Diff-BGM: A Diffusion Model for Video Background Music Generation

Sizhe Li[1] Yiming Qin[2] Minghang Zheng[1] Xin Jin[3,4] Yang Liu[1*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Yuanpei College, Peking University
[3]Beijing Electronic Science and Technology Institute
[4]Beijing Institute for General Artificial Intelligence
{lisizhe, yangliu, minghang}@pku.edu.cn ymk4474@gmail.com jinxinbesti@foxmail.com

Abstract

When editing a video, a piece of attractive background music is indispensable. However, the video background music generation tasks face several challenges, for example, the lack of suitable training datasets, and the difficulties in flexibly controlling the music generation process and sequentially aligning the video and music. In this work, we first propose a high-quality music-video dataset BGM909 with detailed semantics annotation and shot detection to provide multi-modal information about the video and music. We then present evaluation metrics that go beyond assessing music quality, including a metric for evaluating diversity and alignment between music and video with retrieval precision metrics. Finally, we propose the Diff-BGM framework to automatically generate the background music for a given video, which uses different signals to control different aspects of the music during the generation process, i.e., uses dynamic video features to control music rhythm and semantic features to control the melody and atmosphere. We propose to align the video and music sequentially by introducing a segment-aware cross-attention layer. Experiments verify the effectiveness of our proposed method.

1. Introduction

With the rapid development of multimedia and social platforms, videos become a common way to convey feelings and record lives. When creating videos, to make the video more attractive, a piece of suitable and melodious background music is crucial. However, it is not easy for those who do not have much knowledge of music or video editing to select or create proper or perfectly matched music. What's more, the copyright protection issue has also caused broader public concern. As a result, it is pragmatic to automatically generate background music for a given video.

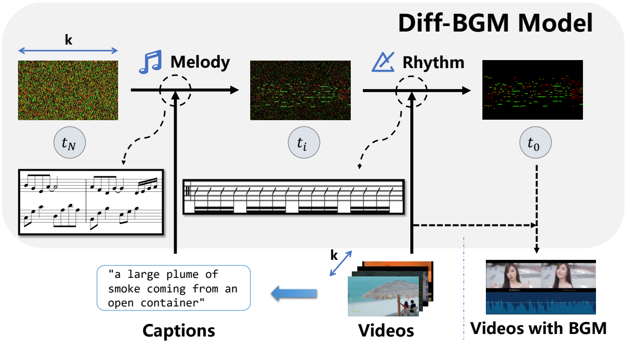


Figure 1. **Overview of the background music generation process of Diff-BGM.** At different stages of the generation process, Diff-BGM uses different features of videos or captions to control the generation of the music rhythm or melody. We find that the video dynamic feature has more control over the rhythm and the semantic feature has more control over the melody.

Existing works [1, 2, 5, 11, 18–22, 24, 25, 29, 32, 34, 38] have started focusing on music generation and achieved good results. Besides, some works [7, 35, 36] focus on generating music for human-centric videos. In comparison, free-style video background music generation tasks present more challenges. First, to generate proper background music for a video, the model needs to consider multiple aspects of information in the video to **control different aspects of the music**. In a piece of music, several elements work together to make it pleasant and concordant to listen to. For example, using different rhythms when composing makes the music sound different in dynamic, and different melodies often reflect different atmospheres. Faster music often gives a liveliness or intense feeling, which is suitable for videos that change quickly while slower music tends to be more soothing and is suitable as a warm-style video soundtrack. Sad video clips should correspond to heavy styles and melodies while cheerful videos are matched by upbeat music. *We conclude that visual dynamic changes are*

[*] Corresponding author

linked to music rhythm(the time when the notes appear) and visual semantics influence the melody and atmosphere of the music. However, most existing transformer-based models [4, 39] cannot intuitively reflect the music generation process with corresponding control signals and lack good interpretability. Models for human-centric videos also abstract rhythm information from human motion, which are also unsuitable for freestyle videos. Secondly, compared to music generation, video-conditioned background music generation requires models to **temporally align** the video and the music. For example, if a video is composed of many transitions of shot, then each transition should have a more prominent sound to indicate the sudden visual changes.

According to the above challenges, we are the first to consider both fronts. However, there lack suitable datasets. Open-source datasets previously used for other music generation tasks either lack corresponded free-style video samples [8, 30], or fail to provide complete annotations of audio or video information [10, 15, 17, 37, 39], making them unsuitable for video background music generation. Details are shown in Tab. 1. As a result, we collect a video-music dataset named BGM909. Compared to existing datasets, BGM909 has several advantages for background music generation. Firstly, we provide high-quality music files, along with comprehensive annotations for various aspects of the audio, such as chords, beats, key signatures, and more. These annotations assist in helping the model learn to analyze music structure and composition. Secondly, we offer videos that align with the audio content. Specifically, for song audios, we provide their official MV videos, ensuring semantic consistency between music and video. Moreover, our videos undergo manual editing and human checks to ensure perfect temporal alignment with the music. Additionally, we provide detailed annotations for video including fine-grained natural language descriptions and video shot transitions. To evaluate the quality of the generated music, we also provide new metrics to measure the music quality and the video-music correspondence.

To tackle the proposed challenges, we propose a framework named **Diff**usion-based **B**ack**G**round **M**usic generation(**Diff-BGM**) to generate video-aligned background music. For the first challenge, we use diffusion-based models as our framework. It is a recursive process to generate background music so that we can use different signals to control different aspects of music. As shown in Fig. 1, to make the music style and atmosphere correspond with the given video, we visualize the music generation process and involve the semantic feature of the video to control the style and melody of the generated music. Then we use the dynamic video feature to control the generation of music rhythm so that the timing information in the two modalities is aligned. For the second one, we propose a segment-aware cross-attention layer to improve the diffusion framework and sequentially align the video and music. The purpose of temporal alignment is to synchronize the music and video for each segment. Therefore, we introduce cross-attention within the diffusion model and apply time encoding to both modalities. We believe that music generation should be influenced by short-term contexts within the video. Hence, we designed a specific mask to constrain the attention mechanism, obtaining better temporal alignment.

Our contributions are summarized as follows: (1) We present BGM909, a high-quality video-music dataset with detailed annotations for background music generation. Also, we provide evaluation metrics to measure the video-music correspondence and diversity. (2) We propose Diff-BGM, the first diffusion-based network for background music generation. It controls the generation process in stages from different dimensions of the video and increases the interpretability of the generation process. (3) Both objective and subjective evaluation demonstrate that Diff-BGM generates high-quality background music and surpasses the state-of-the-art model.

2. Related Work

2.1. Music Generation

In recent years, music generation has attracted much attention, and many models working on music generation have been proposed. [5, 11, 22, 32, 34] propose to generate music based on transformer and gain satisfying generation results. However, transformer-based models often rely on manually designed tokens to encode music, leading to limited generative ability. With the development of diffusion models, it demonstrates remarkable performance not only in visual tasks but also in music generation. Some works propose diffusion-based model to utilize its excellent generative ability. [19–21] focus on generating music by exerting control on music while [2, 25, 29] use text as condition to generate music. [2, 25, 29, 38] build bridge between music and other several modals(visual, text, etc). However, those methods do not consider the time alignment between music and the input conditions(like video). As a result, they cannot solve the video background music generation task.

2.2. Background Music Generation

Some methods [7, 35, 36] focus on generating music for human-centric videos (i.e. dance or sports videos), in which the rhythm is largely dependent on human motions and is not accessible in freestyle videos. The task of video background music generation was first proposed by CMT [4] and has gained more and more attention. Existing background music generation methods mostly use the transformer-based framework and establish the relationship between video and music. CMT [4] first encodes the chords and notes to give the representation of a piece of music and train the model

Dataset	Video	Audio[1]	MIDI	Style	Chord	Melody	Beat	Caption	Shot dec.	Size
MAESTRO [8]	✗	✓	✓	✗	✗	✗	✗	—	—	1,276
POP909 [30]	✗	✗	✓	✓	✓	✓	✓	—	—	909
HIMV-200k [10]	✓	✓	✗	✗	✗	✗	✗	✗	✗	200,500
TikTok [37]	✓	✓	✗	✗	✗	✗	✗	✗	✗	445
AIST++ [17]	✓	✓	✗	✓	✗	✗	✗	✗	✗	1,408
URMP [15]	✓	✓	✓	✗	✗	✗	✗	✗	✗	44
SymMV [39]	✓	✓	✓	✓	✓	✓	✗	✗	✗	1,140
BGM909(Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	909

Table 1. **Comparison between different music datasets.** Our proposed BGM909 contains a considerable amount of video-music pairs, being able to be applied to the background music generation task. Different from existing datasets, BGM909 provides various musical annotations, metadata, captions and shot detection. Two popular music datasets are shown in the first two rows for reference.

to understand the logic of music, then it establishes three rhythmic relations (*e.g.* motion speed of the video corresponds to note density in the audio) between the video and background music to narrow the gap between the two forms of expression. Other than only using the rule-based rhythmic relationships, V-MusProd [39] and Video2Music [12] focus on semantic-level correspondence. They extract the semantic feature of the video to control the style of the generated music in the multi-modal transformer blocks. However, transformer-based methods suffer from the same challenges that it is hard to control the process of end-to-end generation thus leading to poor interpretability. We propose a diffusion-based framework to generate background music for videos to make full use of the generative ability of diffusion models. Besides, we use different features to control different aspects during the generation process and conduct temporal alignment between video and music.

3. Dataset

Due to the lack of high-quality open-source datasets for background music generation task, we collect a new video-music dataset BGM909 based on POP909 [30] containing 909 pieces of piano version music and their corresponding well-aligned videos. Compared to existing datasets, BGM909 has several advantages. (1) We provide high-quality MIDI files of music, and detailed annotations such as chords, beats, styles etc. (2) The content of the videos aligns with the music. Specifically, we provide the official MV videos for each song music to ensure semantic coherence. (3) We also manually edit and check the video-music pairs to ensure perfect temporal alignment. (4) Detailed annotations for videos including fine-grained natural language descriptions and shot transitions are provided for further study. Tab. 1 shows the comparison of BGM909 with other existing video-music datasets.

3.1. Data Collection

Tons of music-video pairs are available on the Internet. However, it is hard to gain high-quality noiseless audio from those videos. As a result, we start with the existing well-annotated music in POP909 [30] dataset. For each MIDI file in POP909, we collect its metadata and use the song title and singer as keywords to search for the corresponding official video on YouTube. After downloading the videos, we remove those that only had static interfaces or lyrics. Then for those left videos, we manually edit the video to align it with the MIDI file temporally (*e.g.* some audios are not played at the beginning of the videos). To ensure dataset quality, we manually check the collected video-music pairs.

3.2. Data Annotation

Melody, Bridge and Piano. Each MIDI file contains three tracks of melody, bridge, and piano, representing the lead melody transcription, the secondary melody and the main body of the accompaniment separately.

Chord, Beat and Key Signature. A chord is a number of notes played at the same time, specific notes comprise harmony chords and play an important role in setting the base tone of the music. Beats mean the length of each note and are basic components of music rhythm. Key signature represents the tonality of the music. We provide chord, beat, and key signatures separately, including beat and downbeat annotations, start and end time for each chord, and chord names. Detailed algorithms can be found in [30].

Natural Language Descriptions. We provide fine-grained natural language descriptions for each video in BGM909. We generate 10 description sentences for each 8 frames in each video with a pre-trained BLIP [16] model. It also serves as a lightweight substitute for video features, capable of fully expressing the semantic information of the video. Besides, the captions can be extended to other tasks like text-to-music generation.

Camera Shot Detection. We extract each shot of the video based on the camera switching. Music often stands out at the transitions in the video, therefore, captured shot transitions often have a significant impact on the rhythm and variations of the music. The timing of shot transitions is also a focal point to pay attention to during music generation. Shot detection assists in training the alignment between music and video. Videos have 62 shots on average.

Styles. We provide GPT with the song's name and the associated artist and obtain the style classification of each audio to further improve the dataset and prove the generality of BGM909. The songs are from 646 different artists and are divided into 8 different styles in total.

Metadata. We also provide extra metadata for the music in BGM909 compared with POP909, like lyrics, genre, and rhythmic pattern. The metadata information is useful for data analysis and may be used in future research works like music-to-video generation.

4. Method

We propose a novel music generation pipeline named Diff-BGM following the principles of latent diffusion models [27] to deal with the video background music generation task. The framework of Diff-BGM is shown in Fig. 2(left), which contains a Music Process Module, a Video Process Module, a Generative Model and an Output Generation Module. The music process module takes original midi files as input and generates corresponding piano rolls to represent the music. The video process module takes the original videos as input. To guide the generation process, we extract the visual features of the video, generate captions for the video and extract language features for those captions. The generative module is a diffusion model, which uses the extracted features as conditions to generate a new piano roll. In the end, after gaining the generated piano roll, the output generation module is used to process the piano roll and generate the corresponding music. In order to control different stages of music generation using different features, we introduced a feature selector, as shown in Fig. 2(right). To consider the timing in the video and the music, and align the video and the generated music, we introduce segment-aware cross-attention to align the video and music.

4.1. Polyffusion Baseline Revisited

We use Polyffusion [21] as our baseline. It is trained on midi data and outputs a midi file for unconditional generating. Polyffusion is a diffusion-based music generation framework, with a Denoising Diffusion Probabilistic Model [9] as its generative baseline and provides a complete midi process algorithm. We follow the structure and data process algorithm proposed by Polyffusion.

Music Process. Polyffusion divides the input midi music into 8-bar (32-beat, $T_0 = 128$ time steps) segments and transfers it into image-like piano roll representation $x \in \mathbb{R}^{2\times T_0 \times P}$, which is a 2-channel binary tensor. The MIDI pitch ranges 0...127 so we gain $P = 128$ pitch bins. In the piano roll representation, entry $a(c, t, p)$ represents at time step t and MIDI pitch p whether there is a note onset($c = 0$) or sustain($c = 1$).

Generative model. Polyffusion uses a latent diffusion model [27] as generator to generate piano rolls for videos. It contains a diffusion process and a denoising process. In the diffusion process, the structure of data x_0 is broken up step by step by iteratively adding Gaussian noises in N steps:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

$$q(x_{1:N}|x_0) = \prod_{t=1}^{N} q(x_t|x_{t-1}) \tag{2}$$

where $\beta_1, \beta_2, ..., \beta_N$ are a set of variance scheduling parameters, x_0 is the clean input piano roll. Then in the denoising process, the model learns to reconstruct the original structure of x_0 from the noisy input $x_N \sim \mathcal{N}(0, \mathbf{I})$. It is defined as a Markov chain with learned Gaussian transitions:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)) \tag{3}$$

$$p_\theta(x_{0:N}) = p(x_N) \prod_{t=1}^{N} p_\theta(x_{t-1}|x_t) \tag{4}$$

The training process is performed by optimizing the following target:

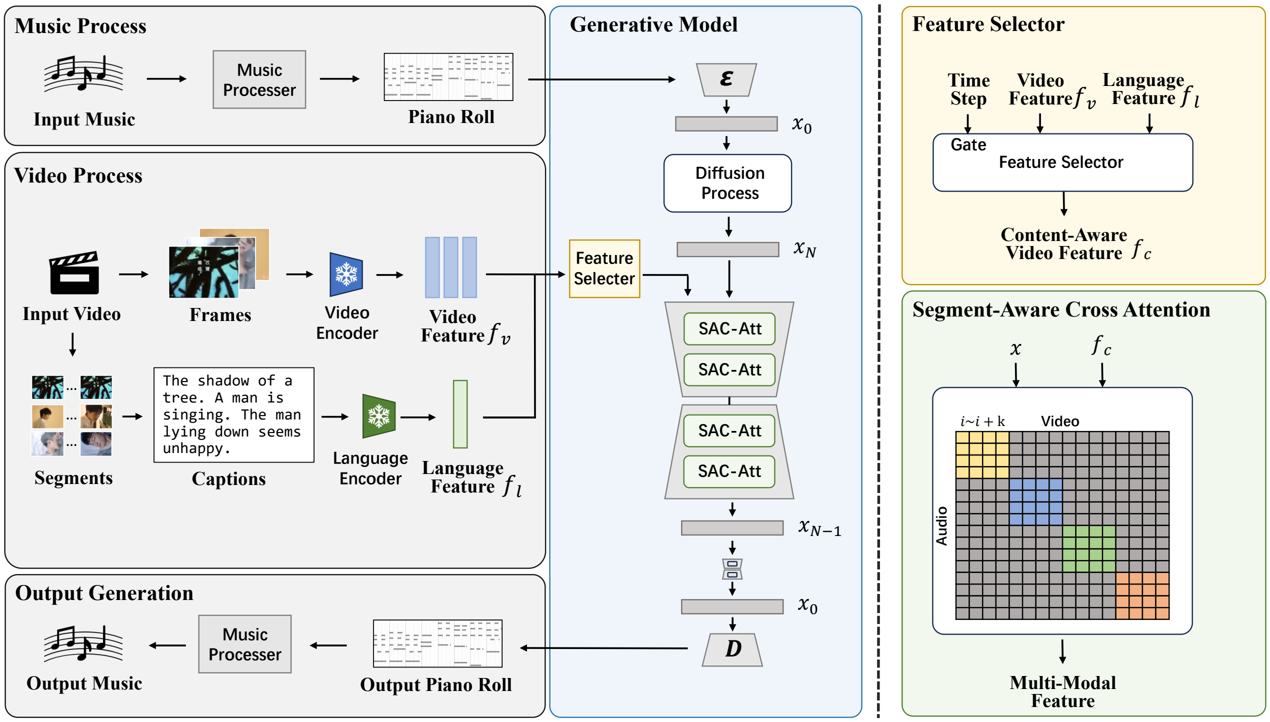
$$L(\theta) = \mathbb{E}_{x_0,\epsilon,t}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2] \tag{5}$$

where ϵ_θ represents the model parameters, t is uniform between 1 and N, $\epsilon \sim \mathcal{N}(0, I)$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

Although Polyffusion has the ability to generate music unconditionally, it still cannot generate background music for a given video and has not addressed the two challenges mentioned in Sec. 1, namely the inability to achieve conditional control and alignment between music and video. Therefore, to achieve temporal alignment between music and video, we made improvements upon Polyffusion.

4.2. Video Process

Diff-BGM model takes videos V as input. We first sample T frames and use a pre-trained video encoder to extract the visual feature $F_v \in \mathbb{R}^{T\times d_1}$ of each frame. Besides, we segment the video into T segments $V = \{S_1, S_2, ..., S_T\}$ and generate natural language captions C_i according to the video content for each segment S_i. For the captions, we use a pre-trained language encoder to extract the language features $F_l \in \mathbb{R}^{T\times d_2}$ for the captions.



The shadow of a tree. A man is singing. The man lying down seems unhappy.

Figure 2. **Illustration of our Diff-BGM model.** We process the input music and video, gain piano rolls to represent the music and extract visual features to represent the videos. In order to get richer semantic information, we segment the video and generate captions then extract language features. The backbone of the generation model is a diffusion model, with the processed visual and language features as conditions to guide the generation process. We propose a feature selector to choose features to control the generation process. And to better align the timing of the music and video, we design segment-aware cross-attention layer to grasp the timing feature in different modalities.

4.3. Feature Selection.

We analyze the timestep intervals defined as [23] to show the type of attribute in the music controlled by each interval. By observing the unconditional music generation process, we found that models tend to generate the melody, which is influenced by the semantics, and then generate the rhythm of the music, which is related to the dynamic feature of the video. As a result, at different timestep intervals, we use different features as conditions to describe the video.

The final condition feature F_c is represented as:

$$F_c = \begin{cases} F_l, & \text{TimeStep} > t_0 \\ F_v, & \text{TimeStep} \leq t_0 \end{cases} \tag{6}$$

where t_0 is a hyper-parameter representing the key time step during the denoising process decreasing from N to 0, Timestep represents the current step of denoising, F_c is the condition feature of the video.

4.4. Sequential Attention

We aim to generate music for given videos, which means the style and atmosphere of the music should match the semantic content of the video. Besides, when shot changes or noticeable motion changes exist, there should be a homologous response in the music. Obviously, unconditional diffusion and denoising process cannot achieve this goal. Firstly, as we require the generated music to match the given video in terms of rhythm, melody, and other aspects, we introduce video features as conditions. We also propose a segment-aware cross-attention layer to fuse the video features with music features in the latent space of the diffusion model, ensuring that the generation process is consistently guided by the video, resulting in music related to the video. Additionally, to achieve fine-grained temporal alignment between music rhythm and the video, we applied time encoding to both and introduced specially designed masks to conduct sequential attention. These enable the music generation process to incorporate small-scale video features as context, facilitating precise alignment between the two and generating high-quality background music.

In order to align video and music sequences and understand the context information in both modalities, we follow Latent Diffusion [27] and design a segment-aware cross-attention layer. The input noisy latent representation x_t serves as Query, while the condition feature F_c serves as

Key and Value. Then the attention can be represented as:

$$Attn = \frac{QK^\top}{\sqrt{d_{key}}} \tag{7}$$

where Q, K denote Query and Key separately, d_{key} is the dimension of the condition feature. However, to align the two modalities in timing, long-term context is not so important as short-term context for music is often associated with the current clip of the video. As a result, a special mask is designed to only pay attention to short-term context and neglect long-term context. As shown in Fig. 2, we divide adjacent k frames into short-term contexts, then only the features of those k adjacent frames can influence the generated music at each time spot. The mask is given as follows:

$$\mathcal{M}ask_{i,j} = \begin{cases} 1, & k \cdot \gamma \le i, j < k \cdot (\gamma + 1) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $\mathcal{M}ask \in \mathbb{R}^{T \times T}$ represents the designed attention mask, $\gamma \in \{0, 1, .., \frac{T}{k}\}$ represents the number of contexts.

As a result, the output of the segment-aware cross-attention layer is as follows:

$$x_{out} = \underset{seq}{softmax}\left(\mathcal{M}ask\left(\frac{QK^\top}{\sqrt{d_{key}}}\right)\right)V \tag{9}$$

where V denotes Value. Then the output x_{out} combines the short-term context features of both video and music so that the model is able to generate video-aligned music.

4.5. Train and Inference

Following the strategy above, the final training objective is

$$L_{cond}(\theta) = \mathbb{E}_{x_0, F_c, \epsilon, t}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, F_c)||^2]. \tag{10}$$

where x_0 represents the original clean piano roll, F_c denotes the condition feature, t denotes the time step. During inference, Diff-BGM receives a random noise as input x_N and uses the video dynamic and semantic features as conditions. We can control the generation process by flexibly adjusting the key time step t_0 to select the condition features and generate diverse music for a video.

5. Experiments

5.1. Implementation Details

To make a fair comparison, we follow previous work [9, 21] to use a Gaussian noise schedule and the noise prediction objective in Sec. 4.1 for all experiments. Our segment-aware cross-attention layers are set as [27]. The diffusion step N is set to 1,000. Diff-BGM model converges around 100 epochs on Adam Optimizer [13] with a constant learning rate 5e-5. We use official pre-trained Video CLIP [33] as video encoder to extract visual features. We choose the BLIP [16] model to segment the video and use official pre-trained bert-base-uncased model [3] as the language encoder. The visual and language encoders keep frozen during training. In the segment-aware cross-attention module, we set k to 8 and t_0 to 200.

5.2. Objective Evaluation

Metrics. As for evaluating metrics for the task of generating background music for videos, they have not been fully refined to date. Therefore, building upon existing metrics, we have proposed additional metrics to assess the generated results of background music as follows:

- **Music Quality.** We choose the same metrics as [4, 31] to evaluate music quality, including *Pitch Class Histogram Entropy(PCHE)* which measures the uncertainty of the distribution of the notes and reflects the quality of tonality, *Grooving Pattern Similarity(GPS)* which measures the quality of the rhythmicity, and *Structureness Indicator(SI)* which captures the repetition in the music by measuring the overall structure and reflects the catchiness and the emotion-provoking nature [14]. On SymMV [39] dataset, scale consistency(SC) is also used to evaluate the music quality. Note that the overall quality is not indicated by how high or low these metrics are, but instead by their *closeness* to the real music data.
- **Diversity** We propose a metric to evaluate the diversity of the generated music. We randomly divide the generated music into two subsets, S_d samples in each set. The diversity of the generated music is defined as:

$$\text{Diversity} = \frac{1}{S_d}\sum_{i=1}^{S_d}||v_i - v_i'||_2 \tag{11}$$

  where v_i, v_i' represent the music feature of the $i-$th sample in the two subsets separately.
- **Music Retrieval** We propose a new metric to measure the music-video consistency. Given a piece of generated music $\hat{m}$ and the ground truth music of its condition video m, we randomly select $M-1$ pieces of music m_i. Here we use Musicnn [26] to extract music feature for each generated item. If the ground-truth music ranks in the top-K place, then we consider it a successful retrieval. All generated samples are used to calculate the successful retrieval rate as the final precision score $P@K$. Here we set $M = 64, K = 5, 10, 20$. Since the ground truth music is related to the given video, the proposed retrieval precision metric is able to measure how well the generated music aligns with the given video.

Results. The results on BGM909 test set are shown in Tab. 2. Compared with CMT [4], our Diff-BGM surpasses it on both music quality metrics and video-music correspondence, and has a gain of $4.91\%, 8.15\%, 10.39\%$

Methods	Music Quality			Video-Music Correspondence			Diversity↑
	PCHE→	GPS→	SI→	P@5↑	P@10↑	P@20↑	
Real(BGM909)	2.717	0.708	0.486	—	—	—	6.664
CMT [4]	2.398	0.594	0.761	10.56%	18.59%	36.40%	6.025
Riffusion [6]	2.556	0.509	0.412	9.12%	15.71%	34.29%	6.335
Unconditional	3.189	0.595	0.528	8.08%	15.89%	31.93%	**6.421**
+ Video feature	2.835	0.514	0.396	**13.44%**	23.54%	43.20%	5.742
+ Feature selection	**2.721**	**0.789**	0.523	11.00%	20.79%	38.47%	5.246
+ SAC-Attn (Diff-BGM)	2.840	0.601	**0.521**	13.28%	**23.91%**	**44.10%**	5.153

Table 2. **Objective evaluation results on BGM909 test set.** We evaluate both music quality and video-music correspondence with several metrics, where P indicates precision, the higher P is better. PCHE indicates Pitch Class Histogram Entropy, GPS indicates Grooving Pattern Similarity and SI means Structureness Indicator, where closer to Real is better.

Methods	Music Quality		Video-Music Correspondence			
	PCHE→	SC→	P@5↑	P@10↑	P@20↑	AR↓
Real(official test set)	2.633	0.986	—	—	—	—
CMT [4]	2.444	0.990	8.9	17.7	31.0	33.4
V-MusProd [39]	2.607	0.983	15.7	24.6	44.8	25.4
Real(our test set)	2.538	0.882	—	—	—	—
Diff-BGM	2.738	0.878	**19.0**	**28.6**	**47.6**	**19.4**

Table 3. Objective Evaluation on SymMV test set. We evaluate music quality with Pitch Class Entropy and Scale Consistency and evaluate video-music correspondence, where P represents precision, AR represents average rank of the ground truth video

.

Metrics		Rates	
		Experts	Non-Exp.
Music Quality	M. Melody	75.0%	81.5%
	M. Rhythm	63.9%	74.4%
Video-Music Correspondence	V. Content	75.0%	80.4%
	V. Rhythm	70.4%	75.0%
Expertise	Chord	70.4%	—
	Accom.	78.7%	—
Overall	Overall Pref.	77.8%	83.9%

Table 4. **Subjective evaluation.** The preference rates for Diff-BGM against CMT [4] are shown in music quality metrics, video-music correspondence metrics, and expertise metrics.

on the retrieval metircs, which proves that Diff-BGM generates higher-quality music and has a better understanding of the correspondence between video and music.

Results on SymMV [39] dataset are shown in Tab. 3. For comparison, we have re-divided the train/val/test sets based on the instructions provided in [39], ensuring that the size of each set aligns with the proposed official sizes[2]. Methods in the first block in Tab. 3 are evaluated on the official SymMV test set, while methods in the second block are on the test split we obtained [3]. Since the test split is not identical [4], direct numerical comparisons of music quality metrics cannot be made. However, it's important to note that retrieval-based metrics (for video-music correspondence evaluation) can still be compared in a relatively fair manner, given that the size of the retrieval pool is consistent. As shown in Tab. 3, Diff-BGM outperforms CMT and V-MusProd in video-music correspondence metrics by a large margin, indicating that Diff-BGM can effectively align video and music during the generation process.

5.3. Subjective Evaluation

The best way to evaluate a generative model today remains using user study, and it is widely adopted in previous works [4, 28, 31, 39]. We conduct the user study by designing and sending out questionnaires. We invite 46 people to participate in the user study, 18 of them are experts with expert knowledge in music, 28 are non-experts. We choose videos from different categories then use Diff-BGM and CMT to generate music for each video separately, present them randomly for blindness, and require the participants to compare the two generation results in several aspects and give preference scores separately. For some videos are long, the questionnaire takes about 25 minutes to complete.

[2] As of now, SymMV has not publicly disclosed complete information, including the partitioning of train/val/test sets and the alignment timestamps between video and audio.

[3] We are unable to present the performance results of V-MusProd on the new split since their code is not publicly accessible.

[4] The numbers of real data on the two splits are different

	M. Mel	M. Rhy	V. Content	V. Rhy	Chord	Accom.	Overall	Conf.
Human	3.48±0.99	3.46±0.99	3.45±1.25	3.28±1.06	3.26±0.97	3.39±1.09	3.48±1.02	3.90
Riffusion[6]	2.97±1.11	2.86±1.02	2.74±1.30	2.68±1.15	2.89±1.02	2.86±1.08	2.94±1.13	3.88
CMT [4]	2.94±1.04	2.97±1.13	2.74±1.22	2.72±1.19	2.81±1.02	2.78±1.12	2.88±1.10	3.73
ours	**3.29**±0.87	**3.22**±0.96	**3.14**±1.03	**3.16**±1.09	**3.11**±1.03	**3.22**±0.97	**3.33**±0.99	**3.78**

Table 5. **Subjective evaluation.** Experts are asked to score 6 music generated by different models and also created by humans.

Methods	Music Quality		
	PCHE→	GPS→	SI→
Only Video	2.835	0.514	0.396
Only Language	2.849	**0.641**	0.521
Video+Language	2.840	0.626	0.536
Language+Video	**2.781**	0.601	**0.517**

Table 6. **Ablation studies on feature selector.** We use different features in lines 1-2 and different feature orders in lines 3-4 to control the generation process. Closer to Real is better.

Metrics. For each video, participants are required to listen to two music pieces and compare them from several aspects as [39]: (1)Music Melody: the richness of the musical melody; (2)Music Rhythm: the structure consistency of rhythm; (3) Content Correspondence: the correspondence between music and video content; (4) Rhythm Correspondence: the correspondence between music and video rhythm; (5) Overall Preference. Besides, the experts are asked to evaluate two extra metrics related to music theory: (6) Chord Quality: the quality, composition and degree of harmony of generated chords; (7) Accompaniment Quality: the richness and quality of the generated accompaniment.

Results. The result is provided in Tab. 4, showing the preference rate of Diff-BGM against CMT (the percentage of participants who consider music generated by Diff-BGM better than CMT). It shows that in all metrics and user groups, Diff-BGM outperforms CMT(>50%), indicating that Diff-BGM generates higher-quality music and better understands video-music correspondence. We also include preference scores of more models to compare as shown in Tab. 5. It can be observed that in every aspect, Diff-BGM is the closest to artificial(line 1). Note the *Human*-created music score is only 3.5. It reflects how much improvement is needed to get human-level creation.

5.4. Ablation Studies

We conduct ablation studies on different components of our Diff-BGM as shown in Tab. 2. Unconditional means that we use the baseline diffusion model to generate music for each given video. For we do not add any conditions or restrictions, the generation results have the highest diversity(6.421). However, for the lack of control signal and temporal alignment, the quality and correspondence are not so good. Then we add video feature and feature selector(row 5-6) to the base model. When adding more signals to control the generation process, the quality of the generated music keeps improving and the diversity keeps decreasing. The metric of PCHE has a gain of 0.468, indicating a more clear melody. Besides, with the introduction of video feature, the video-music correspondence score P@20 has a gain of 11.27, indicating that the music contains information from the video. In the last row, segment-aware cross-attention layers are added to the model, which focuses on the alignment between music and video and improves the retrieval score. However, when we force the music to pay attention to only short-term context of the video, the music quality decreases. The results indicate that the quality of music and its correlation with video mutually influence each other when we aim to exert control over the music.

Besides, we conduct an ablation study on the feature selector as shown in Tab. 6. In the first two rows, we only use features from one modality (either video dynamic or semantics) to control the generation process. We find that when only using language features as condition, the results gain the highest GPS marks (0.641), which means that the structure of generated music is closest to real one and captions facilitate the generation of musical structures. And in the last two rows, we attempt to use different feature orders to control the generation at different stages. The results indicate that early-stage usage of video semantic features followed by dynamic features yields the best music quality, aligning with the viewpoint that the model generates melody first and then produces rhythm.

6. Conclusion

In this paper, we propose the Diff-BGM framework to tackle the video background music generation task and new metrics to measure video-music correspondence and music diversity. We also provide a high-quality dataset, BGM909, comprising temporally and semantically aligned video-music pairs and fine-grained annotations for shots and semantics. We address the issue of poor interpretability in existing generative models by using different features to control various stages of music generation. We introduce segment-aware cross-attention to temporally align music and video and generate music corresponding to video content and rhythm. Experiments verify Diff-BGM's capability of generating high-quality background music for videos.

References

[1] 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 Christian Havnø Frank. Musiclm: Generating music from text. *ArXiv*, abs/2301.11325, 2023. 1

[2] K. Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. *ArXiv*, abs/2308.01546, 2023. 1, 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 6

[4] Shangzhe Di, Zeren Jiang, Sihan Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2, 6, 7, 8

[5] Hao-Wen Dong, K. Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick. Multitrack music transformer. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2022. 1, 2

[6] Seth* Forsgren and Hayk* Martiros. Riffusion - Stable diffusion for real-time music generation. 2022. 7, 8

[7] Chuang Gan, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision*, 2020. 1, 2

[8] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the maestro dataset. *ArXiv*, abs/1810.12247, 2018. 2, 3

[9] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 4, 6

[10] Sungeun Hong, Woobin Im, and Hyun S. Yang. Content-based video-music retrieval using soft intra-modal structure constraint. *arXiv: Computer Vision and Pattern Recognition*, 2017. 2, 3

[11] 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. In *AAAI Conference on Artificial Intelligence*, 2021. 1, 2

[12] Jaeyong Kang, Soujanya Poria, and Dorien Herremans. Video2music: Suitable music generation from videos using an affective multimodal transformer model. *ArXiv*, abs/2311.00968, 2023. 3

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[14] Daniel J. Levitin. *This is Your Brain on Music: The Science of a Human Obsession*. Dutton Penguin, 2006. 6

[15] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21:522–535, 2016. 2, 3

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 3, 6

[17] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13381–13392, 2021. 2, 3

[18] Ang Lv, Xuejiao Tan, Peiling Lu, Wei Ye, Shikun Zhang, Jiang Bian, and Rui Yan. Getmusic: Generating any music tracks with a unified representation and diffusion framework. *ArXiv*, abs/2305.10841, 2023. 1

[19] Kinyugo Maina. Msanii: High fidelity music synthesis on a shoestring budget. *ArXiv*, abs/2301.06468, 2023. 2

[20] Giorgio Mariani, Irene Tallini, Emilian Postolache, Michele Mancusi, Luca Di Cosmo, and Emanuele Rodolà. Multi-source diffusion models for simultaneous music generation and separation. *ArXiv*, abs/2302.02257, 2023.

[21] Lejun Min, Junyan Jiang, Gus G. Xia, and Jingwei Zhao. Polyffusion: A diffusion model for polyphonic score generation with internal and external controls. *ArXiv*, abs/2307.10304, 2023. 2, 4, 6

[22] Pedro Neves, José Fornari, and João Batista Florindo. Generating music with sentiment using transformer-gans. In *International Society for Music Information Retrieval Conference*, 2022. 1, 2

[23] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. *ArXiv*, abs/2303.11306, 2023. 5

[24] Matthias Plasser, Silvan David Peter, and Gerhard Widmer. Discrete diffusion probabilistic models for symbolic music generation. *ArXiv*, abs/2305.09489, 2023. 1

[25] Manos Plitsis, Theodoros Kouzelis, Georgios Paraskevopoulos, Vassilis Katsouros, and Yannis Panagakis. Investigating personalization methods in text to music generation. *ArXiv*, abs/2309.11140, 2023. 1, 2

[26] Jordi Pons and Xavier Serra. musicnn: Pre-trained convolutional neural networks for music audio tagging. *ArXiv*, abs/1909.06654, 2019. 6

[27] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 4, 5, 6

[28] Ludan Ruan, Y. Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10219–10228, 2022. 7

[29] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Scholkopf. Moûsai: Text-to-music generation with long-context latent diffusion. 2023. 1, 2

[30] Ziyu Wang, K. Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Xianbin Gu, and Gus G. Xia. Pop909: A pop-song dataset for music arrangement generation. In *International Society for Music Information Retrieval Conference*, 2020. 2, 3

[31] Shih-Lun Wu and Yi-Hsuan Yang. The jazz transformer on the front line: Exploring the shortcomings of ai-composed music through quantitative measures. In *International Society for Music Information Retrieval Conference*, 2020. 6, 7

[32] Shih-Lun Wu and Yi-Hsuan Yang. Musemorphose: Full-song and fine-grained piano music style transfer with one transformer vae. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1953–1967, 2021. 1, 2

[33] Hu Xu, Gargi Ghosh, Po-Yao (Bernie) Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Conference on Empirical Methods in Natural Language Processing*, 2021. 6

[34] Botao Yu, Peiling Lu, Rui Wang, Wei Hu, Xu Tan, Wei Ye, Shikun Zhang, Tao Qin, and Tie-Yan Liu. Museformer: Transformer with fine- and coarse-grained attention for music generation. *ArXiv*, abs/2210.10349, 2022. 1, 2

[35] Jiashuo Yu, Yaohui Wang, Xinyuan Chen, Xiao Sun, and Y. Qiao. Long-term rhythmic video soundtracker. *ArXiv*, abs/2305.01319, 2023. 1, 2

[36] 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. 1, 2

[37] 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. 2, 3

[38] Ye Zhu, Yuehua Wu, Kyle Olszewski, Jian Ren, S. Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal music and image generation. In *International Conference on Learning Representations*, 2022. 1, 2

[39] Le Zhuo, Zhaokai Wang, Baisen Wang, Yue Liao, Chenxi Bao, Stanley Peng, Songhao Han, Aixi Zhang, Fei Fang, and Si Liu. Video background music generation: Dataset, method and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15637–15647, 2023. 2, 3, 6, 7, 8