Video Echoed in Music: Semantic, Temporal, and Rhythmic Alignment for Video-to-Music Generation

Xinyi Tong,[1,2,3] **Yiran Zhu,**[3] **Jishang Chen,**[1,2,3] **Chunru Zhan,**[3] **Tianle Wang,**[1,2] **Sirui Zhang,**[1,2] **Nian Liu,**[2] **Tiezheng Ge**[3], **Duo Xu**[2], **Xin Jin**[2], **Feng Yu**[1], **Song-Chun Zhu**[2,4*]

[1]Central Conservatory of Music, Beijing, China
[2]Beijing Institute for General Artificial Intelligence, Beijing, China
[3]Alibaba Group, Beijing, China
[4]Peking University, Beijing, China
tongxinyi@mail.ccom.edu.cn, jinxinbesti@foxmail.com, s.c.zhu@pku.edu.cn

Abstract

Video-to-Music generation seeks to generate musically appropriate background music that enhances audiovisual immersion for videos. However, current approaches suffer from two critical limitations: 1) incomplete representation of video details, leading to weak alignment, and 2) inadequate temporal and rhythmic correspondence, particularly in achieving precise beat synchronization. To address the challenges, we propose **V**ideo **E**choed in **M**usic (VeM), a latent music diffusion that generates high-quality soundtracks with semantic, temporal, and rhythmic alignment for input videos. To capture video details comprehensively, VeM employs a hierarchical video parsing that acts as a music conductor, orchestrating multi-level information across modalities. Modality-specific encoders, coupled with a storyboard-guided cross-attention mechanism (SG-CAtt), integrate semantic cues while maintaining temporal coherence through position and duration encoding. For rhythmic precision, the frame-level transition-beat aligner and adapter (TB-As) dynamically synchronize visual scene transitions with music beats. We further contribute a novel video-music paired dataset sourced from e-commerce advertisements and video-sharing platforms, which imposes stricter transition-beat synchronization requirements. Meanwhile, we introduce novel metrics tailored to the task. Experimental results demonstrate superiority, particularly in semantic relevance and rhythmic precision.

Demo&Code — https://vem-paper.github.io/VeM-page/

Introduction

Music, akin to video, evokes sensory perception and emotional responses. This intrinsic relationship underscores the integration to enhance the audiovisual experience. However, music pieces raise copyright issues and manual composition is time-consuming. Thus, Video-to-Music(V2M) generation presents a promising solution with applications in film, advertising, gaming, and short-form video production.

The V2M task aims to generate background music that exhibits semantic, temporal, and rhythmic alignment with the given video. This involves three critical aspects: 1) **High fidelity** ensures that music is indistinguishable from human-composed pieces, serving as a fundamental benchmark for music generation. 2) **Semantic alignment**, whereby music accurately reflects thematic, emotional, and narrative elements in videos. 3)**Temporal synchronization** emphasizes alignment with temporal dynamics. **Rhythmic consistency**, as a distinctive dimension of temporal alignment, accentuates junctures by synchronizing video transitions with music beats, ensuring transition-beat matching.

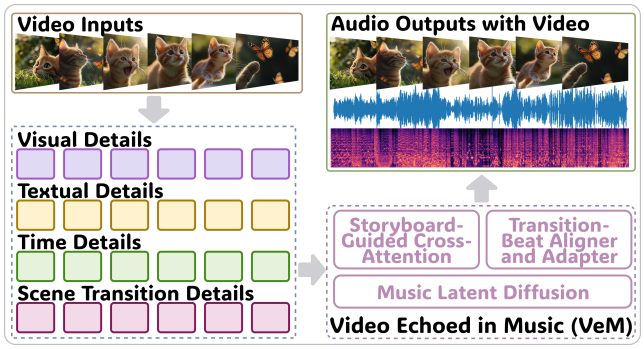

Figure 1: **Task overview.** The proposed latent music diffusion, VeM, achieves video-music alignment by integrating multimodal details from videos as conditions.

Recent research has advanced in these areas. 1) For music quality, some focus on symbolic representations to meet human-composed standards, but audio synthesis with engines restricts timbral diversity (Di et al. 2021; Zhuo et al. 2023; Xie et al. 2025). More efforts (Agostinelli et al. 2023; Copet et al. 2023) shift towards waveform directly, facilitating superior auditory feedback, which we have also adopted. 2) Existing semantic alignment methods fall into two broad categories. The first employs rule-based or learnable visual features to guide generation (Yu et al. 2023; Li et al. 2024c; Xie et al. 2025; Tian et al. 2025a). However, the features provide coarse video understanding, potentially imposing insufficient constraints. Although MuVi (Li et al. 2024a) and VidMusician (Li et al. 2024d) advance with visual adapters, they neglect global semantic invariance over time. The second category leverages visual-language models to extract textual

*Corresponding authors.

descriptions (Tian et al. 2024; Tong et al. 2024; Wang et al. 2024a; Zhou et al. 2025), reducing the task to text-to-music and largely bypassing visual features. The inherent limitations of text hinder temporal details, leading to poor synchronization. 3) For temporal synchronization, recent methods employ local semantics to involve temporal variations through video clips (Li et al. 2024c; Zuo et al. 2025) or textual timestamps (Zhang and Fuentes 2025; Zhou et al. 2025), but typically overlook fine-grained temporal details. More works focus on rhythmic consistency by aligning partial visual dynamics with musical rhythms, including optical flow (Di et al. 2021; Kang, Poria, and Herremans 2024), visual embedding variation (Zhuo et al. 2023; Lin et al. 2024; Li et al. 2024d; Xie et al. 2025), and human-centric motion (Zhu et al. 2022; Li et al. 2024b; You et al. 2024). These specific dynamics fail to explicitly capture the rhythmic cues.

The most pertinent research, Video-to-Audio, generating sound effects from videos, also emphasizes temporal consistency (Ruan et al. 2023; Liu et al. 2024; Luo et al. 2024; Xing et al. 2024; Wang et al. 2024b; Rong et al. 2025). However, applying the strategy directly to music presents challenges. Sound effects align with discrete visual events, whereas music exhibits intrinsic rhythmic periodicity with recurring beats, requiring longer alignment spans and smoother transitions. Crucially, salient video transitions typically coincide with music beats; arbitrary deviations can disrupt the rhythmic flow and lead to discordance.

In this paper, we propose **V**ideo **e**choed in **M**usic (VeM), a diffusion-based framework to achieve semantic, temporal, and rhythmic alignment for V2M generation. We provide a hierarchical video parsing, serving as a music conductor, which comprehensively orchestrates multilevel details, shown in Fig. 1. Semantic and temporal cues are integrated by a storyboard-guided cross-attention mechanism (SG-CAtt). Rhythmic precision is maintained by frame-level transition-beat aligner and adapter (TB-As), synchronizing video transitions with music beats. Meanwhile, we construct TB-Match, a video-music paired dataset collected from e-commerce advertisements and video-sharing platforms, enforcing stricter synchronization for transitions and beats. We introduce novel evaluation metrics tailored to the task. The experimental results demonstrate superiority in both semantic-temporal relevance and rhythmic precision. The main contributions are claimed as follows:

- A novel perspective that utilizes hierarchical video parsing as a music conductor to orchestrate comprehensive multimodal constraints for video-to-music generation.
- A diffusion-based framework that explicitly integrates multimodal constraints into soundtracks to achieve semantic, temporal, and rhythmic alignment.
- A video-music dataset annotated with fine-grained parsing and evaluation metrics tailored to the task. Both subjective and objective results show the superiority.

Related Works

Diffusion-Based Conditional Music Generation

Recent advances in diffusion models have demonstrated potential for conditional music generation. Riffusion (Forsgren and Martiros 2022), Noise2Music (Huang et al. 2023b), and Moûsai (Schneider et al. 2023) have pioneered open-domain text-to-music generation by diffusion models. AudioLDM2 (Liu et al. 2024) facilitates holistic audio generation, including music, through self-supervised pretraining. DITTO (Novack et al. 2024) leverages distilled diffusion inference-time T-optimization for enhanced generation. Mustango (Melechovsky et al. 2024) and Music ControlNet (Wu et al. 2024) apply various time-varying musical constraints (e.g., chords, rhythms), while MusicMagus (Zhang et al. 2024) and SteerMusic (Niu et al. 2025) explore zero-shot music editing via diffusion. These developments underscore the effectiveness of diffusion models for conditional music generation. Building upon the foundations, we present VeM that extends latent diffusion to video-to-music while retaining the controllability benefits established in conditional music generation.

Video-to-Music Generation

Current approaches for video-to-music alignment employ diverse strategies. The first method, CMT (Di et al. 2021) and subsequent approaches (Yang, Yu, and Wu 2022; Zhuo et al. 2023; Yu et al. 2023; Kang, Poria, and Herremans 2024; Qi, Ni, and Xu 2024) project disentangled visual features (RGB, saliency, motion) onto musical attributes (melody, chord, rhythm), failing to capture visual semantics. Large Language Model-based techniques (Liu et al. 2023; Xu et al. 2024; Tong et al. 2024; Wang et al. 2024a; Zhou et al. 2025) leverage textual representations. Specifically, M^2UGen (Liu et al. 2023) focuses on textual music understanding, while SONIQUE (Zhang and Fuentes 2025) extracts musical tags from unpaired data. AudioX (Tian et al. 2025b) combines visual, textual, and audio features to a multimodal condition. However, textual abstraction inherently loses fine-grained temporal dynamics. Motion-centric methods, such as V2Meow (Su et al. 2024), FilmComposer (Xie et al. 2025), and VMAS (Lin et al. 2024), achieve movement alignment but neglect broader domains. VidMuse (Tian et al. 2024) involves long-short-term temporal dependencies, but suffers from limited generative capacity. Diff-BGM (Li et al. 2024c) addresses clip-level alignment, but only partially adapts to semantic shifts. Recent approaches, MuVi (Li et al. 2024a), VidMusician (Li et al. 2024d), and GVMGen (Zuo et al. 2025), improve local semantic correspondence that involves temporal dynamics but lack explicit temporal position and duration encoding, preventing precise frame-level synchronization. Therefore, substantial opportunities remain for advancing semantic, temporal, and rhythmic alignment in video-to-music generation.

Method

This section introduces the proposed VeM, a latent music diffusion to achieve semantic, temporal, and rhythmic alignment for videos. The pipeline is illustrated in Fig. 2. Hierarchical video parsing acts as a music conductor, providing comprehensive multimodal video details that are represented by modality-specific encoders. Semantic and temporal cues are integrated via SG-CAtt. Fine-grained rhythmic precision is ensured through frame-level TB-As.

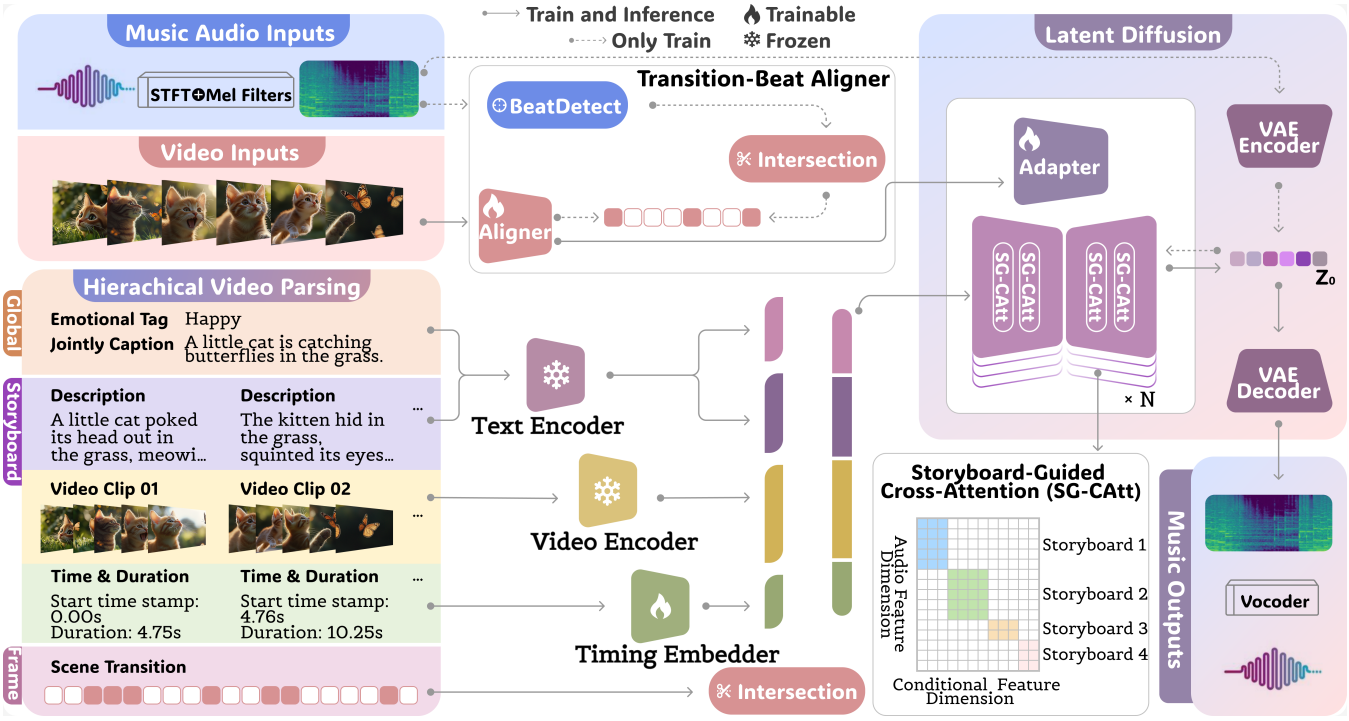

Figure 2: **Illustration of the proposed method.** The **hierarchical video parsing** provides a comprehensive analysis across three levels. Cross-modal features are captured by **modality-specific encoders**, facilitating the semantic and temporal alignment by integrating global and storyboard details into the generative latent via **storyboard-guided cross-attention**. The frame-level **transition-beat aligner and adapter** ensure precise rhythmic synchronization by coupling video scene transitions with detected music beats and adapting to the music latent.

Preliminary

Music Audio Representation. For a music waveform $x \in \mathbb{R}^{L_s}$, where L_s denotes the number of audio samples, we adopt the log Mel-spectrogram $X \in \mathbb{R}^{W \times B}$ as the training target, derived via the Short-Time Fourier Transform (STFT) and Mel-filters, due to its perceptual relevance and dimensionality reduction. W and B represent the time windows and Mel-frequency bins, respectively. A trained variational autoencoder (VAE) encodes X into a latent representation z. We subsequently train a latent diffusion to generate z by iteratively denoising from Gaussian noise ϵ. Finally, the predicted latent z is reconstructed to the Mel-spectrogram by the VAE decoder, followed by waveform synthesis via the vocoder (Kong, Kim, and Bae 2020).

Latent Music Diffusion The Latent Diffusion Model (LDM) (Rombach et al. 2022) comprises the diffusion phase and the denoising phase. The forward diffusion phase is a T-step Markov process that corrupts the input by iteratively adding noise to a standard isotropic Gaussian distribution. Given latent z_{t-1} at step $t-1$, the distribution of z_t at step $t \in 2, ..., T$ is defined as:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta}z_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

where the noise schedule hyperparameter, $\beta_t \in [0, 1]$, regulates the rate at which noise is applied to the data. By recursively substituting $q(z_t|z_{t-1})$, the formulation is derived:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\overline{\alpha}_t}z_0, (1-\overline{\alpha}_t)\epsilon \quad (2)$$

where α_t parameterizes $1-\beta_t$ and $\overline{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ represents the cumulative noise level at timestep t. $z_T \sim \mathcal{N}(0, \mathbf{I})$ indicates the final state at step T follows a standard isotropic Gaussian distribution. $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ denotes noise addition. During the reverse process, we implement a Transformer-UNet (T-UNet) architecture, which is crafted to optimize the noise estimation objective:

$$\mathcal{L} = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(0,1), t, c}[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2] \quad (3)$$

The process iteratively generates the prior z_0 according to:

$$p_\theta(z_{0:T}|c) = p(z_T) \prod_{s=t}^{T} p_\theta(z_{t-1}|z_t, c) \quad (4)$$

$$p_\theta(z_{t-1}|z_t, c) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, n, c), \sigma_t^2 \mathbf{I}) \quad (5)$$

where $\epsilon_\theta(z_t, t, c)$ is the predicted noise, μ_θ and σ_t denote parameterized mean and variance, and c stands for conditions provided to the model. During the training phase, T-UNet is optimized to learn a backward transition from the prior distribution $\mathcal{N}(0, \mathbf{I})$ to the target z, conditioned on the input c. In this paper, we structure hierarchical video representation captured by modality-specific encoders as condition signals.

Hierarchical Video parsing — Music Conductor

For comprehensive video analysis, five key elements are supposed to be determined: 1) the overarching theme, at-

mosphere, and emotional impact; 2) smooth video segmentation into coherent video shots; 3) narrative and visual compositions within each shot; 4) temporal boundaries and duration of each shot; 5) precise timing of frame-level visual changes. The five details are collectively derived from hierarchical video parsing, as depicted in Fig. 2, where segmented shots are conceptualized as storyboards and frame-level changes as scene transitions.

Hierarchical parsing operates on three levels: global, storyboard, and frame. At the global level, video captions from a video understanding model and emotional tags from a music classification model address Key 1. The storyboard level employs a video segmentation model to extract local visual features, descriptions, start timestamps, and durations, corresponding to Keys 2–4. At the frame level, a scene transition detector ensures precise transitions, enabling fine-grained rhythmic synchronization for Key 5. Details of the aforementioned models are provided in Appendix A. Since video parsing is independent of the training process, we perform it as a preprocessing annotation step, with manual correction and cleaning.

Modality-Specific video representation

To fully leverage the rich parsing details of the video, we employ modality-specific encoders for representation. Textual information is encoded using CLAP (Wu et al. 2023), a pre-trained text-audio contrastive model. Visual content is processed by MAViL (Huang et al. 2023a), which projects videos into a shared video-audio latent space. This strategy ensures consistency between textual and visual embeddings in the audio domain, containing the features of the global video caption f_t^C and the emotional tag f_t^T, the storyboard-level description $f_t^{story_i}$ for the i-th storyboard and the corresponding visual features $f_v^{story_i}$. Temporal details, including the storyboard start time $f_s^{story_i}$ and duration $f_d^{story_i}$, are encoded by a learnable continuous-time MLP operating on seconds. Frame-level scene transitions are represented by a binary timestamp indicator $f_b^{frame-v_i}$, indicating the presence or absence of a transition for each frame.

Storyboard-Guided Cross-Attention

While cross-attention mechanisms are effective for aligning condition signals with generative representations across modalities (Ruan et al. 2023; Tian et al. 2025b), existing implementations exhibit critical limitations in temporal modeling. For example, the segment-aware approach (Li et al. 2024c) involves local temporal cues, but suffers from rigid segment divisions that neglect natural semantic boundaries. Thus, we propose storyboard-guided cross-attention (SG-CAtt) that explicitly preserves semantic alignment and simultaneously ensures temporal synchronization.

To incorporate global information f_t^C and f_t^T into each individual storyboard i, we concatenate global features with storyboard-specific features:

$$f_{att}^i = \left\{ f_t^C \| f_t^T \| f_t^{story_i} \| f_v^{story_i} \| f_s^{story_i} \| f_d^{story_i} \right\} \quad (6)$$

For a video with N number of storyboards, the conditional feature is $F_{att} = \left\{ f_{att}^1, f_{att}^2, ..., f_{att}^N \right\}$ and serves as the Value and Key within cross-attention. The Query is provided by the latent representation z_t of the diffusion model (Vaswani 2017). The temporal boundaries are defined by the start time s^i and duration d^i of the storyboard. To constrain the fusion between the condition and the latent operated solely within relevant storyboards, we introduce a storyboard mask that restricts attention to the interval $[s^i, s^i + d^i]$:

$$sMask_{x,y} = \left\{ \begin{array}{ll} 1, & s^i \leq x, y < s^i + d^i \\ 0, & else \end{array} \right\} \quad (7)$$

where x and y represent the temporal indices of music latent and conditional features, respectively. As shown in Fig. 2, the mask delineates rectangular regions due to the varying sequence lengths of each storyboard. The SG-CAtt is defined as:

$$Attention(Q, K, V) = softmax(sMask \odot \frac{QK^T}{\sqrt{d_{key}}}) \cdot V \quad (8)$$

where $\odot$ denotes element-wise multiplication. Within the T-UNet architecture, the self-attention layers in the final transformer blocks at each level are replaced by the SG-CAtts. To enforce consistent guidance across T-UNet levels, we apply uniform up-sampling and down-sampling ratios, adjusting feature dimensions of the conditional mask. The SG-CAtt technique facilitates semantic alignment and temporal synchronization at the storyboard level. By concatenating global features, semantic consistency is preserved among all storyboards, while masked cross-attention targets local temporal synchronization within individual storyboard boundaries.

Transition-Beat Aligner and Adapter

To achieve precise rhythmic consistency where visual scene transitions coincide with music beats, we first introduce the transition-beat aligner. As shown in Fig. 2, frame-level video parsing provides scene transitions, denoted by the binary indicator $f_b^{frame-v_i}$, where a value of 1 signifies a transition and 0 indicates its absence. Concurrently, we apply an RNN-based beat detector (Böck et al. 2016) to generate a corresponding binary sequence $f_b^{frame-m_i}$, indicating frame-wise music beats. Both sequences operate at a consistent frame rate of 16 fps. The intersection $f_b^{frame_i} = f_b^{frame-v_i} \cap f_b^{frame-m_i}$ identifies the timestamps where visual transitions align with music beats, thereby ensuring cross-modal rhythmic consistency. To extract the aligned frame-level rhythmic features highlighted by the intersected sequence $\hat{f}_b^{frame_i}$ from visual inputs, a ResNet(2+1)D-18 model (Tran et al. 2018) is trained using binary cross-entropy (BCE) loss over N number of samples:

$$\begin{aligned} \mathcal{L}_{BCE} = -\frac{1}{N} \sum_{i=1}^{N} [& f_b^{frame_i} log(\hat{f}_b^{frame_i}) \\ & + (1 - f_b^{frame_i}) log(1 - \hat{f}_b^{frame_i})] \end{aligned} \quad (9)$$

After training, the transition-beat aligner is capable of predicting a timestamp mask indicating the presence or absence of transition-beat matches in the target music. We extract activations from the penultimate layer and interpolate them to align with the temporal resolution of the music latent

	Au.	Vd.	IS↑	FAD↓	KLD↓	CLAP↑	LB↑	tw-CLAP↑	tw-LB↑	B_{IoU}↑	TB_{IoU}↑
GroundTruth			-	-	-	0.247	0.928	0.252	0.932	1.000	0.559
CMT	×	✓	1.131	7.151	5.540	0.109	0.728	0.113	0.775	0.254	0.213
Diff-BGM	×	✓	1.173	6.940	4.870	0.112	0.781	0.109	0.792	0.227	0.261
M^2UGen	✓	×	1.211	5.902	3.350	0.158	0.892	0.163	0.893	0.307	0.331
VidMuse	✓	✓	1.206	7.437	4.210	0.102	0.704	0.103	0.718	0.335	0.352
GVMGen	✓	✓	1.227	6.137	3.210	0.212	0.899	0.219	0.917	0.465	0.357
Ours	✓	✓	**1.263**	**4.043**	**3.160**	**0.244**	**0.930**	**0.249**	**0.935**	**0.594**	**0.364**

Table 1: **Quantitative results for objective evaluation.** Comparison to five established baselines and the groundtruth with nine quantitative metrics. Here, Au. stands for the audio output capability, and Vd. indicates supporting variable-duration music.

z, which is subsequently processed by the transition-beat adapter. Although concatenation along the channel dimension is feasible, it risks overemphasizing conditional signals, potentially distorting the music latent z of the generative model. Drawing inspiration from adaptive normalization layers (AdaLNs) (Xu et al. 2019), we propose the transition-beat adapter to ensure precise alignment between generated music and designated rhythmic features. Specifically, we normalize the music feature z_i into a scale γ_i and a shift β_i based on AdaLNs with the two zero-initialized convolution layers, where γ_i and β_i are learned from the transition-beat aligner. The adaptive normalization layers are integrated into each encoder block of the U-TNet architecture, with γ_i and β_i modulating z_i by a linear projection:

$$z_i = z_i + \gamma_i \cdot z_i + \beta_i \tag{10}$$

Train and Inference

In the training phase, we first pre-train the music reconstruction VAE model and the transition-beat aligner independently (dashed lines in Fig. 2). We then freeze these components, along with the frozen text and video encoders. Subsequently, the full latent diffusion is trained with only the trainable time embedder, facilitating the model to focus on semantic and temporal details from hierarchical video representation. The transition-beat module is excluded in this stage to prioritize conditioned music generation. Finally, we integrate the pre-trained aligner into the framework and jointly optimize the adapter to refine rhythmic consistency. The training configurations are provided in Appendix C.1.

During inference, latent music diffusion receives random noise as the initial z_T. Hierarchical video parsing processes input videos to provide conditional information represented by the encoders for the generative latent diffusion. The transition-beat aligner predicts visual features correlated with transition-beat events, which are incorporated into the music latent via the adapter (Fig. 2, solid lines).

Experiments

Dataset and Settings

Dataset. We introduce TB-Match, a high-quality video-music paired dataset comprising around 18,000 samples sourced from e-commerce advertisements and video-sharing platforms. This type of video typically exhibits frequent and highly precise synchronization between scene transitions and music beats, rendering them especially suitable for studying temporal and rhythmic alignment in video-music relationships. Each pair undergoes rigorous hybrid filtering, combining automated quality control (e.g., minimum SNR of 20dB, visual-auditory rhythmic coherence, and emotional consistency) with manual expert curation to ensure strong video-music relevance. The details of the dataset can be found in Appendix B. Furthermore, we incorporate the M^2UGen (Liu et al. 2023) dataset, contributing 13,000 video-music pairs, resulting in approximately 280 hours of total training data. For evaluation, we reserve a validation set of 1,000 TB-Match samples, ensuring no overlap with training data. For the universality study, we supplement the SymMV dataset (Zhuo et al. 2023), Sora-generated silent videos (Brooks et al. 2024), and other random data.

Implementation. We leverage a pre-trained VAE and vocoder (Liu et al. 2024), fine-tuning for our specific task. Modality-specific encoders, excluding the timing embedder, are frozen during the entire training process. T-UNet architecture adheres to the configuration described in (Liu et al. 2024), employing a 1000-step diffusion process. To handle variable-length inputs, we standardize music clips to durations between 10-60 seconds. Audio signals are downsampled to 16 kHz and transformed into Mel-spectrograms using 60 frequency bins with a hop size of 256. The video inputs are processed at 16 fps.

Baseline Models. We conduct a comparative evaluation with five state-of-the-art methods: GVMGen (Zuo et al. 2025), VidMuse (Tian et al. 2024), M^2UGen (Liu et al. 2023), Diff-BGM (Li et al. 2024c) and CMT (Di et al. 2021). GVMGen employs hierarchical attentions to align spatial-temporal video-music features. VidMuse adopts long-short-term modeling to capture the temporal dependencies. M^2UGen leverages LLMs to handle cross-modal relationships. Diff-BGM addresses semantic and temporal alignment at the clip level, and the CMT adapts rhythmic features to the generated music. The output of M^2UGen is restricted to approximately 10 seconds. For fair comparison, we loop the shorter segments to match the duration of the videos. Diff-BGM and CMT produce variable-length MIDI representations, which we convert to waveform audio via high-quality synthesizers to ensure format consistency across all evaluated methods.

	Preference Rate		Preference Score			
	Top-1		MOS-Q		MOS-A	
	Expert	Non-expert	Expert	Non-expert	Expert	Non-expert
CMT(Di et al. 2021)	3.625%	2.000%	$5.622_{\pm 0.213}$	$6.139_{\pm 0.329}$	$4.680_{\pm 0.247}$	$4.924_{\pm 0.189}$
Diff-BGM(Li et al. 2024c)	2.250%	2.125%	$5.406_{\pm 0.185}$	$5.935_{\pm 0.314}$	$4.387_{\pm 0.243}$	$4.530_{\pm 0.212}$
M^2UGen(Liu et al. 2023)	5.375%	5.125%	$5.340_{\pm 0.162}$	$5.863_{\pm 0.307}$	$5.814_{\pm 0.221}$	$6.127_{\pm 0.205}$
VidMuse(Tian et al. 2024)	4.250%	2.750%	$4.767_{\pm 0.234}$	$4.992_{\pm 0.128}$	$5.467_{\pm 0.229}$	$5.270_{\pm 0.210}$
GVMGen(Zuo et al. 2025)	11.125%	10.125%	$5.418_{\pm 0.223}$	$5.693_{\pm 0.262}$	$6.467_{\pm 0.197}$	$6.374_{\pm 0.251}$
Ours	**73.375%**	**77.875%**	$\mathbf{6.892}_{\pm 0.173}$	$\mathbf{7.537}_{\pm 0.195}$	$\mathbf{7.341}_{\pm 0.174}$	$\mathbf{7.852}_{\pm 0.260}$

Table 2: **Qualitative results for subjective evaluation.** The preference rates in the Top-1 rank and the preference scores in MOS-Q and MOS-A with CI95 for expert and non-expert groups.

Objective Evaluation

Metrics. This section outlines the quantitative metrics employed to evaluate the generated music on four dimensions: musical quality, semantic alignment, temporal synchronization, and rhythmic consistency.

Music Quality. We adopt three metrics to evaluate fidelity in generation tasks (Agostinelli et al. 2023). Inception Score (IS) measures the diversity and the perceptual clarity of generated spectrograms compared to the groundtruth. Fréchet Audio Distance (FAD) quantifies the distance between the embedding distributions of generated and reference samples. Kullback-Leibler Divergence (KLD) assesses similarity by comparing probability distributions derived from activations of a pre-trained Musicnn model (Pons and Serra 2019).

Semantic Alignment. The Contrastive Language-Audio Pretraining (CLAP) score (Wu et al. 2023) quantifies the semantic alignment between audio signals and corresponding textual descriptions. To directly assess visual-audio consistency, we employ the pre-trained LanguageBind model (Zhu et al. 2024), which projects video and music into a unified textual latent space. The cosine distance between embeddings is calculated to produce the LanguageBind (LB) score.

Temporal Synchronization. The video-music semantics remain consistent over time for temporal synchronization. Since VeM explicitly captures temporal dynamics through storyboard sequences, we compute time-weighted CLAP and LB scores (tw-CLAP and tw-LB). The weight of each storyboard i is proportional to its relative duration (d_i/d_{total}, *storyboard duration / total duration*).

Rhythmic Consistency. Rhythmic consistency requires that video transitions align with music beats. Assuming that the ideal video-music pairs are well-synchronized, we introduce the Beats Intersection over Union (IoU) metric, B_{IoU}. It measures the overlap, within a specified threshold, between the number of detected beats in generated music B_{syn} and that in the groundtruth B_{gt}, defined as:

$$B_{IoU} = \frac{B_{gt} \cap B_{syn}}{B_{gt} \cup B_{syn}} \tag{11}$$

Furthermore, we present the Transitions-Beats IoU metric, TB_{IoU}, which calculates the intersection within a threshold between the video transition timestamps T_v and the music beat timestamps B_m. The temporal threshold in both the beat and the transition detectors is 0.5 seconds, and the detectors are detailed in Section 3.2. The score is defined as:

$$TB_{IoU} = \frac{T_v \cap B_m}{T_v \cup B_m} \tag{12}$$

	SymMV		Sora		Others	
	LB↑	TB_{IoU}↑	LB↑	TB_{IoU}↑	LB↑	TB_{IoU}↑
CMT	0.912	0.314	0.758	0.671	0.578	0.337
Diff	0.643	0.253	0.898	0.667	0.589	0.325
M^2U	0.925	0.296	1.029	0.725	0.885	0.332
Vid	0.787	0.312	0.982	0.785	0.670	0.400
GVM	0.910	0.260	1.084	0.814	0.887	0.391
Ours	**0.989**	**0.331**	**1.106**	**0.829**	**0.895**	**0.453**

Table 3: **Universality evaluation on other data with three quantitative metrics.** Diff, M^2U, Vid and GVM stand for Diff-BGM, M^2UGen, VidMuse and GVMGen, respectively.

Quantitative Results. Table 1 presents the comparative evaluation with five baselines in nine quantitative metrics, where the audio output (Au.) and the variable duration (Vd.) are emphasized. Our approach consistently outperforms existing methods, showcasing improvements in music quality, semantic alignment, temporal synchronization, and rhythmic consistency. VeM surpasses not only audio-based (GVMGen, VidMuse, and M^2UGen) but also MIDI-based methods (CMT and Diff-BGM), which is particularly notable for two reasons: 1) the inherent decoupling of MIDI allows the integration of fine-grained musical details during generation, and 2) the generated MIDI is converted to audio via high-quality synthesizers, effectively reducing auditory noise. Meanwhile, the time-weighted CLAP and LB scores exceed their non-weighted counterparts in our approach, demonstrating the local semantic and temporal alignment within storyboards. Overall, the proposed method exhibits superior quality, enhancing the audio-visual experience.

Universality Study. To assess universality, we conduct experiments in external domains, distinct from our training set. As shown in Table 3, VeM outperforms baselines across diverse inputs, indicating its effectiveness even in zero-shot scenarios. Partial results are presented due to space constraints, and complete results are provided in Appendix D.1.

HVP-Cond	SG-CAtt	TB-As	IS ↑	FAD↓	KLD↓	CLAP↑	LB↑	tw-CLAP↑	tw-LB↑	B_{IoU}↑	TB_{IoU}↑
×	×	×	0.823	6.692	4.714	0.180	0.624	0.188	0.625	0.221	0.197
×	×	✓	0.772	7.217	5.097	0.172	0.639	0.181	0.643	0.433	0.283
✓	✓	×	1.191	4.382	3.608	0.231	0.890	0.236	0.882	0.403	0.265
✓	×	×	1.140	5.712	3.869	0.218	0.735	0.227	0.742	0.383	0.220
✓	✓	✓	**1.263**	**4.043**	**3.160**	**0.244**	**0.930**	**0.249**	**0.935**	**0.594**	**0.364**

Table 4: **Ablation study on three components** including hierarchical video parsing conditions (HVP-Cond), storyboard-guided cross-attention (SG-CAtt), and transition-beat aligner and adapter (TB-As). Nine quantitative metrics are employed.

Subjective Evaluation

Due to the subjective nature of video-music alignment evaluation, we conduct a human study with 50 participants, divided into expert and non-expert groups. The expert group consists of 5 film production experts and 25 professional musicians. The non-experts include 20 amateur viewers. 16 video samples are involved, each with 6 variations featuring soundtracks generated by different methods. Participants watch the 6 versions in a randomized order. The preference rate is reported as the probability that a soundtrack receives the top rank (Top-1). Meanwhile, participants evaluate each video on two 10-point Likert scales (1 = worst, 10 = best) to assess music quality and video-music alignment. Results are reported as mean-opinion-scores for quality (MOS-Q) and alignment (MOS-A), along with 95% confidence intervals (CI95). The details are provided in Appendix D.2.

Qualitative Results. Table 2 presents the comprehensive subjective evaluation, demonstrating the consistent superiority of the proposed method. Specifically, VeM achieves the highest Top-1 preference rate among both expert and non-expert participants. For mean opinion scores, MOS-Q and MOS-A scores indicate superior perceived music quality and video-music alignment. The performance advantages across evaluator backgrounds underscore the effectiveness.

Ablation Study

We conduct ablation studies to analyze the contribution of each component within the proposed framework. The components include hierarchical video parsing conditions (HVP-Cond), storyboard-guided cross-attention (SG-CAtt), transition-beat aligner and adapter (TB-As). Table 4 details five ablated variants. The unconditional generation removes all conditional signals. To assess the impact of TB-As, we exclude both HVP-Cond and SG-CAtt. We further evaluate the combined influence of HVP-Cond and SG-CAtt by omitting the fine-grained rhythmic synchronization from TB-As. The effectiveness of SG-CAtt is tested by substituting it with standard cross-attention. Lastly, we present the results for the complete VeM model that incorporates all components.

The variant utilizing only TB-As (w/TB-As) achieves the highest transition-beat alignment measured by B_{IoU} and TB_{IoU}, highlighting the importance of the TB-As module for fine-grained rhythmic synchronization. Compared to the variant with only TB-As, the one incorporating both HVP-Cond and SG-CAtt (w/HVP-Cond & SG-CAtt) outperforms on the rest metrics, indicating the substantial contribution of HVP-Cond and SG-CAtt to the semantic and temporal alignment. Replacing SG-CAtt with standard cross-attention (w/HVP-Cond) results in degenerate performance, confirming the superiority of the SG-CAtt mechanism. The complete VeM demonstrates the best overall performance, validating the cumulative contribution of each component.

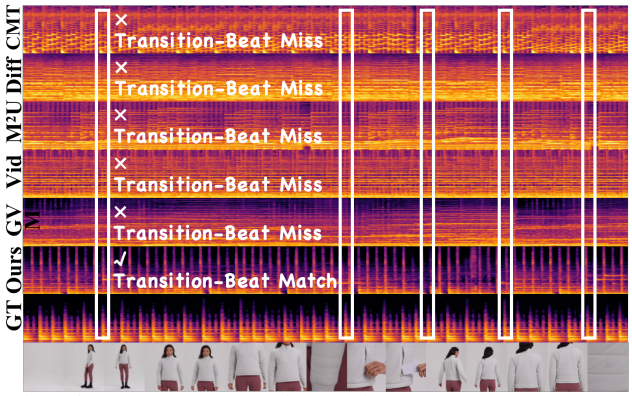

Figure 3: **Visualized comparison** shows Mel-spectrograms alongside the video frames from different methods.

Visualization of Generated Music

Fig. 3 visualizes Mel-spectrograms of audio samples alongside the video frames. Compared with baselines, VeM exhibits greater consistency with the groundtruth spectrogram, particularly in preserving temporal and rhythmic dynamics corresponding to salient visual scene transitions, highlighted by the white bounding boxes in Fig. 3.

Conclusion

In this paper, we propose VeM, a latent music diffusion to generate high-quality soundtracks semantically, temporally, and rhythmically aligned with video. VeM leverages hierarchical video parsing to comprehensively capture rich details for generation. Storyboard-guided cross-attention facilitates semantic alignment and temporal synchronization. Fine-grained rhythmic precision is achieved by the transition-beat aligner and adapter. Experimental results on a constructed video-music dataset with novel evaluation metrics showcase superior performance. Future work will explore video-integrated music editing and investigate more sophisticated alignment techniques.

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Böck, S.; Korzeniowski, F.; Schlüter, J.; Krebs, F.; and Widmer, G. 2016. Madmom: A new python audio and music signal processing library. In *Proceedings of the 24th ACM international conference on Multimedia*, 1174–1178.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. 2024. Video generation models as world simulators. *URL https://openai.com/research/video-generation-models-as-world-simulators*, 1: 8.

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 47704–47720.

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.

Forsgren, S.; and Martiros, H. 2022. Riffusion-Stable diffusion for real-time music generation. *URL https://riffusion.com/about*, 6.

Huang, P.-Y.; Sharma, V.; Xu, H.; Ryali, C.; fan, h.; Li, Y.; Li, S.-W.; Ghosh, G.; Malik, J.; and Feichtenhofer, C. 2023a. MAViL: Masked Audio-Video Learners. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 20371–20393. Curran Associates, Inc.

Huang, Q.; Park, D. S.; Wang, T.; Denk, T. I.; Ly, A.; Chen, N.; Zhang, Z.; Zhang, Z.; Yu, J.; Frank, C.; et al. 2023b. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*.

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.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33: 17022–17033.

Li, R.; Zheng, S.; Cheng, X.; Zhang, Z.; Ji, S.; and Zhao, Z. 2024a. MuVi: Video-to-Music Generation with Semantic Alignment and Rhythmic Synchronization. *arXiv preprint arXiv:2410.12957*.

Li, S.; Dong, W.; Zhang, Y.; Tang, F.; Ma, C.; Deussen, O.; Lee, T.-Y.; and Xu, C. 2024b. Dance-to-Music Generation with Encoder-based Textual Inversion. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Li, S.; Qin, Y.; Zheng, M.; Jin, X.; and Liu, Y. 2024c. Diff-BGM: A Diffusion Model for Video Background Music Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27348–27357.

Li, S.; Yang, B.; Yin, C.; Sun, C.; Zhang, Y.; Dong, W.; and Li, C. 2024d. VidMusician: Video-to-Music Generation with Semantic-Rhythmic Alignment via Hierarchical Visual Features. *arXiv preprint arXiv:2412.06296*.

Lin, Y.-B.; Tian, Y.; Yang, L.; Bertasius, G.; and Wang, H. 2024. VMAS: Video-to-Music Generation via Semantic Alignment in Web Music Videos. *arXiv preprint arXiv:2409.07450*.

Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2024. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Liu, S.; Hussain, A. S.; Sun, C.; and Shan, Y. 2023. Multi-modal Music Understanding and Generation with the Power of Large Language Models. *arXiv preprint arXiv:2311.11255*.

Luo, S.; Yan, C.; Hu, C.; and Zhao, H. 2024. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36.

Melechovsky, J.; Guo, Z.; Ghosal, D.; Majumder, N.; Herremans, D.; and Poria, S. 2024. Mustango: Toward Controllable Text-to-Music Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 8286–8309.

Niu, X.; Cheuk, K. W.; Zhang, J.; Murata, N.; Lai, C.-H.; Mancusi, M.; Choi, W.; Fabbro, G.; Liao, W.-H.; Martin, C. P.; et al. 2025. SteerMusic: Enhanced Musical Consistency for Zero-shot Text-Guided and Personalized Music Editing. *arXiv preprint arXiv:2504.10826*.

Novack, Z.; McAuley, J.; Berg-Kirkpatrick, T.; and Bryan, N. J. 2024. Ditto: Diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2401.12179*.

Pons, J.; and Serra, X. 2019. Musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv preprint arXiv:1909.06654*.

Qi, F.; Ni, L.; and Xu, C. 2024. Harmonizing Pixels and Melodies: Maestro-Guided Film Score Generation and Composition Style Transfer. *arXiv preprint arXiv:2411.07539*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rong, Y.; Wang, J.; Yang, S.; Lei, G.; and Liu, L. 2025. AudioGenie: A Training-Free Multi-Agent Framework for Diverse Multimodality-to-Multiaudio Generation. *arXiv preprint arXiv:2505.22053*.

Ruan, L.; Ma, Y.; Yang, H.; He, H.; Liu, B.; Fu, J.; Yuan, N. J.; Jin, Q.; and Guo, B. 2023. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10219–10228.

Schneider, F.; Kamal, O.; Jin, Z.; and Schölkopf, B. 2023. Moûsai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*.

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.

Tian, S.; Zhang, C.; Yuan, W.; Tan, W.; and Zhu, W. 2025a. XMusic: Towards a Generalized and Controllable Symbolic Music Generation Framework. *arXiv preprint arXiv:2501.08809*.

Tian, Z.; Jin, Y.; Liu, Z.; Yuan, R.; Tan, X.; Chen, Q.; Xue, W.; and Guo, Y. 2025b. AudioX: Diffusion Transformer for Anything-to-Audio Generation. *arXiv preprint arXiv:2503.10522*.

Tian, Z.; Liu, Z.; Yuan, R.; Pan, J.; Liu, Q.; Tan, X.; Chen, Q.; Xue, W.; and Guo, Y. 2024. VidMuse: A simple video-to-music generation framework with long-short-term modeling. *arXiv preprint arXiv:2406.04321*.

Tong, X.; Chen, S.; Yu, P.; Liu, N.; Qv, H.; Ma, T.; Zheng, B.; Yu, F.; and Zhu, S.-C. 2024. Video Echoed in Harmony: Learning and Sampling Video-Integrated Chord Progression Sequences for Controllable Video Background Music Generation. *IEEE Transactions on Computational Social Systems*.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6450–6459.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, B.; Zhuo, L.; Wang, Z.; Bao, C.; Chengjing, W.; Nie, X.; Dai, J.; Han, J.; Liao, Y.; and Liu, S. 2024a. Multimodal Music Generation with Explicit Bridges and Retrieval Augmentation. *arXiv preprint arXiv:2412.09428*.

Wang, Y.; Guo, W.; Huang, R.; Huang, J.; Wang, Z.; You, F.; Li, R.; and Zhao, Z. 2024b. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in Neural Information Processing Systems*, 37: 128118–128138.

Wu, S.-L.; Donahue, C.; Watanabe, S.; and Bryan, N. J. 2024. Music controlnet: Multiple time-varying controls for music generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2692–2703.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. *arXiv preprint arXiv:2211.06687*.

Xie, Z.; He, Q.; Zhu, Y.; He, Q.; and Li, M. 2025. FilmComposer: LLM-Driven Music Production for Silent Film Clips. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13519–13528.

Xing, Y.; He, Y.; Tian, Z.; Wang, X.; and Chen, Q. 2024. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7151–7161.

Xu, J.; Sun, X.; Zhang, Z.; Zhao, G.; and Lin, J. 2019. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32.

Xu, T.; Li, J.; Chen, X.; Yao, X.; and Liu, S. 2024. Mozart's Touch: A Lightweight Multi-modal Music Generation Framework Based on Pre-Trained Large Models. *arXiv preprint arXiv:2405.02801*.

Yang, X.; Yu, Y.; and Wu, X. 2022. Double Linear Transformer for Background Music Generation from Videos. *Applied Sciences*, 12(10).

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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

Zhang, Y.; Ikemiya, Y.; Xia, G.; Murata, N.; Martínez-Ramírez, M. A.; Liao, W.-H.; Mitsufuji, Y.; and Dixon, S. 2024. MusicMagus: zero-shot text-to-music editing via diffusion models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*.

Zhou, Z.; Mei, K.; Lu, Y.; Wang, T.; and Rao, F. 2025. Harmonyset: A comprehensive dataset for understanding video-music semantic alignment and temporal synchronization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3152–3162.

Zhu, B.; Lin, B.; Ning, M.; Yan, Y.; Cui, J.; HongFa, W.; Pang, Y.; Jiang, W.; Zhang, J.; Li, Z.; Zhang, C. W.; Li, Z.; Liu, W.; and Yuan, L. 2024. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. In *the Twelfth International Conference on Learning Representations*.

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 Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 182–199. Springer Nature Switzerland.

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*.