music generation: Dataset, method and evaluation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15591–15601, 2023. 1, 3, 4, 5