

Diff-V2M: A Hierarchical Conditional Diffusion Model with Explicit Rhythmic Modeling for Video-to-Music Generation

Shulei Ji^{1,2}, Zihao Wang^{1,3}, Jiaxing Yu¹, Xiangyuan Yang⁴,
Shuyu Li¹, Songruoyao Wu¹, Kejun Zhang^{1,2*}

¹Zhejiang University

²Innovation Center of Yangtze River Delta, Zhejiang University

³Carnegie Mellon University

⁴Xi'an Jiaotong University

{shuleiji, carlwang, yujx}@zju.edu.cn, ouyang_xy@stu.xjtu.edu.cn, {lsyxary, wsry, zhangkejun}@zju.edu.cn

Abstract

Video-to-music (V2M) generation aims to create music that aligns with visual content. However, two main challenges persist in existing methods: (1) the lack of explicit rhythm modeling hinders audiovisual temporal alignments; (2) effectively integrating various visual features to condition music generation remains non-trivial. To address these issues, we propose Diff-V2M, a general V2M framework based on a hierarchical conditional diffusion model, comprising two core components: visual feature extraction and conditional music generation. For rhythm modeling, we begin by evaluating several rhythmic representations, including low-resolution mel-spectrograms, tempograms, and onset detection functions (ODF), and devise a rhythmic predictor to infer them directly from videos. To ensure contextual and affective coherence, we also extract semantic and emotional features. All features are incorporated into the generator via a hierarchical cross-attention mechanism, where emotional features shape the affective tone via the first layer, while semantic and rhythmic features are fused in the second cross-attention layer. To enhance feature integration, we introduce timestep-aware fusion strategies, including feature-wise linear modulation (FiLM) and weighted fusion, allowing the model to adaptively balance semantic and rhythmic cues throughout the diffusion process. Extensive experiments identify low-resolution ODF as a more effective signal for modeling musical rhythm and demonstrate that Diff-V2M outperforms existing models on both in-domain and out-of-domain datasets, achieving state-of-the-art performance in terms of objective metrics and subjective comparisons. Demo and code are available at <https://Tayjs197.github.io/Diff-V2M-Demo/>.

Introduction

Music not only activates the auditory system but also modulates visual perception through its functional connections with the visual cortex (Koelsch 2014). As such, background music serves as a critical element in enhancing the overall impact and expressiveness of videos. However, traditional background music composition often relies on manual editing or customized production, which is both costly and inflexible. With the rapid rise of video streaming platforms

like YouTube and TikTok, alongside the emergence of video generative models such as Sora (OpenAI 2024) and Veo (Google DeepMind 2024), the demand for personalized audiovisual content has surged. In this context, video-to-music generation has emerged as a rapidly growing research topic.

In recent years, video-to-music generation has attracted increasing research attention, enabling background music creation tailored to diverse video domains (Ji et al. 2025; Wang et al. 2025). Early works primarily targeted **human-centric videos**, such as silent instrument performances and dance clips. Studies on instrument performance videos (Koepke et al. 2020; Su, Liu, and Shlizerman 2020; Gan et al. 2020) generated music by modeling visual cues from performers, while works on dance videos (Su, Liu, and Shlizerman 2021; Zhu et al. 2022, 2023) usually extracted human motion features to guide music generation. Moving beyond human-centric videos, more recent research (Di et al. 2021; Su et al. 2024; Tian et al. 2025b; Zuo et al. 2025) expanded to **general videos** such as music videos and movie trailers by extracting multi-perspective visual features and incorporating multi-condition guidance to steer music generation. In parallel, some studies (Zhang and Fuentes 2025; Li et al. 2025a) incorporated video understanding into large language models (LLMs), translating videos into textual prompts that condition text-to-music generation pipelines.

Despite recent advances, existing video-to-music generation methods lack explicit modeling of musical rhythm, which is crucial for achieving precise audiovisual temporal alignment. Existing approaches model visual dynamics through scene detection (Kang, Poria, and Herremans 2024), optical flow (Di et al. 2021), frame differences (Zhuo et al. 2023; Kang, Poria, and Herremans 2024), or frame-level visual features (Su et al. 2024; Tian et al. 2025b). However, these strategies still require the model to implicitly learn the mapping from visual dynamics to musical rhythm. Others (Zhang and Fuentes 2025; Li et al. 2025a) attempt to translate videos into textual prompts, which often fail to preserve fine-grained temporal dynamics. There remains a lack of a unified and effective musical rhythmic representation that can support consistent temporal alignment in general video-to-music generation.

The other key challenge lies in how to effectively integrate diverse visual features to guide music generation. The fu-

*Corresponding author.

sion of multi-perspective features from videos, such as emotional, semantic, and rhythmic features, remains non-trivial. In prior studies, progressive fusion strategies (Zhuo et al. 2023; Liang et al. 2024) often involve multi-stage architectures that increase computational overhead, while simple concatenation (Kang, Poria, and Herremans 2024; Tan et al. 2023) fails to capture the underlying dependencies between features. Alternatively, large language models (LLMs) have been leveraged to convert video content into text-based conditions, thereby simplifying inputs and bypassing explicit feature fusion. However, textual descriptions struggle to capture dynamic visual cues, limiting the temporal alignment between video and generated music.

To address the aforementioned challenges, we propose **Diff-V2M**, a hierarchical conditional diffusion transformer framework designed for general video-to-music generation. Inspired by TiVA (Wang et al. 2024), which uses low-resolution mel-spectrograms as audio layouts to support temporal synchronization, we systematically explore and compare several rhythmic representations, including low-resolution mel-spectrograms, tempograms, and onset detection functions (ODF) (Bello et al. 2005). To ensure robust rhythm conditioning, a rhythm predictor is trained to infer rhythmic representations from video and is jointly optimized with the music generator during training following the proposed scheduled conditioning training strategies. In addition to rhythmic features, Diff-V2M extracts color histograms (Zhuo et al. 2023; Afifi, Brubaker, and Brown 2021) as emotional features and CLIP features (Radford et al. 2021) as semantic cues, enabling emotional and semantic alignment between video and music. Moreover, to condition the generator on these three features, we design a hierarchical conditional module. Specifically, emotional features are first integrated through a cross-attention layer to guide the overall affective tone. Subsequently, semantic and rhythmic features are processed independently via parallel cross-attention and adaptively fused using a set of timestep-aware fusion strategies, including feature-wise linear modulation (FiLM) (Perez et al. 2018) and weighted fusion.

The main contributions of this paper are as follows:

- We introduce three rhythmic representations to model temporal alignment for video-to-music generation and identify low-resolution ODF as the most effective.
- We propose Diff-V2M, a conditional diffusion transformer framework tailored for the general video-to-music generation task. It integrates emotional, semantic, and rhythmic features via hierarchical cross-attention, enhanced by timestep-aware FiLM and weighted fusion strategies for effective multi-feature conditioning.
- Extensive experiments demonstrate that Diff-V2M outperforms the state-of-the-art models in terms of objective and subjective evaluation on both in-domain and out-of-domain datasets.

Related Work

Visual Understanding

Video-to-music generation leverages video understanding models to extract diverse visual features, including emo-

tional, semantic, and rhythmic features. Semantic features are typically obtained using pretrained models such as CLIP (Radford et al. 2021), VideoCLIP (Xu et al. 2021), ViViT (Arnab et al. 2021), and VideoMAE (Tong et al. 2022). Rhythmic features are derived from motion dynamics using models like GCN (Gan et al. 2020; Liang et al. 2024), TCN (Pedersoli and Goto 2020), and I3D (Zhu et al. 2022, 2023; Carreira and Zisserman 2017), as well as handcrafted approaches based on optical flow (Di et al. 2021) and frame differences (Zhuo et al. 2023; Kang, Poria, and Herremans 2024). Emotional features are commonly represented by frame-level color histograms (Zhuo et al. 2023) or CLIP-based emotion probability distribution vectors (Kang, Poria, and Herremans 2024). In this paper, we use CLIP to extract frame-wise semantic features and model visual emotion using color histograms. To explicitly model musical rhythm, we introduce a rhythm predictor that estimates rhythmic representations directly from video.

Music Generation

Music generation can be broadly categorized into symbolic and audio-domain approaches (Briot, Hadjeres, and Pachet 2017; Ji, Luo, and Yang 2020). For symbolic music generation, models such as Transformers, Variational Autoencoders (VAEs), and Generative Adversarial Networks (GANs), along with their variants, have been widely adopted (Ji, Yang, and Luo 2023). As music generation has evolved from unimodal to cross-modal tasks (Li et al. 2025b), such as text-to-music and vision-to-music, audio generation has gained increasing popularity due to its enhanced expressive capacity and the relative ease of collecting large-scale datasets. Autoregressive models like MusicLM (Agostinelli et al. 2023) and MusicGen (Copet et al. 2023), as well as latent diffusion models (LDMs) such as AudioLDM (Liu et al. 2023) and Stable Audio (Evans et al. 2025), have achieved notable success in text-to-music generation. These models provide a strong foundation for audio-based cross-modal music generation. In this paper, we adopt an audio LDM as the backbone and advance it with a hierarchical conditioning mechanism that incorporates emotional, semantic, and rhythmic features from videos.

Video-to-Music Generation

Video-to-music generation can be broadly categorized by video type into human-centric videos (e.g., dance videos) and general videos (e.g., movie clips). Early studies on human-centric videos focused on silent music performance (Koepke et al. 2020; Su, Liu, and Shlizerman 2020; Gan et al. 2020), while recent studies have extensively explored dance-to-music generation by leveraging motion or keypoint-based features to control rhythm and style (Su, Liu, and Shlizerman 2021; Zhu et al. 2022, 2023; You et al. 2024; Liang et al. 2024; Tan et al. 2023; Li et al. 2024a; Yu et al. 2023). For general videos, methods typically extract diverse visual features (e.g., emotional, semantic, and rhythmic features) to guide music generation (Di et al. 2021; Su et al. 2024; Zhuo et al. 2023; Kang, Poria, and Herremans 2024; Li et al. 2024b; Tian et al. 2025a). Additionally, some approaches employ textual prompts or video captions as extra

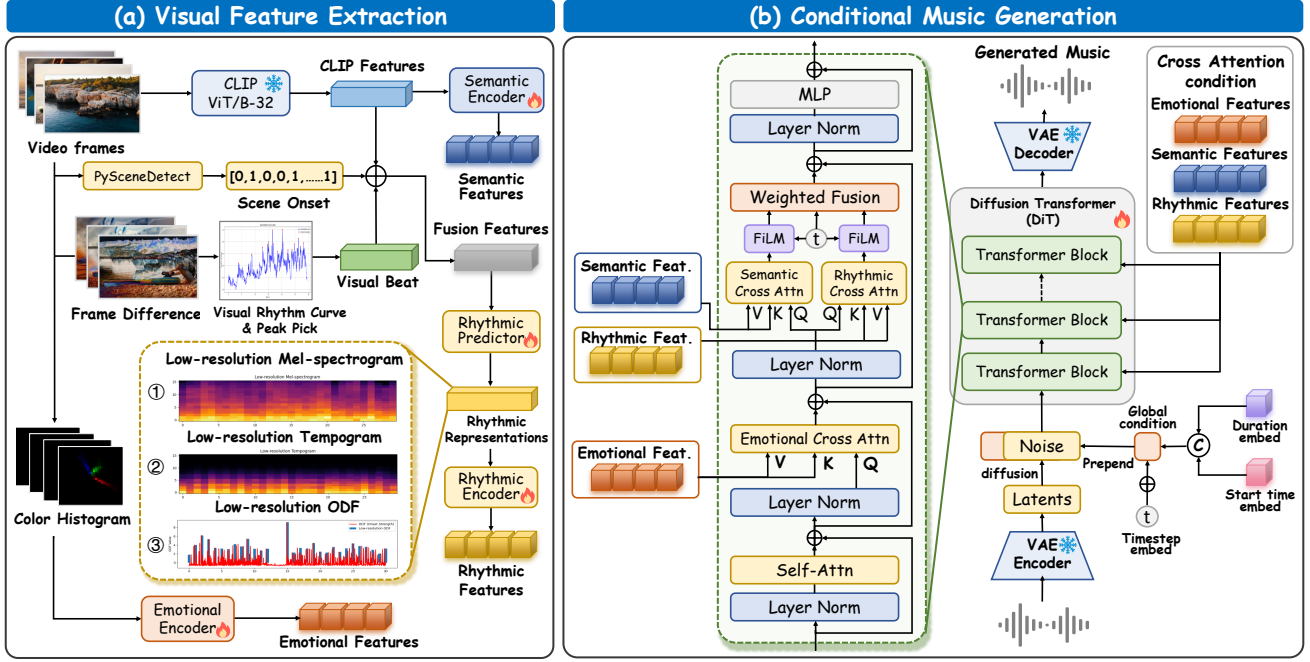


Figure 1: The architecture of Diff-V2M, consisting of two core modules: (a) visual feature extraction that derives emotional, semantic, and rhythmic features; and (b) conditional music generation built on a DiT-based LDM, which integrates multi-view features via hierarchical cross attention and timestep-aware fusion strategies.

high-level control signals (Su et al. 2024; Li et al. 2024b), while others use large language models (LLMs) to convert visual inputs into textual prompts for text-to-music generation (Zhang and Fuentes 2025; Li et al. 2025a; Liu et al. 2024). Despite these advances, existing approaches lack explicit modeling of musical rhythm and effective conditioning mechanisms for multiple visual features. Consequently, we propose a novel framework capable of generating music for diverse general videos by predicting generalizable rhythmic representations and integrating multiple video-driven features through a hierarchical conditioning module.

Methodology

Generalizable Rhythmic Representations

Low-resolution mel-spectrograms have been proven effective for temporal control in video-to-sound effect generation (Wang et al. 2024). Motivated by this, we explore their effectiveness in video-to-music generation for the first time. In addition, we investigate tempograms and onset detection functions (ODF) as alternative rhythmic representations. To facilitate learning and improving efficiency, all representations are dimensionally reduced, as detailed below. An illustration of the three types is provided in Figure 1(a).

Low-resolution Mel-spectrogram are an effective control signal for coarse-to-fine audio generation (Wang et al. 2024). Given a raw Mel-spectrogram Mel_{raw} of size $[M_{\text{raw}}, C_{\text{raw}}]$, we normalize and downsample it to a low-resolution version Mel_{LR} with target resolution $[M, C]$:

$$\text{Mel}_{LR} = \text{Resize}(\text{Norm}(\text{Mel}_{\text{raw}}); M, C) \quad (1)$$

where M_{raw} and C_{raw} denote the original number of frames and frequency dimensions, and M and C are their reduced counterparts.

Low-resolution Tempogram. A tempogram is a time-tempo representation that captures the local tempo of an audio signal as it evolves over time. Following the same strategy as for Mel_{LR} , we normalize and downsample the raw tempogram $\text{Tem}_{\text{raw}} \in \mathbb{R}^{M_{\text{raw}} \times B_{\text{raw}}}$ to obtain a compact form $\text{Tem}_{LR} \in \mathbb{R}^{M \times B}$:

$$\text{Tem}_{LR} = \text{Resize}(\text{Norm}(\text{Tem}_{\text{raw}}); M, B) \quad (2)$$

where M_{raw} and B_{raw} denote the original number of frames and tempo bins, and M and B are their reduced counterparts. This compact representation preserves the overall tempo contour while simplifying model learning.

Low-resolution ODF. The onset detection function (ODF) converts audio into a one-dimensional time series that reflects the likelihood or intensity of note onsets over time. Compared to mel-spectrograms and tempograms, ODF provides cleaner rhythmic cues by emphasizing critical rhythmic events. Given a raw ODF curve $\mathbf{o} = [o_1, o_2, \dots, o_T]$, where T is the number of audio frames, we apply peak detection to identify onset peaks $\mathcal{P} = \{(t_i, o_{t_i})\}_{i=1}^N$, where each t_i is the time (in seconds) of a detected peak and o_{t_i} is the corresponding onset strength. We then map each detected peak to its nearest second and construct a second-level vector as the low-resolution ODF, i.e., $\text{ODF}_{LR} = [o_1, o_2, \dots, o_M]$, where M is the total number of seconds. For each second, we keep o_m the maximum onset strength if any peak exists, otherwise set $o_m = 0$.

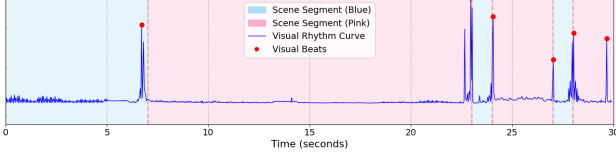


Figure 2: An example illustrating explicit video scene transitions, the visual rhythm curve, and the visual beats.

Architecture of Diff-V2M

As illustrated in Figure 1, **Diff-V2M** consists of two main modules: visual feature extraction and conditional music generation. The feature extraction module includes a **semantic encoder**, a **rhythmic encoder**, an **emotional encoder**, and a **rhythmic predictor**. The generation module features an LDM-based **music generator**, which employs hierarchical cross-attention and feature fusion mechanisms for multi-feature conditioning. Each component is detailed in the follow-up.

Visual Feature Extraction and Encoding. Following prior work (Zhuo et al. 2023), we adopt frame-wise color histograms (Afifi, Brubaker, and Brown 2021) to capture the underlying emotion of videos. Frame-level semantic features are obtained using a pretrained CLIP model (Radford et al. 2021), while rhythmic features are obtained from one of the three representations introduced in Section . To match the input dimensions of the generator’s conditioning module, each feature is projected through a dedicated encoder composed of linear layers.

Rhythmic Predictor. To infer rhythmic features without relying on audio at inference, we introduce a decoder-only transformer as the rhythmic predictor that takes as input: (i) CLIP features, (ii) scene transition embeddings, and (iii) visual beat vectors.

To capture macro-level visual changes, scene transitions are detected via PySceneDetect (Castellano 2018), yielding a binary vector $\mathbf{e} = [e_1, e_2, \dots, e_M] \in \{0, 1\}^M$ that marks scene boundaries per second, where M is the video length in seconds and $e_m = 1$ indicates the start of a new scene at second m , and $e_m = 0$ otherwise. For fine-grained visual dynamics, frame-wise differences are aggregated over time to form a visual rhythm curve. Peaks are detected to obtain a second-level visual beat vector $\mathbf{v} = [v_1, v_2, \dots, v_M] \in \mathbb{R}^M$, where v_m denotes the peak beat intensity around the second m , or zero if no peak is detected. Figure 2 provides an example illustrating explicit video scene transitions, the visual rhythm curve, and the visual beats selected based on peak detection. These two vectors provide complementary rhythmic cues. Although differing in granularity, they often correlate in high-activity scenes.

To align dimensions, $\mathbf{e}$ is passed through an embedding layer $\text{Embed}(\cdot)$ and $\mathbf{v}$ through a linear projection $\text{Linear}(\cdot)$. The resulting vectors are summed with the frame-level CLIP features C_s to form the input sequence:

$$\mathbf{X} = C_s + \text{Embed}(\mathbf{e}) + \text{Linear}(\mathbf{v}) \quad (3)$$

The resulting sequence $\mathbf{X}$ is fed into the rhythmic predictor to estimate the target rhythmic representations introduced in Section , enabling audio-free rhythm prediction from visual input at inference time.

DiT-based Conditional Music Generator. We adapt Stable Audio Open (Evans et al. 2025), an LDM-based audio generator, for video-to-music generation. A VAE encodes raw waveforms into latent representations $\mathbf{z}_a$, enabling efficient generation. A conditional diffusion model G is trained to predict the added noise ϵ from the noisy latents $\mathbf{z}_a^t$, conditions $\mathbf{C}$, and diffusion timestep t :

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t, \mathbf{z}_a^0, \epsilon} \left[\|\epsilon - G(\mathbf{z}_a^t, \mathbf{C}, t)\|_2^2 \right], \quad (4)$$

In this paper, condition $\mathbf{C}$ includes emotional, semantic, and rhythmic features extracted from video, along with global embeddings (i.e., music start time and duration). A diffusion transformer equipped with a hierarchical cross-attention module is devised for integrating multiple video-driven features.

Hierarchical Cross-attention Module. As illustrated in Figure 1(b), the hierarchical cross-attention module first incorporates emotional features to shape the overall mood of the generated music. Next, semantic and rhythmic features are attended to via parallel cross-attention layers, preventing information entanglement and enabling more precise capture of content-relevant and tempo-aligned cues. Besides, timestep-aware feature fusion strategies are adopted to adaptively combine semantic and rhythmic branches, which will be elaborated later. This hierarchical design enables flexible integration of emotional tone, semantic meaning, and rhythmic structure for enhanced music generation.

Feature Fusion Strategy. Prior study (Li et al. 2024b) proposes a feature selector that enforces exclusive attention to either semantic or dynamic features at each timestep, disregarding complementary information from other features. This rigid selection limits the model’s ability to leverage multiple features in complex video scenarios, especially when both semantic and rhythmic information are crucial. To overcome this, we introduce novel fusion strategies that adaptively balance semantic and rhythmic contributions for improved multi-feature integration. Two fusion methods are designed: weighted fusion and FiLM-based fusion.

(1) Weighted fusion. A gating network conditioned on the diffusion timestep t outputs a scalar weight $\alpha \in [0, 1]$ to balance semantic features $\mathbf{h}_{\text{sem}}$ and rhythmic features $\mathbf{h}_{\text{rhy}}$. The fused feature is computed as:

$$\alpha = \sigma(f_{\text{gate}}(t)), \quad \mathbf{h}_{\text{fuse}} = \alpha \cdot \mathbf{h}_{\text{sem}} + (1 - \alpha) \cdot \mathbf{h}_{\text{rhy}} \quad (5)$$

where $f_{\text{gate}}(\cdot)$ denotes the gating network and $\sigma(\cdot)$ is the sigmoid function.

(2) FiLM-based fusion. To enhance fine-grained feature modulation beyond weighted fusion, we propose a Feature-wise Linear Modulation (FiLM)-based fusion mechanism. FiLM applies learnable, timestep-dependent scaling and shifting to each feature dimension, allowing precise and dimension-wise adjustment over semantic and rhythmic features. For each input feature $\mathbf{h} \in \mathbb{R}^{B \times T \times D}$, two MLP

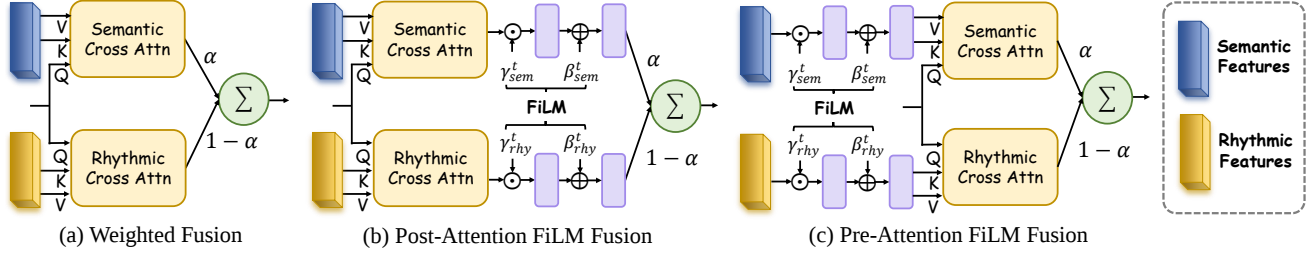


Figure 3: The illustration of different fusion strategies for semantic and rhythmic features.

networks generate timestep-aware modulation parameters $\gamma^t, \beta^t \in \mathbb{R}^{B \times 1 \times D}$, which are then applied as follows:

$$\begin{aligned} \text{FiLM}_{\text{sem}}(\mathbf{h}_{\text{sem}}) &= \gamma_{\text{sem}}^t \cdot \mathbf{h}_{\text{sem}} + \beta_{\text{sem}}^t \\ \text{FiLM}_{\text{rhy}}(\mathbf{h}_{\text{rhy}}) &= \gamma_{\text{rhy}}^t \cdot \mathbf{h}_{\text{rhy}} + \beta_{\text{rhy}}^t \end{aligned} \quad (6)$$

Furthermore, we investigate the optimal position for applying the above fusion strategy, as shown in Figure 3. Three design variants are explored:

- **Weighted fusion (Figure 3(a)):** Semantic and rhythmic attention outputs are first computed independently and then combined via a timestep-aware weighted fusion (Eq. (5)). When $\alpha = 0.5$, this fusion becomes *additive fusion* of two features. When $\alpha = 1$ or 0, the fusion degrades to *single feature selection* (Li et al. 2024b).
- **Post-attention FiLM fusion (Figure 3(b)):** Each attention output is individually modulated via FiLM, followed by weighted fusion.
- **Pre-attention FiLM fusion (Figure 3(c)):** FiLM is applied to the semantic and rhythmic features prior to parallel cross-attention, and the resulting attention outputs are then combined via weighted fusion.

This combination of hierarchical conditioning module and feature fusion design facilitates flexible and fine-grained interaction between features, allowing the model to effectively leverage complementary cues in diverse video-to-music generation scenarios.

Training with Scheduled Conditioning

To mitigate the training-inference discrepancy caused by using ground-truth rhythmic representations C_{r}^{gt} during training and predicted representations $C_{\text{r}}^{\text{pred}}$ during inference, we adopt a scheduled conditioning strategy that gradually substitutes ground-truth rhythmic representations with predicted ones during training. Specifically, we define a probability schedule $p_{\text{pred}}(e) \in [0, 1]$ to control the use of predicted rhythmic representations at epoch e :

$$p_{\text{pred}}(e) = \begin{cases} 0, & \text{if } e < e_1 \\ \frac{e - e_1}{e_2 - e_1}, & \text{if } e_1 \leq e < e_2 \\ 1, & \text{if } e \geq e_2 \end{cases} \quad (7)$$

where $e_1 = 10$ and $e_2 = 30$ in our setup. At each epoch e , a Bernoulli variable $q \sim \text{Bernoulli}(p_{\text{pred}}(e))$ determines whether to use predicted or ground-truth rhythmic representation. This ensures a smooth transition from teacher-forced

Dataset	Training	Validation	Test
BGM909	8510	1074	1061
SymMV	9898	1260	1245
V2M-Bench	0	0	1426

Table 1: The statistical distribution of the adopted datasets

training to fully relying on the predicted rhythms, ensuring robustness at inference time when only predicted features are available. This training strategy is inspired by *Scheduled Sampling* (Bengio et al. 2015), but differs in that the replaced variable is a conditioning signal rather than an autoregressive input. Last but not least, the rhythmic predictor is trained jointly with the generator, ensuring co-adaptation and better alignment with generation objectives.

Experiments

Datasets

We employ BGM909 (Li et al. 2024b) and SymMV (Zhuo et al. 2023) datasets for training our models. BGM909 is built upon the POP909 dataset (Wang et al. 2020), which includes 909 piano arrangements of Chinese pop songs accompanied by temporally aligned music videos. SymMV is a large-scale dataset curated from YouTube, comprising 1,181 video-music pairs across more than 10 genres, with a total duration of 78.9 hours.

We preprocess all datasets by removing vocals and normalizing audio loudness. Silent segments longer than 3 seconds are discarded. The remaining data is segmented into clips up to 30 seconds with a 10-second hop size. Both datasets are split into training, validation, and test sets with an 8:1:1 ratio. In addition, we include V2M-Bench (Tian et al. 2025b) as an out-of-domain test set to evaluate the generalization performance. V2M-Bench contains 300 video-music pairs with a total duration of 9 hours, covering a diverse range of genres including movie trailers, ads, documentaries, and vlogs. The statistics of the processed datasets are shown in Table 1.

Implementation Details

Video frames and audio are sampled at 1 FPS and 44.1 kHz, respectively. Rhythmic features have shape $[M, d]$, where M denotes the video duration in seconds, $d = 16$ when using low-resolution Mel-spectrograms or tempograms, and $d = 1$ for low-resolution ODF. Diff-V2M uses a frozen VAE from

Stable Audio Open (Evans et al. 2025), while the DiT-based diffusion model is trained from scratch. The DiT is optimized with the v-objective (Salimans and Ho 2022) to predict noise. During inference, we use a 250-step DDIM sampler with classifier-free guidance (scale 3.0). Training employs AdamW optimizer with a learning rate of 1×10^{-4} , betas (0.9, 0.999), weight decay 1×10^{-3} , and an InverseLR scheduler (power 0.5). All models are trained for 50 epochs on 2 NVIDIA A100 GPUs.

Evaluation Metrics

Following the state-of-the-art method (Tian et al. 2025b), we quantitatively evaluate our model using several metrics that assess the fidelity and diversity of the generated music, including **Frechet Audio Distance (FAD)**, **Frechet Distance (FD)**, **Kullback–Leibler divergence (KL)** (Liu et al. 2023), **Density (Den.)** and **Coverage (Cov.)** (Naeem et al. 2020). We also use the **ImageBind Score (IB)** (Girdhar et al. 2023) to assess alignment between video and generated music. For subjective evaluation, we consider four criteria (Tian et al. 2025b), i.e., (1) **Audio Quality**: perceptual clarity and fidelity of the audio; (2) **Musicality**: the aesthetic quality of the music, distinct from audio quality; (3) **Video-Music Alignment**: how well the music matches the visuals; and (4) **Overall Assessment**: overall generation quality.

Comparison Models

We compare Diff-V2M with the following methods:

- **CMT** (Di et al. 2021) establishes the rhythmic relationships between video and background music, then proposes a Controllable Music Transformer (CMT) for local rhythmic and global genre/instrument control.
- **Video2Music** (Kang, Poria, and Herremans 2024) extracts semantic, scene offset, motion, and emotion features from music videos and proposes an Affective Multimodal Transformer (AMT) to generate music given a video.
- **MuMu-LLaMA** (Liu et al. 2024) combines ViViT and LLaMA (Touvron et al. 2023) with multimodal adapters, then projects audio tokens from LLaMA as conditions for text-to-music generation.
- **GVMGen** (Zuo et al. 2025) encodes audio into discrete tokens using EnCodec (Défossez et al. 2023) and predicts discrete audio tokens with hierarchical spatial and temporal cross-attention to align visual features with music.
- **VidMuse** (Tian et al. 2025b) predicts the discrete audio tokens by incorporating local and global visual cues, and employs long-short-term modeling to ensure coherence between the video and music.

Note that the first two approaches generate symbolic music, while the others directly generate musical audio.

Experimental Results

Comparison of Rhythmic Representations. Table 2 compares the performance of different rhythmic representations, evaluated on both the mixed test set of two datasets and the V2M-Bench dataset. The low-resolution

Models	Metrics					
	FAD↓	FD↓	KL↓	Den.↑	Cov.↑	IB↑
Mixed Test Set						
Mel _{LR}	2.0376	13.1304	0.9341	0.3690	0.3380	0.1653
Tem _{LR}	2.3163	13.0197	0.9551	0.3210	0.3270	0.1636
ODF _{LR}	2.0175	12.8566	0.9432	0.3694	0.3370	0.1831
V2M-Bench						
Mel _{LR}	2.0310	24.8375	1.2638	0.6916	0.3822	0.1810
Tem _{LR}	2.0230	20.2481	1.2939	0.5170	0.4137	0.1748
ODF _{LR}	1.8129	21.3321	1.2405	0.6360	0.3717	0.1887

Table 2: The comparison of different rhythmic representations. Mel_{LR}, Tem_{LR}, and ODF_{LR} refer to low-resolution Mel-spectrogram, tempogram and ODF, respectively. The best results are highlighted in **bold**.

Strategies	Metrics					
	FAD↓	FD↓	KL↓	Den.↑	Cov.↑	IB↑
Weighted Fusion	2.3625	11.5332	0.9246	0.3106	0.3540	0.1729
Additive Fusion	2.0175	12.8566	0.9432	0.3694	0.3370	0.1831
Feature Selection	2.0894	12.4990	0.8875	0.3228	0.3720	0.1800
PreAttn FiLM	2.0812	13.9032	0.9507	0.3192	0.3500	0.1640
PostAttn FiLM	2.1340	12.1286	0.9052	0.3682	0.3510	0.1808
w/ FS	1.5175	10.9567	0.8575	0.3756	0.3990	0.1812

Table 3: The comparison of different feature fusion strategies on the mixed test set. FS denotes feature selection.

ODF (ODF_{LR}), offering a simpler and more direct rhythmic representation, consistently outperforms other representations. Low-resolution Mel-spectrogram (Mel_{LR}), despite prior use in video-to-sound effect task (Wang et al. 2024), is less effective here. Thus, ODF_{LR} is used as the default rhythmic representation in subsequent experiments. Note that simple additive fusion is employed in this comparison.

Comparison of Feature Fusion Strategies. The comparison results of the fusion strategies for the semantic and rhythmic features within the hierarchical conditioning module are presented in Table 3. Note that the weighted fusion following FiLM employs a simple additive fusion strategy. Among post-attention strategies, both feature selection and post-attention FiLM outperform weighted fusion and perform on par with additive fusion. Pre-attention FiLM is less effective than its post-attention counterpart. The best performance is achieved by combining post-attention FiLM with feature selection.

Quantitative Comparisons with Other Methods. Quantitative results in Table 4 show that the proposed Diff-V2M significantly outperforms prior methods on the in-domain test sets. On the out-of-domain V2M-Bench dataset, Diff-V2M also achieves the best overall performance, particularly in audio quality. However, it lags slightly behind GVMGen in video-music alignment, likely because GVMGen was trained on larger, more diverse video datasets similar to V2M-Bench. Additionally, Figure 4 reports average inference time over 10 runs for generating 30-second music clips. CMT and Video2Music are faster as they generate simpler symbolic music. While audio-based models are slower, Diff-V2M achieves the shortest inference time among them.

Models	Metrics					
	FAD↓	FD↓	KL↓	Den.↑	Cov.↑	IB↑
Mixed Test Set						
GT	0.0000	0.0000	0.0000	1.1178	0.9980	0.2154
CMT	8.9265	47.7599	1.0984	0.0422	0.0080	0.0820
Video2Music	31.1963	106.7122	1.7220	0.0010	0.0010	0.0312
MuMu-LLaMA	<u>2.8410</u>	27.1160	1.2481	0.1074	0.0900	0.1448
GVMGen	4.3354	27.9350	1.1633	0.0900	0.0830	0.1760
VidMuse	3.4376	21.0391	0.9361	0.1496	0.1300	0.1804
Diff-V2M (ours)	1.5175	10.9567	0.8575	0.3756	0.3990	0.1812
V2M-Bench						
GT	0.0000	0.0000	0.0000	1.0632	0.9930	0.2379
CMT	10.9118	65.5007	1.4109	0.0895	0.0189	0.1255
Video2Music	32.4392	128.9256	2.1149	0.0094	0.0007	0.0388
MuMu-LLaMA	3.8593	40.4072	1.4866	0.3637	0.1971	0.1714
GVMGen	<u>2.1528</u>	21.5488	1.2060	0.3853	0.2938	0.2030
VidMuse	2.5875	22.0282	1.0138	0.2181	0.2020	0.1963
Diff-V2M (ours)	1.7612	<u>22.0164</u>	1.2684	0.5083	0.4032	<u>0.1967</u>

Table 4: Comparison with existing methods. The best results are highlighted in **bold**, and the second-best are underlined.

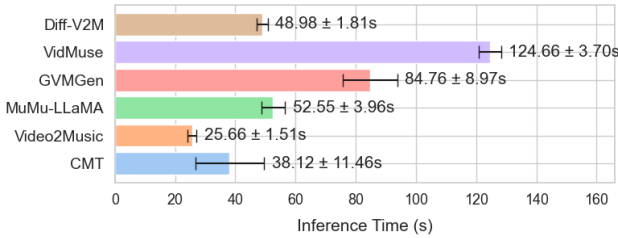


Figure 4: The comparison of inference time for different methods in generating soundtracks for 30-second videos.

Subjective Evaluation. We conducted an A/B test to subjectively compare different methods. A total of 30 participants (20 amateurs and 10 experts) were invited to ensure that each pair of methods was compared 20 times. Participants were instructed to choose their preferred sample based on the subjective criteria. The evaluation results are presented in Figure 5, where the value at position matrix[i][j] indicates the percentage (0–100) that the method in row i was preferred over that in column j . Diff-V2M outperforms all baselines in over half of the pairwise comparisons, except when compared with ground-truth (GT) samples. These results highlight the subjective superiority of Diff-V2M.

Ablation Studies. We further conduct ablation studies on Diff-V2M’s training strategies: (1) w/o C_r , remove rhythmic features; (2) w/o C_e , remove emotional features; (3) w/o C_r & C_e : remove both; (4) w/o Visual Rhythm (VR): the rhythm predictor takes only CLIP features as input, excluding dynamic visual inputs including scene onset and visual beat; (5) w/o Joint: train the rhythm predictor and music generator separately, leading to a training–inference mismatch; (6) w/o Scheduler: jointly train the predictor and generator from the beginning without applying scheduled conditioning. Table 5 shows that Diff-V2M achieves the best performance by incorporating rhythmic and emotional features as well as adopting the scheduled training strategy. Removing VR degrades performance, likely due to reduced rhythm prediction accuracy. Interestingly, excluding C_r slightly improves ImageBind (IB) score, possibly because IB empha-

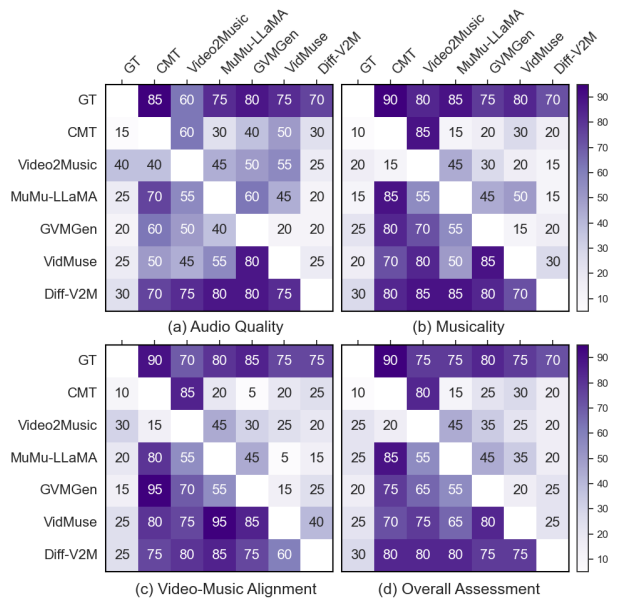


Figure 5: A/B test results of the subjective comparisons.

Ablation	Metrics					
	FAD↓	FD↓	KL↓	Den.↑	Cov.↑	IB↑
Diff-V2M	1.5175	10.9567	0.8575	0.3756	0.3990	0.1812
w/o C_r	1.8264	11.9535	<u>0.8692</u>	0.3476	0.3610	0.1893
w/o C_e	1.6761	12.8922	0.9306	0.3664	0.3480	0.1814
w/o C_r & C_e	1.7086	9.7514	0.8775	0.3636	0.3900	0.1814
w/o VR	2.2177	13.6060	0.9147	0.3220	0.3300	0.1800
w/o Joint	1.8787	13.3894	0.8743	0.3798	0.3880	0.1814
w/o Scheduler	<u>1.6159</u>	<u>10.6706</u>	0.9105	0.3402	0.3630	<u>0.1860</u>

Table 5: Ablation results of different training strategies for Diff-V2M on the mixed test set.

sizes semantic alignment, which may benefit from the absence of rhythm-related interference during generation.

Conclusion

We propose Diff-V2M, a general video-to-music generation framework designed to handle diverse videos. To explicitly model musical rhythm, we introduce a simple yet effective rhythmic representation and develop a predictor to estimate it from video. Diff-V2M builds upon an audio LDM, featuring a hierarchical cross-attention conditioning module to integrate multiple video-derived features, along with novel fusion strategies to adaptively combine semantic and rhythmic cues. Extensive comparisons on both mixed in-domain and out-of-domain test sets demonstrate the superiority of Diff-V2M in both quantitative and qualitative evaluations.

Limitation. Despite promising performance, our method has limitations. First, relying on scene cuts and inter-frame differences may overlook subtle motion cues, leading to sub-optimal rhythm alignment in human-centric videos. Second, the model lacks explicit control over musical attributes such as genre and emotion, limiting its adaptability in scenarios requiring style or affective manipulation. These limitations highlight valuable directions for future research.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62272409).

References

- Afifi, M.; Brubaker, M. A.; and Brown, M. S. 2021. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7941–7950.
- Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzett, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.
- Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.
- Bello, J. P.; Daudet, L.; Abdallah, S.; Duxbury, C.; Davies, M.; and Sandler, M. B. 2005. A tutorial on onset detection in music signals. *IEEE Transactions on speech and audio processing*, 13(5): 1035–1047.
- Bengio, S.; Vinyals, O.; Jaitly, N.; and Shazeer, N. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28.
- Briot, J.-P.; Hadjeres, G.; and Pachet, F.-D. 2017. Deep learning techniques for music generation—a survey. *arXiv preprint arXiv:1709.01620*.
- Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.
- Castellano, B. 2018. Pyscenedetect: Intelligent scene cut detection and video splitting tool.
- Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36: 47704–47720.
- Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2023. High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research*, 2023.
- Di, S.; Jiang, Z.; Liu, S.; Wang, Z.; Zhu, L.; He, Z.; Liu, H.; and Yan, S. 2021. Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2037–2045.
- Evans, Z.; Parker, J. D.; Carr, C.; Zukowski, Z.; Taylor, J.; and Pons, J. 2025. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Gan, C.; Huang, D.; Chen, P.; Tenenbaum, J. B.; and Torralba, A. 2020. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision*, 758–775. Springer.
- Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15180–15190.
- Google DeepMind. 2024. Veo. <https://deepmind.google/models/veo>. Accessed: July 30, 2025.
- Ji, S.; Luo, J.; and Yang, X. 2020. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801*.
- Ji, S.; Wu, S.; Wang, Z.; Li, S.; and Zhang, K. 2025. A Comprehensive Survey on Generative AI for Video-to-Music Generation. *arXiv preprint arXiv:2502.12489*.
- Ji, S.; Yang, X.; and Luo, J. 2023. A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Computing Surveys*, 56(1): 1–39.
- Kang, J.; Poria, S.; and Herremans, D. 2024. Video2music: Suitable music generation from videos using an affective multimodal transformer model. *Expert Systems with Applications*, 249: 123640.
- Koelsch, S. 2014. Brain correlates of music-evoked emotions. *Nature reviews neuroscience*, 15(3): 170–180.
- Koepke, A. S.; Wiles, O.; Moses, Y.; and Zisserman, A. 2020. Sight to sound: An end-to-end approach for visual piano transcription. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1838–1842. IEEE.
- Li, J.; Xu, T.; Chen, X.; Yao, X.; Han, J.; and Liu, S. 2025a. Mozart’s Touch: a lightweight multimodal music generation framework based on pre-trained large models. In *International Conference on AI-Generated Content (AIGC 2024)*, volume 13649, 198–207. SPIE.
- Li, S.; Dong, W.; Zhang, Y.; Tang, F.; Ma, C.; Deussen, O.; Lee, T.-Y.; and Xu, C. 2024a. Dance-to-music generation with encoder-based textual inversion. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.
- Li, S.; Ji, S.; Wang, Z.; Wu, S.; Yu, J.; and Zhang, K. 2025b. A Survey on Music Generation from Single-Modal, Cross-Modal, and Multi-Modal Perspectives. *arXiv preprint arXiv:2504.00837*.
- Li, S.; Qin, Y.; Zheng, M.; Jin, X.; and Liu, Y. 2024b. Diffbgm: A diffusion model for video background music generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27348–27357.
- Liang, X.; Li, W.; Huang, L.; and Gao, C. 2024. DanceComposer: dance-to-music generation using a progressive conditional music generator. *IEEE Transactions on Multimedia*, 26: 10237–10250.
- Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D. P.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 21450–21474. PMLR.

- Liu, S.; Hussain, A. S.; Wu, Q.; Sun, C.; and Shan, Y. 2024. Mumu-llama: Multi-modal music understanding and generation via large language models. *arXiv preprint arXiv:2412.06660*, 3(5): 6.
- Naeem, M. F.; Oh, S. J.; Uh, Y.; Choi, Y.; and Yoo, J. 2020. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, 7176–7185. PMLR.
- OpenAI. 2024. Sora. <https://openai.com/sora>. Accessed: July 30, 2025.
- Pedersoli, F.; and Goto, M. 2020. Dance Beat Tracking from Visual Information Alone. In *ISMIR*, 400–408.
- Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Salimans, T.; and Ho, J. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Su, K.; Li, J. Y.; Huang, Q.; Kuzmin, D.; Lee, J.; Donahue, C.; Sha, F.; Jansen, A.; Wang, Y.; Verzetti, M.; et al. 2024. V2meow: Meowing to the visual beat via video-to-music generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4952–4960.
- Su, K.; Liu, X.; and Shlizerman, E. 2020. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems*, 33: 3325–3337.
- Su, K.; Liu, X.; and Shlizerman, E. 2021. How does it sound? *Advances in Neural Information Processing Systems*, 34: 29258–29273.
- Tan, V.; Nam, J.; Nam, J.; and Noh, J. 2023. Motion to dance music generation using latent diffusion model. In *SIGGRAPH Asia 2023 Technical Communications*, 1–4.
- Tian, S.; Zhang, C.; Yuan, W.; Tan, W.; and Zhu, W. 2025a. XMusci: Towards a Generalized and Controllable Symbolic Music Generation Framework. *IEEE Transactions on Multimedia*.
- Tian, Z.; Liu, Z.; Yuan, R.; Pan, J.; Liu, Q.; Tan, X.; Chen, Q.; Xue, W.; and Guo, Y. 2025b. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18782–18793.
- Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Video-mae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Wang, X.; Wang, Y.; Wu, Y.; Song, R.; Tan, X.; Chen, Z.; Xu, H.; and Sui, G. 2024. Tiva: Time-aligned video-to-audio generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 573–582.
- Wang, Z.; Bao, C.; Zhuo, L.; Han, J.; Yue, Y.; Tang, Y.; Huang, V. S.-J.; and Liao, Y. 2025. A Survey on Vision-to-Music Generation: Methods, Datasets, Evaluation, and Challenges. In *ISMIR 2025 Hybrid Conference*.
- Wang, Z.; Chen, K.; Jiang, J.; Zhang, Y.; Xu, M.; Dai, S.; and Xia, G. 2020. POP909: A Pop-Song Dataset for Music Arrangement Generation. In *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, 38–45.
- Xu, H.; Ghosh, G.; Huang, P.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 6787–6800. Association for Computational Linguistics.
- You, F.; Fang, M.; Tang, L.; Huang, R.; Wang, Y.; and Zhao, Z. 2024. Momu-diffusion: On learning long-term motion-music synchronization and correspondence. *Advances in Neural Information Processing Systems*, 37: 127878–127906.
- Yu, J.; Wang, Y.; Chen, X.; Sun, X.; and Qiao, Y. 2023. Long-term rhythmic video soundtracker. In *International Conference on Machine Learning*, 40339–40353. PMLR.
- Zhang, L.; and Fuentes, M. 2025. Sonique: Video background music generation using unpaired audio-visual data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Zhu, Y.; Olszewski, K.; Wu, Y.; Achlioptas, P.; Chai, M.; Yan, Y.; and Tulyakov, S. 2022. Quantized gan for complex music generation from dance videos. In *European Conference on Computer Vision*, 182–199. Springer.
- Zhu, Y.; Wu, Y.; Olszewski, K.; Ren, J.; Tulyakov, S.; and Yan, Y. 2023. Discrete Contrastive Diffusion for Cross-Modal Music and Image Generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Zhuo, L.; Wang, Z.; Wang, B.; Liao, Y.; Bao, C.; Peng, S.; Han, S.; Zhang, A.; Fang, F.; and Liu, S. 2023. Video background music generation: Dataset, method and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15637–15647.
- Zuo, H.; You, W.; Wu, J.; Ren, S.; Chen, P.; Zhou, M.; Lu, Y.; and Sun, L. 2025. Gvmgen: A general video-to-music generation model with hierarchical attentions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23099–23107.

Appendix

Architecture of Diff-V2M

The training flow of Diff-V2M is shown in Algorithm 1.

Algorithm 1: Diff-V2M

Input: Video clip V
Output: Generated musical audio A
Waveform Encoding:
Latent representation $z = \text{VAE Encoder}(A)$
Visual Feature Extraction:
Extract semantic features: $C_s = \text{CLIP}(V)$
Extract emotional features: $C_e = \text{ColorHist}(V)$
Extract visual rhythm: $C_{VR} = \text{VisualDynamics}(V)$
Rhythm Prediction:
Predict rhythm features:
 $C_r = \text{RhythmPredictor}(C_s + C_{VR})$
Apply scheduled conditioning:
Sample $q \sim \text{Bernoulli}(p_{\text{pred}}(e))$ at epoch e
if $q = 1$ **then**
 use C_r
else
 use ground-truth rhythm C_r^{gt}
Hierarchical Conditional Diffusion:
Add noise to encoded latent: $z^t = \text{AddNoise}(z)$
for $\text{timestep } t = 1$ **to** 0 **do**
 Global Conditioning:
Concatenate global metadata: $C_g = [g_{\text{start}}; g_{\text{dur}}]$
Compute time-modulated global embedding:
 $C_g^t = \text{MLP}([C_g + \text{Embed}(t)])$
Prepend C_g^t to z and apply RoPE
 $\tilde{z} = \text{RoPE}([C_g^t; z^t])$
 for $\text{Block } l = 1$ **to** N **do**
 Cross Attention 1 (emotional condition)
 $\mathbf{h}_{\text{emo}}^{(l)} \leftarrow \text{CrossAttn}(\mathbf{h}_{\text{self}}^{(l)}, \mathbf{K}_e, \mathbf{V}_e)$
 Cross Attention 2 (semantic and rhythmic conditions)
 $\mathbf{h}_{\text{sem}}^{(l)} = \text{CrossAttn}(\mathbf{h}_{\text{emo}}^{(l)}, \mathbf{K}_s, \mathbf{V}_s)$
 $\mathbf{h}_{\text{rhy}}^{(l)} = \text{CrossAttn}(\mathbf{h}_{\text{emo}}^{(l)}, \mathbf{K}_r, \mathbf{V}_r)$
 Feature fusion followed by FFN
 $\mathbf{h}^{(l+1)} \leftarrow \text{FFN}(\text{Fuse}(\mathbf{h}_{\text{sem}}^{(l)}, \mathbf{h}_{\text{rhy}}^{(l)}, t))$
 Denoise latent: $z \leftarrow \text{DiT}_{\theta}(z^t, \mathbf{C}, t)$
Waveform Reconstruction:
Generate audio $A = \text{VAE Decoder}(z)$
return A

DiT-based Conditional Music Generator In this paper, conditions $\mathbf{C}$ include features extracted by specialized encoders, i.e., emotional, semantic, and rhythmic features, along with global embeddings including music start time g_{start} and duration g_{dur} . These two embeddings are concatenated into a single vector and added to a learnable timestep embedding corresponding to the diffusion step t . The time-modulated global embedding is then prepended as a special

token to the input sequence. To enhance temporal awareness within the attention mechanism, we further adopt rotary positional embeddings (RoPE), dynamically generating the position encoding matrix based on the current sequence length. The enhanced sequence representation is subsequently fed into the diffusion transformer.

Hierarchical Cross-attention Module. Let $\mathbf{h}^{(l)}$ be the input sequence at the l -th Transformer block. After the self-attention module, the emotional context is first integrated as:

$$\begin{aligned} \mathbf{h}_{\text{self}}^{(l)} &= \text{SelfAttn}(\mathbf{h}^{(l)}) \\ \mathbf{h}_{\text{emo}}^{(l)} &= \text{CrossAttn}(\mathbf{h}_{\text{self}}^{(l)}, \mathbf{K}_e, \mathbf{V}_e). \end{aligned} \quad (8)$$

where $\mathbf{K}_e$ and $\mathbf{V}_e$ are obtained through linear projections of emotional features C_e . Using $\mathbf{h}_{\text{emo}}^{(l)}$ as the updated query, we compute cross attention with semantic and rhythmic features in parallel:

$$\begin{aligned} \mathbf{h}_{\text{sem}}^{(l)} &= \text{CrossAttn}(\mathbf{h}_{\text{emo}}^{(l)}, \mathbf{K}_s, \mathbf{V}_s), \\ \mathbf{h}_{\text{rhy}}^{(l)} &= \text{CrossAttn}(\mathbf{h}_{\text{emo}}^{(l)}, \mathbf{K}_r, \mathbf{V}_r). \end{aligned} \quad (9)$$

where $\mathbf{K}_s$ and $\mathbf{V}_s$, and $\mathbf{K}_r$ and $\mathbf{V}_r$, are obtained through linear projections of the semantic features C_s and the rhythmic features C_r , respectively. Timestep-aware feature fusion modules then adaptively combine both branches before the final feed-forward network (FFN), yielding the $(l+1)$ -th input sequence:

$$\mathbf{h}^{(l+1)} = \text{FFN}(\text{Fuse}(\mathbf{h}_{\text{sem}}^{(l)}, \mathbf{h}_{\text{rhy}}^{(l)}, t)). \quad (10)$$

Feature Selection According to Diff-BGM (Li et al. 2024b), 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. Based on this, a feature selector is devised to enforce exclusive attention to a single feature at each time step, neglecting the complementary information from other features. This rigid selection limits the model’s ability to leverage multiple features in complex video scenarios, especially when both semantic and rhythmic information are crucial.

Diff-BGM selects conditioning features during the denoising process from timestep N to 0 based on a hyper-parameter t_0 . Specifically, language features are used when $t_0 > 200$, and video features when $t_0 \leq 200$. In this paper, we follow a similar strategy. Since our model adopts a continuous noise schedule with timesteps normalized to $t \in [0, 1]$, we select semantic features when $t_0 > 0.2$, and rhythmic features when $t_0 \leq 0.2$.

Algorithm. The overall training workflow of Diff-V2M is presented in **Algorithm 1**. Note that the self-attention processes within the transformer block are omitted, as follows:

$$\mathbf{h}_{\text{self}}^{(l)} = \begin{cases} \text{SelfAttn}(\tilde{z}), & \text{if } l = 1 \\ \text{SelfAttn}(\mathbf{h}^{(l)}), & \text{if } l > 1 \end{cases} \quad (11)$$

Strategies	Metrics					
	FAD↓	FD↓	KL↓	Den.↑	Cov.↑	IB↑
Weighted Fusion	2.3050	18.5891	1.3059	0.5196	0.4299	0.1764
Additive Fusion	1.8129	21.3321	1.2405	0.6360	0.3719	0.1887
Feature Selection	1.9527	25.7266	1.2601	0.5149	0.3864	0.1867
PreAttn FiLM	2.3738	23.7385	1.2626	0.7844	0.3941	0.1703
PostAttn FiLM	1.9276	20.6385	1.2583	0.6196	0.3717	0.1923
w/ FS	1.7612	22.0164	1.2684	0.5083	0.4032	0.1967

Table 6: The comparison of different feature fusion strategies on the V2M-Bench dataset. FS denotes feature selection. The best results are highlighted in **bold**.

Ablation	Metrics					
	FAD↓	FD↓	KL↓	Den.↑	Cov.↑	IB↑
Diff-V2M	1.7612	22.0164	1.2684	0.5083	0.4032	0.1967
w/o C_r	<u>1.7671</u>	22.5636	1.2595	0.5613	0.3822	0.1913
w/o C_e	1.9417	24.5074	1.3307	0.4659	0.3576	<u>0.1959</u>
w/o C_r & C_e	1.7797	<u>21.2935</u>	<u>1.2345</u>	0.6144	<u>0.4004</u>	0.1906
w/o VR	2.0195	22.4100	1.2864	0.4300	0.3892	0.1861
w/o Joint	3.3245	30.2149	1.2305	<u>0.5798</u>	0.2854	0.1753
w/o Scheduler	1.7706	19.9180	1.3037	0.4861	0.3850	0.1910

Table 7: Ablation results of different training strategies for Diff-V2M on V2M-Bench dataset. The best results are highlighted in **bold**, and the second-best are underlined.

Experiments

Comparison of Inference Time. We compare the average inference time over 10 runs for generating 30-second music clips using different methods. All models were evaluated on a single A10 GPU (20 GB), except for MuMu-LLaMA, which was executed on an A100 GPU (80 GB) due to its higher memory requirements. Note that GVMGen is limited to generating background music for videos of up to 25 seconds in duration.

Comparison of Feature Fusion Strategies. Table 6 presents a comparison of feature fusion strategies on the V2M-Bench dataset. Among them, post-attention FiLM combined with feature selection achieves the best overall performance in terms of audio quality and video-music alignment, while weighted fusion yields higher diversity in generated music. Since V2M-Bench serves as an out-of-domain dataset, the best scores are distributed across different strategies rather than concentrated in a model.

Ablation Studies. Ablation study results on the V2M-Bench dataset are shown in Table 7. Consistent with the findings on the in-domain mixed test sets, Diff-V2M achieves the best performance by incorporating rhythmic and emotional features and leveraging the scheduled conditioning training strategy.