This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

VidMuse: A Simple Video-to-Music Generation Framework with Long-Short-Term Modeling

Zeyue Tian[*1], Zhaoyang Liu[*1], Ruibin Yuan[1], Jiahao Pan[1],
Qifeng Liu[1], Xu Tan[2], Qifeng Chen[†1], Wei Xue[†1], Yike Guo[1]

[1]Hong Kong University of Science and Technology
[2]Microsoft Research Asia

Abstract

In this work, we systematically study music generation conditioned solely on the video. First, we present a large-scale dataset by collecting 360K video-music pairs, including various genres such as movie trailers, advertisements, and documentaries. Furthermore, we propose VidMuse, a simple framework for generating music aligned with video inputs. VidMuse stands out by producing high-fidelity music that is both acoustically and semantically aligned with the video. By incorporating local and global visual cues, VidMuse enables the creation of coherent music tracks that consistently match the video content through Long-Short-Term modeling. Through extensive experiments, VidMuse outperforms existing models in terms of audio quality, diversity, and audio-visual alignment. The code and datasets are available at https://vidmuse.github.io/

1. Introduction

Music, as an essential element of video production, can enhance humans' feelings and convey the theme of the video content. Along with the development of social media platforms *i.e*., YouTube and TikTok, some studies [11, 46, 52] have shown that a piece of melodious music can vastly attract the audience's attention and interest in watching the video. It thus leads to a great demand for studying video-to-music generation [13, 20, 26, 29, 33, 65].

Nevertheless, music creation for a video is a challenging task, which needs to understand both music theory and video semantics. It would be very time-consuming to produce a piece of suitable music for video in a hand-crafted manner. Therefore, it is desirable when we can automatically generate high-quality music for different genres of videos. Currently, most of works [10, 19, 30, 31, 61, 79] have made significant achievements, especially in text-to-music generation, but the video-to-music generation still remains to be further studied. Specifically, existing works on video-conditioned music generation mainly focus on specific scenarios, such as dance videos [37, 87], or on the symbolic music, *i.e*., MIDI [13, 33, 73, 89]. However, these works are unable to generate more diverse musical styles and are also difficult to generalize to various video genres. Moreover, Hong *et al*. [29] build a music–video retrieval dataset from YouTube-8M [1], albeit with limited video genres. Despite that there are also some prominent works [32, 65] employing multi-modal inputs to generate music for the video, it is still worth studying that **conditioned solely on the visual input, whether it is possible to generate diverse and harmonious music for various genres of videos.**

Motivated by this, we first construct a large-scale dataset termed *V2M*, equipped with a comprehensive benchmark to evaluate the state-of-the-art works thoroughly. The video-music pairs are collected from YouTube with various genres, *e.g*., movie trailers, advertisements, documentaries, vlogs, *etc*. In order to ensure the quality of our dataset, we establish a multi-step pipeline illustrated in Fig. 1 to systematically clean and pre-process data. The videos with low quality or composed of static images are filtered out. The proposed dataset contains three subsets: *V2M-360K* for pre-training, *V2M-20K* for finetuning, and *V2M-bench* for evaluation. We believe that *V2M* is able to facilitate the development of video-to-music generation.

Furthermore, on top of *V2M*, we propose a simple yet effective method, termed as *VidMuse*, to generate music only conditioned on the visual input. Instead of predicting the intermediate musical symbols such as MIDI or retrieving the music from the database, the proposed VidMuse, integrates both local and global visual cues to generate background music consistent with the video in an end-to-end manner. The core techniques in our method are a *Long-Short-Term*

[*]Equal Contribution
[†]Corresponding Authors

Visual Module (LSTV-Module) and a *Music Token Decoder*. Specifically, the LSTV-Module aims to learn the spatial-temporal representation of videos, which is the key to generating music aligned with the video. On the one hand, the long-term module models the entire video, capturing global context to understand the whole video. It contributes to the coherence of generated music at the video level. On the other hand, the short-term module focuses on learning the fine-grained cues at the clip level, which plays a vital role in generating diverse music. The integration of two modules can improve the quality and visual consistency of generated music. In addition, the Transformer-based music token decoder is an autoregressive model, converting video embeddings obtained by LSTV-Module into music tokens. We formulate music generation as a task of next token prediction, which has been widely validated by the NLP community. The predicted music tokens are further decoded into the music signals by a high-fidelity neural audio compression model.

The main contributions of this work are as follows:

- We construct a large-scale video-to-music dataset, *i.e.*, *V2M*, which contains about 360k video-music pairs with high quality, covering various genres and themes. To the best of our knowledge, this is the largest and most diverse dataset for this task, which can facilitate future research.
- We propose a simple yet effective method, VidMuse, for video-to-music generation. The proposed method integrates both local and global cues in the video, enabling the generation of high-fidelity music tracks that are not only musically coherent but also semantically aligned with the video content.
- We benchmark several state-of-the-art works against our method on the V2M-bench via a series of subjective and objective metrics for a thorough evaluation. As demonstrated in experiments, VidMuse achieves state-of-the-art performance on *V2M-bench*, outperforming existing models in terms of audio quality, diversity, and audio-visual consistency.

2. Related Work

We review the existing works related to video-to-music generation, which mainly fall into four categories:

Video Representation. Various methods have been proposed to learn the spatio-temporal representation [4, 18, 41, 42, 48, 67, 69, 84] for videos. They aim to capture the contextual features of video frames, which is beneficial for video understanding. Recent advances primarily concentrate on developing video transformers [4, 43, 47, 67]. These transformer-based methods achieve superior generalized performance on various video understanding tasks, such as video classification and temporal action localization. Among them, Tong *et al*. [67] extend masked autoencoders [25] from the image to the video, exhibiting the strong generalized performance in downstream tasks. Benefiting from the advance in multi-modal large language models, lots of works [39, 54, 84] of interactive video understanding have been proposed, which built upon the large language models (LLMs) [68, 85] and showcase the visual reasoning capabilities for video understanding.

Audio-Visual Alignment. Audio-visual alignment [3, 8, 23, 60, 63, 74, 76] aims to align the feature between audio, vision in the semantics level, which plays a vital role in tasks of audio-visual understanding and generation. For example, CAV-MAE [23] is an audio-visual MAE that integrates the contrastive learning and masked modeling method. Currently, many works go beyond exploring audio-visual alignment. ImageBind [22] extends CLIP [56] to more modalities, including audio, depth, thermal, and IMU data, which paves the way for cross-modal retrieval and generation. In addition, Wu *et al*. [74] employ LLMs with multi-modal adaptors to support any modal data as input and output, showing strong capabilities in universal multi-modal understanding. These methods transcend audio-visual alignment and dramatically advance the development of multi-modal representation learning.

Conditional Music Generation. Despite that there are lots of methods [15, 24, 30, 40, 45, 50, 53] studying unconditional music generation, in this paper, we mainly focus on reviewing the methods of conditional music generation, which are more related to our work. Many researchers [2, 10, 12, 19, 31, 61, 79, 81] make their endeavours on text-to-music generation. Similar to Stable Diffusion [57], these works [19, 31, 61, 79] try to adapt diffusion models for music generation. M^2UGen [32] is a multi-modal music understanding and generation system that leverages large language models to process video, audio, and text. Video2Music[33] can generate music that matches the content and emotion of a given video. Moreover, the proposed V2Meow [65] and MeLFusion [9] conditioned on video and image, respectively, can generate music that supports style control via text prompts. In contrast to previous video-to-music works [20, 33, 65, 89], our VidMuse utilizes a short-term module and a long-term module to model local and global visual cues in videos. As a result, it can generate high-fidelity music aligned with the video.

Video-to-music Datasets. Many multimodal datasets [1, 6, 7, 21, 27, 36, 49, 51, 62, 64, 66, 77, 86] have been released, but there is still a lack of datasets for video-to-music generation. Hong *et al*. [29] construct the HIMV-200K with video-music pairs and aim to retrieve music for the video from the database. However, this dataset exhibits limited video genres and also suffers from the issue of data quality, as stated in [89]. We observe that several works [13, 38, 73, 89] aim to facilitate MIDI music generation. However, this musical form imposes restrictions on diversity for the music generation. Other

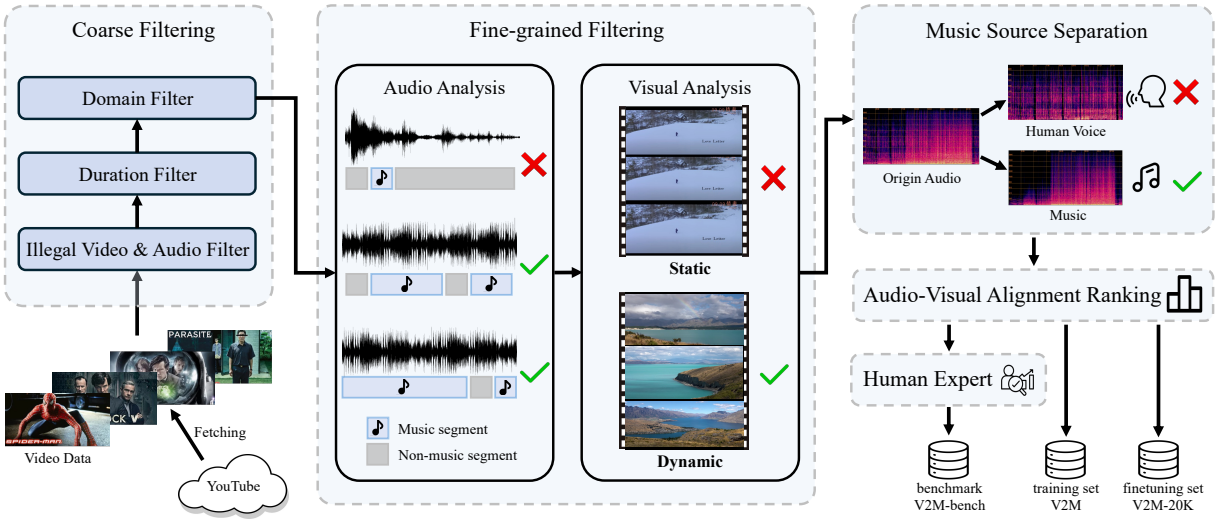
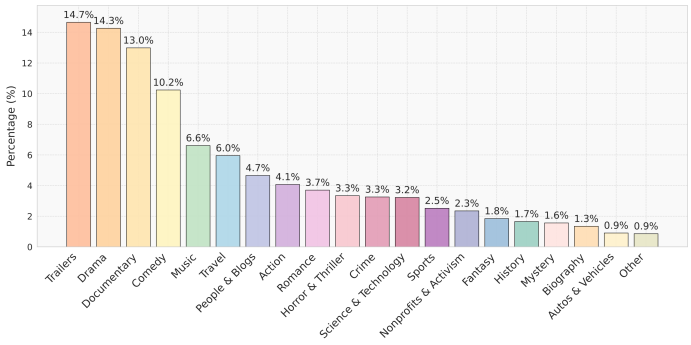
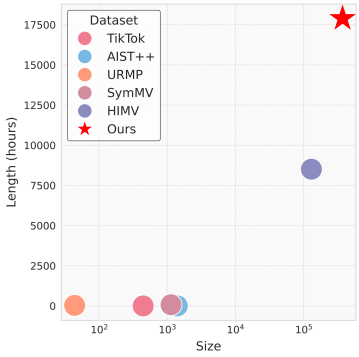


Figure 1. **Dataset Construction.** To ensure data quality, *V2M* goes through rule-based coarse filtering and content-based fine-grained filtering. Music source separation is applied to remove speech and singing signals in the audio. After processing, human experts curate the benchmark subset, while the remaining data is used as the pretraining dataset. The pretrain data is then refined using Audio-Visual Alignment Ranking to select the finetuning dataset.



(a) Video genre distribution.



(b) Dataset Comparisons.

Figure 2. **Statistics of our dataset.** (a) The distribution of video genres in our dataset, (b) Comparisons with other related datasets in terms of scale of datasets. Please zoom in for details.

datasets [37, 87] focus on generating music for dance videos only and have limited data size, which limits their applicability for general video-to-music models. As a result, this work constructs a large-scale video-to-music dataset where the music directly in wav format is diverse. We establish a rigorous pipeline for data collection and cleaning, which ensures the quality and diversity of our dataset. We expect the model can learn the music with more diverse forms from this dataset.

3. Dataset

In this section, we build a multi-step pipeline to clean and process source videos from YouTube to ensure data quality. After that, we construct a large-scale video-to-music generation dataset, *i.e.*, *V2M*, with a benchmark. The constructed dataset stands out for its significant size, high quality, and rich diversity, including a wide range of genres such as movie trailers, advertisements, documentaries, vlogs, *etc*. This comprehensive and diverse dataset aims to facilitate the video-to-music generation.

3.1. Dataset Collection

To quickly collect a large scale of video-music pairs, we curate a series of query sets to retrieve corresponding videos from YouTube. In addition, we find that the music in the movie trailer usually showcases rich diversity and high quality. Therefore, we also aggregate a vast array of video information from the IMDb Non-Commercial Datasets, including video types, names, release dates, etc. Queries are

formulated based on the titles of these selections and retain the videos released after 2000, as videos from earlier periods are less likely to be of good quality. In the process of video crawling, we only keep the top 2 search results, resulting in a collection of around 400K videos, ranging from movies to documentaries. Besides, several existing datasets already contain video-music pairs, such as HIMV-200K [29], subsets of YouTube-8M [1] labeled with "Music" and "Trailer" tags. We incorporate these datasets into our collection to further expand its scope. After merging all sources, our final dataset comprises about 600K videos, spanning diverse genres and categories.

3.2. Dataset Construction

The raw videos may include many low-quality samples. To address this, we develop a series of rigorous steps to filter out undesirable data and obtain a clean set. The overall pipeline of data processing is depicted in Fig. 1. The following steps outline our approach: (1) The process begins with coarse filtering, where we remove videos lacking audio or video tracks, videos that are too short or too long, those containing inappropriate content such as violence or explicit material, and those from categories like *Interview* and *News*, which generally have background music not aligned with the visual content. (2) Following that, we perform fine-grained filtering to retain videos with substantial music content and dynamic visual elements. We use an audio analysis model [34] to identify music segments, ensuring a sufficient portion of the audio is classified as music. In parallel, we analyze the visuals [72] to exclude videos consisting mainly of static images. (3) To further refine the dataset, we apply music source separation [59] to isolate the music component by removing vocal tracks, enhancing the overall audio quality. (4) Finally, we rank the videos based on their audio-visual alignment scores [22] to ensure a high level of semantic correlation between the audio and visual modalities. The resulting videos are then split into training (*V2M*), fine-tuning (*V2M-20K*), and benchmark (*V2M-bench*) subsets. For details on dataset construction, please refer to the Appendix 9).

3.3. Data Analysis

The above data pipeline yields three data splits. Specifically, the training set comprises ∼360K video-music pairs, around 1.8×10^4 hours. The finetuning dataset consists of ∼20K pairs, about 6×10^2 hours. The benchmark dataset contains 300 pairs, with a cumulative duration of 9 hours. Fig. 2a showcases the genre distribution of our training data, highlighting its comprehensive diversity. This diversity ensures a rich and varied dataset for the model training. As shown in Fig. 2b, we compare with other related datasets, demonstrating its advantage in data scale.

Dataset Necessity. Some existing video-music pair datasets have been released [13, 29, 37, 73, 87, 89], but some of them [13, 73, 89] aim to facilitate MIDI music generation, which limits the form of music. Datasets like [37, 87] focus on generating music for dance videos only and have limited data size. The dataset constructed by [29] includes video-music pairs but exhibits limited video genres and suffers from data quality issues. In addition, the evaluation metrics used in these video-to-music benchmarks [13, 29, 37, 38, 73, 80, 87, 89] are divergent, making it difficult to fairly and thoroughly assess performance of methods on different benchmarks. Motivated by this, we develop the multi-step pipeline and curate a large-scale dataset *V2M* as well as a benchmark for the video-to-music generation.

4. Method

4.1. Architecture of VidMuse

In this section, we elaborate on the framework of VidMuse, which leverages **LSTV-Module** to generate music aligned with video content. The proposed pipeline is shown in Fig. 3, including (1) Visual Encoder, (2) LSTV-Module, (3) Music Token Decoder, and (4) Audio Codec.

Visual Encoder. To generate music conditioned on the video, we first need to extract the high-level features from a stack of frames. Given an input video, the visual encoder extracts feature representations $\mathbf{X} \in \mathbb{R}^{N \times P \times D}$. Here, N is the number of input frames, P refers to the sequence length with the class token, and D denotes the size of the feature vectors. Currently, there are lots of visual encoders available, including 2D [16], 3D [4, 67] and multi-modal [56] models, which will be validated in the Sec. 5.5.

LSTV-Module. Generating music for videos with variable length still presents significant challenges, especially for a long video, because sometimes it is difficult to directly model on whole video due to hardware limitations. Prior studies opt to generate music in segment level [13, 32, 33]. However, such a manner often lacks sufficient context information. Music should vary in expression depending on the context. Even the same video segment may lead to distinct musical interpretations when being in different contexts. By incorporating global guidance, it can enhance the alignment of the generated music with the overall video content. To capture both local and global visual cues, the visual features extracted from the visual encoder are thus fed into the LSTV-Module. Specifically, the short-term module takes segment-level embeddings as input, aiming to capture local dependencies $\mathbf{X}_s \in \mathbb{R}^{N_s \times P \times D}$ to ensure that the generated music aligns with short-term variations in the video, while long-term module models on video-level embeddings, providing context $\mathbf{X}_l \in \mathbb{R}^{N_l \times P \times D}$ to guide the short-term module in generating more visually coherent music. N_s and N_l is the number of frames sampled from the video.

To capture both global and local visual cues, we leverage

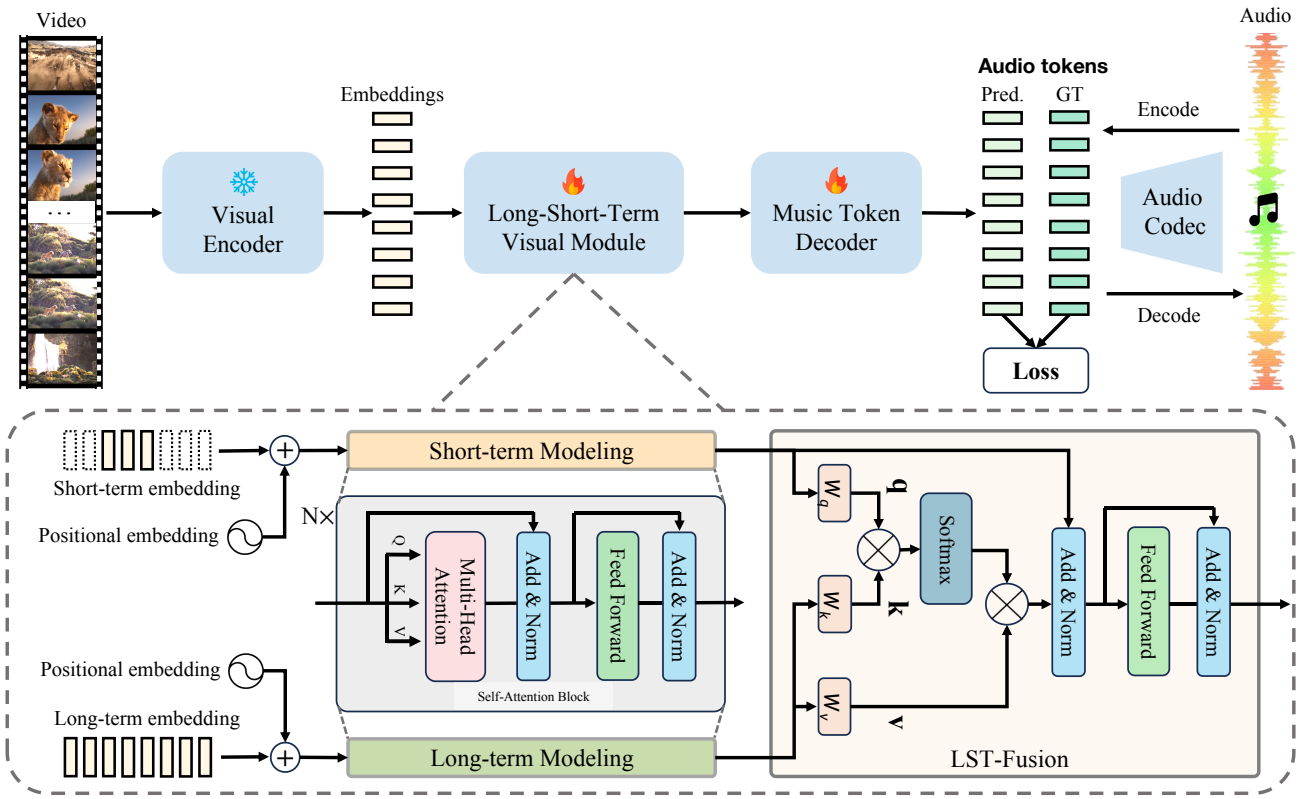


Figure 3. **Overview of the VidMuse Framework.** This pipeline outlines the key components for generating music aligned with video content: (1) Visual Encoder for extracting visual features, (2) Long-Short-Term Visual Module for integrating local and global cues, (3) Music Token Decoder for generating music tokens, and (4) Audio Codec for the conversion between audio and audio tokens.

the cross-attention [70] in both Long-term and Short-term Modeling. Long-term modeling extracts long-range dependencies, while short-term modeling focuses on local details. This results in refined long-term features $\mathbf{F}_l \in \mathbb{R}^{N_l \times P \times D}$ and short-term features $\mathbf{F}_s \in \mathbb{R}^{N_s \times P \times D}$.

To incorporate global guidance for generating segment-based music, we design LST-Fusion. It integrates long-term and short-term features by utilizing the cross-attention mechanism $\mathrm{CA}(\cdot)$ with Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$), which can be mathematically formulated as:

$$\begin{aligned} \mathbf{Z}' &= \mathrm{CA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}), \\ \text{where } \mathbf{Q} &= \mathbf{F}_s, \mathbf{K} = \mathbf{F}_l, \mathbf{V} = \mathbf{F}_l, \end{aligned} \tag{1}$$

This mechanism allows the model to query global information rather than generating music based solely on local visual features. It guarantees that the generated music is more consistent with the video content. After the cross-attention, a linear layer projects $\mathbf{Z}'$ to $\mathbf{Z} \in \mathbb{R}^{N_s \times P \times M}$, where M represents the input vector dimension of the music token decoder in the next step. In addition, we also explore different implementations discussed in experiments.

Music Token Decoder. We adopt an autoregressive approach to predict the music tokens $\bar{\mathbf{Y}}$ conditioned on the video segment. Music token decoder is implemented by a transformer decoder with a linear classifier. We set the latent vector size of the transformer decoder to M, allowing it to scale up or down the model's size. The decoder incorporates a cross-attention mechanism that receives the visual signal $\mathbf{Z} \in \mathbb{R}^{N_s \times P \times M}$, where N_s is the number of frames sampled in the video segment. At each time step t (where $t = 1, \ldots, T$), the decoder predicts the logits of current token $\bar{\mathbf{Y}}_t \in \mathbb{R}^{K \times C}$ based on previous tokens and visual context $\mathbf{Z}$. Here, K denotes the number of codebooks, and C represents the vocabulary size.

Audio Codec. It can convert an audio segment into discretized codebooks and, conversely, decode codebooks back into audio. The size of codebooks is $K \times T$, where T denotes the length of the video. Given the Audio codec $\mathcal{C}$, we denote the Encoder as $\mathcal{C}_{\text{encode}}(\cdot)$ and the Decoder as $\mathcal{C}_{\text{decode}}(\cdot)$. In training, we need to encode the ground truth audio $\mathbf{A}$ into discretized tokens serving as supervise signals for the next token prediction. In the inference phase, the predicted tokens will be then decoded into music signals.

Table 1. **Comparison with naive baselines and state-of-the-art methods.**

Methods	Metrics					
	KL ↓	FD ↓	FAD ↓	density ↑	coverage ↑	Imagebind ↑
GT	0.000	0.000	0.000	1.167	1.000	0.241
Caption2Music	1.081	40.199	2.485	0.378	0.486	0.191
Video2Music [33]	1.782	144.881	18.722	0.103	0.023	0.136
CMT [13]	1.220	85.704	8.637	0.080	0.070	0.124
M^{2}UGen [32]	0.997	52.246	5.104	0.608	0.433	0.181
M^{2}UGen* [32]	0.965	52.041	5.003	0.633	0.430	0.180
VM-NET [29]	0.899	67.480	6.252	0.986	0.383	0.147
VidMuse	**0.734**	**29.946**	**2.459**	**1.250**	**0.730**	**0.202**

4.2. Training

Given a video segment with corresponding ground-truth audio $\mathbf{A}$, we train our model using a next-token prediction approach. The video segments are processed through the Visual Encoder and the LSTV-Module to generate visual features, which are then fed into the Music Token Decoder to produce the predicted logits $\bar{\mathbf{Y}} \in \mathbb{R}^{K\times T\times C}$. Here, K denotes the number of codebooks, T is the sequence length (number of timesteps), and C is the vocabulary size for codebooks. Next, the ground-truth audio $\mathbf{A}$ is encoded by the Audio Codec to obtain the target one-hot vector: $\mathbf{Y} = \mathcal{C}_{\text{encode}}(\mathbf{A})$, where $\mathbf{Y} \in \mathbb{R}^{K\times T\times C}$. The value of $\mathbf{Y}_{k,t,c}$ is 1 when c equals the ground-truth token index in codebook k at timestep t, and 0 otherwise.

Our objective is to minimize the cross-entropy loss between the predicted probabilities $\bar{\mathbf{Y}}$ and the ground-truth tokens $\mathbf{Y}$. The cross-entropy loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = -\frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{c=1}^{C}\mathbf{Y}_{k,t,c}\log\bar{\mathbf{Y}}_{k,t,c}, \tag{2}$$

where $\bar{\mathbf{Y}}_{k,t,c}$ is the predicted probability of class c at codebook k and timestep t. Conditioned on the video, we train the model by predicting the next token with this loss.

5. Experiments

In this section, we elaborate on the implementation details of our experiments, and conduct massive experiments to thoroughly evaluate the efficacy of our proposed method from both subjective and objective perspectives. This is expected to provide insights for video-to-music generation.

5.1. Implementation details

Since this work does not focus on audio encoding and decoding, we use Encodec [17] for 32 kHz monophonic audio as our default compression model and use the pretrained transformer model proposed in MusicGen [10]. The training stage utilizes the AdamW optimizer [44] with a batch size of 5 samples per GPU. We sample frames from a continuous 30s video segment at 2 fps for short-term modeling and uniformly sample 32 frames from the entire video for long-term modeling. The hyperparameters are set to $\beta_1 = 0.9$, $\beta_2 = 0.95$, with a weight decay of 0.1 and gradient clipping at 1.0. A cosine learning rate schedule is employed, incorporating a warm-up phase of 4,000 steps and an exponential moving average decay of 0.99. We use 64 H800 GPUs in pretraining and train models with $56K$ steps, which takes about 50 hours. For the finetuning stage, we utilize 32 H800 GPUs and train models with $8K$ steps, which takes about 8 hours. A top-k strategy is applied for sampling, retaining the top 250 tokens with a temperature setting of 1.0. In the inference stage, we set the sliding window size as 30s, and the window's overlap as 0.5s.

5.2. Evaluation Metrics

To quantitatively evaluate the effectiveness of our model, we employ a series of metrics to assess different models in terms of quality, fidelity, and diversity of the generated music. These metrics include the Frechet Audio Distance (FAD), Frechet Distance (FD), Kullback-Leibler Divergence (KL), as well as Density and Coverage [55]. Additionally, we utilize the ImageBind Score [22] to examine the alignment between the video and the generated music. We acknowledge that ImageBind has limitations as it is not specifically trained on music data, but it currently seems to be a possible option for evaluating the semantic alignment between video and generated music. To fairly compare with non-public baseline models [65, 80, 87], we use evaluation metrics from their papers: beats coverage score (BCS), beats hit score (BHS), standard deviations of BCS (CSD) and BHS (HSD), and F1 scores of BCS and BHS.

5.3. Main Results

We benchmark several state-of-the-art methods, serving as baselines to compare with our method: 1) **Caption2Music**, a naive baseline that employs the SpaceTimeGPT to ex-

Table 2. **Results across different benchmarks [37, 80, 87].**

Methods	Metrics				
	BCS ↑	CSD ↓	BHS ↑	HSD ↓	F1 ↑
AIST++ [37]					
D2M-GAN [87]	92.3	–	91.7	–	–
CDCD [88]	93.9	1.2	90.7	**1.5**	–
V2Meow [65]	**100.0**	**0.0**	84.4	25.1	91.5
VidMuse	99.97	0.3	**96.5**	9.4	**98.2**
LORIS [80]					
D2M-GAN [87]	95.6	9.4	88.7	19.0	93.1
CDCD [88]	96.5	9.1	89.3	18.1	92.7
LORIS[80]	**98.6**	**6.1**	90.8	13.9	94.5
VidMuse	96.3	10.6	**95.6**	**8.9**	**95.9**
TikTok [87]					
D2M-GAN [87]	87.1	–	83.9	–	–
CDCD [88]	**91.8**	–	86.3	–	–
VidMuse	79.8	20.0	**97.3**	7.6	87.7

Table 3. **Ablation studies on design choices.**

Methods	KL ↓	FD ↓	FAD ↓	density ↑
VidMuse-STM	0.898	45.752	4.915	1.124
VidMuse-LTM	0.858	53.907	16.074	**1.439**
VidMuse-CAQ_SL	0.843	48.940	3.733	0.947
VidMuse-CAQ_LS	0.919	45.335	2.917	0.562
VidMuse-Slowfast	1.511	84.683	10.029	0.266
VidMuse	**0.738**	**36.171**	**2.369**	1.175

Table 4. **Ablation studies on visual encoders.**

Encoders	KL ↓	density ↑	GFLOPs ↓	Latency ↓	Throughput ↑
ViViT	0.822	**1.433**	451.83	1650 ms	9.12
VideoMAE	0.778	1.074	360.99	452 ms	17.44
CLIP	**0.753**	1.122	**141.24**	**341** ms	**24.16**
ViT	0.876	1.081	562.64	405 ms	23.84

tract the video captions and outputs the music by feeding captions into MusicGen [10]. 2) **Video2Music** [33] and 3) **CMT** [13] which both predict MIDI notes [58] from videos while our method directly generates music signals. 4) **M^{2}UGen** [32], a strong baseline, which leverages a language model to connect vision and language, then use MusicGen [10] to generate music from language. 5) **M^{2}UGen***, a re-trained version of M^{2}UGen using our dataset. 6) **VM-NET** [29], different from above methods, retrieves a piece of music from the database for a given video, while other methods predict music by training on video-music pairs.

In Table 1, VidMuse, with both global and local visual modeling, exhibits impressive performance on all metrics. Specifically, compared with Video2Music [33] and CMT [13], VidMuse shows the superiority in the diversity of generated music based on the density or coverage. It justifies the advantage of directly predicting music signals compared with MIDI notes. Our method even outperforms the strong competitors, *i.e.*, M^{2}UGen. It proves that our method of directly predicting music based on video input can also achieve better performance. Furthermore, compared with a retrieval-based method, *i.e.*, VM-NET [29], VidMuse achieves a higher Imagebind score, indicating that the music generated by the learning-based strategy is more consistent with the video semantics.

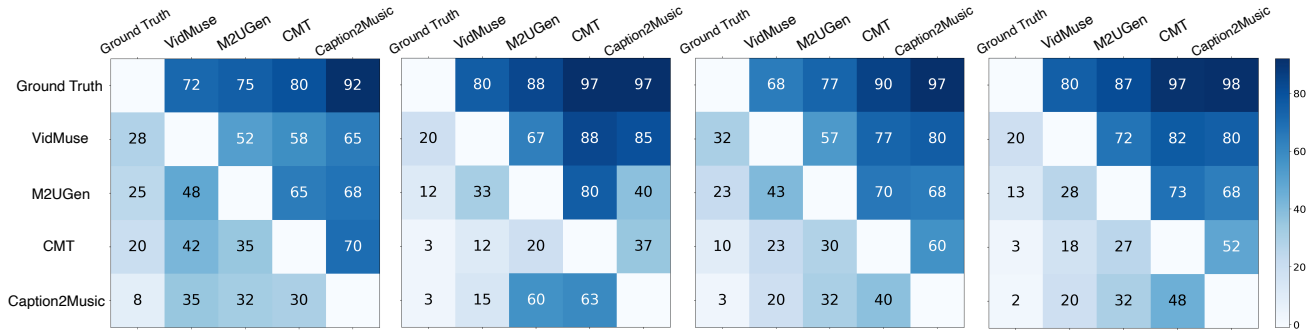
Furthermore, we validate the generalization abilities of our proposed VidMuse on several different types of video-to-music benchmarks. To make fair comparisons and avoid data leakage, we check our dataset and remove the potential repeated data sample from our training set. As shown in Table 2, our VidMuse achieves comparable performance against other methods. This demonstrates that VidMuse does not overfit on our own benchmarks and exhibits strong generalization capabilities.

5.4. Subjective Comparisons with User Study

In the user study, we randomly sample 600 video-music pairs from the benchmark to conduct an A/B test, which is a widely used subjective evaluation method in the music field [14, 81]. This test includes CMT, M^{2}UGen, Caption2Music, Ground Truth (GT), and VidMuse. The test was distributed among 40 participants, ensuring each method was compared against another 60 times. The evaluation criteria are four-fold: 1) **Audio quality**: Refers to the sound quality of the audio track; 2) **Video-music alignment**: Assesses how well the music matches the visual content, e.g., a scene showing a woman crying should ideally be paired with music that sounds sad; 3) **Musicality**: Evaluates the aesthetic quality of the music, distinct from audio quality. For example, a piece of music may have good audio quality, but if it is out of tune, it would be considered to have poor musicality; 4) **Overall assessment**: Comprehensively evaluates the performance for models. Participants are asked to choose the better sample for each criterion. The user study is shown in Fig. 4, where the value at matrix$[i][j]$ ranges from 0 to 100, indicating the % of times listeners preferred the method in i-row compared to the method in j-column. For example, in Fig. 4 (c), the value of matrix[2][4] represents that VidMuse outperforms CMT in 77% of cases in terms of Musicality. Across all criteria, our method surpasses others in more than half of the comparisons, except when compared to the ground truth. Overall, these results thoroughly validate VidMuse's effectiveness through subjective evaluation.

5.5. Ablation Studies

In this section, we conduct ablation studies, aiming to find the optimal design choices.



(a) Audio quality. (b) Video-music alignment. (c) Musicality. (d) Overall assessment.

Figure 4. **A/B test results of the user study.** We design four criteria in Sec. 5.4 to assess the subjective performance.

Justification of Design Choices. To validate the impact of different model design choices on our generation results and verify the effectiveness of our method, we first design two modules: a short-term modeling module (STM) and a long-term modeling module (LTM). VidMuse-STM aims to ablate the contribution of STM by removing LTM, while VidMuse-LTM utilizes only LTM. Based on the results in Table 3, we gain the insight that local information plays a more important role in the generation. By integrating global guidance with local information, we improve the alignment of the generated music with the video content. Second, we implement two variants with **C**ross-**A**ttention with learnable **Q**ueries (**CAQ**) in our framework. Specifically, CAQ_SL first uses a **CAQ** where K and V are short-term features and then uses a **CAQ** where K and V are long-term features. CAQ_LS does it in the opposite order. As shown in Table 3, VidMuse outperforms two variants, demonstrating the efficacy of our manner. Furthermore, we evaluate a baseline that replaces the LSTV-Module with a SlowFast-like mechanism [18], where the slow path models appearance and the fast path captures temporal dynamics. As shown in Table 3, this modification degrades performance, likely because the fast path operates at a higher frame rate while sharing the same temporal receptive field as the slow path, resulting in insufficient global guidance.

Visual Encoder. We here study the impact of various visual encoders. As shown in Table 4, we experiment with different visual encoders, including ViT [16], CLIP [56], VideoMAE [67], and ViViT [4]. For fair comparisons, these encoders all use ViT-B as the backbone. The Latency and throughput are assessed with 30-second videos on the NVIDIA H800 GPU. Latency is measured with a batch size of 1, while throughput is measured with a batch size of 16. Our results show that VidMuse remains robust in processing visual information for music generation across all encoder choices. To balance computational efficiency and generation quality, we select CLIP [56] as the default visual encoder if not stated.

More ablations are provided in the appendix, including effects of the finetuning set, different input settings, *etc*.

6. Conclusion

In this work, we build a rigorous pipeline to collect high-quality and diverse video-music pairs, curating a comprehensive dataset **V2M**. Then, we propose VidMuse, a simple yet effective method for video-to-music generation. Our approach utilizes a Long-Short-Term approach to capture both local and global visual cues in the video, allowing for the generation of contextually rich and musically diverse outputs. To validate our method, we benchmark a series of state-of-the-art methods as baselines to compare with VidMuse. Through comprehensive quantitative studies and qualitative analyses, our method has demonstrated its superiority over the existing methods.

Limitations. Our work achieves a significant advancement in video-to-music generation, but it still has some limitations. First, the current implementation relies on the EnCodec model [17], which sometimes exhibits a noticeable reconstruction loss for different genres of audio, potentially lowering the quality of the generated music. Second, our work solely focuses on video-to-music generation without exploring other input conditions. As such, our future work aims to overcome these limitations by integrating advanced codec technologies to enhance audio reconstruction fidelity, and exploring diverse controls during music generation.

7. Acknowledgment

The research was supported by NSFC (No. 62206234), the Early Career Scheme (ECS-Hong Kong University of Science and Technology 22201322), the Theme-based Research Scheme (T45-205/21-N), and the General Research Fund (No. 16212623) from the Hong Kong RGC, as well as the Generative AI Research and Development Center from InnoHK. We also thank Fortuna for discussions and Kyle for assisting in setting up the user study UI.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1, 2, 4

[2] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 2

[3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 2

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 2, 4, 8

[5] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 4

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 2

[7] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 2

[8] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 431–448. Springer, 2022. 2

[9] Sanjoy Chowdhury, Sayan Nag, KJ Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26826–26835, 2024. 2

[10] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 6, 7

[11] Johanna N Dasovich-Wilson, Marc Thompson, and Suvi Saarikallio. Exploring music video experiences and their influence on music perception. *Music & Science*, 5: 20592043221117651, 2022. 1

[12] Qixin Deng, Qikai Yang, Ruibin Yuan, Yipeng Huang, Yi Wang, Xubo Liu, Zeyue Tian, Jiahao Pan, Ge Zhang, Hanfeng Lin, et al. Composerx: Multi-agent symbolic music composition with llms. *arXiv preprint arXiv:2404.18081*, 2024. 2

[13] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2037–2045, 2021. 1, 2, 4, 6, 7

[14] Chris Donahue, Antoine Caillon, Adam Roberts, Ethan Manilow, Philippe Esling, Andrea Agostinelli, Mauro Verzetti, Ian Simon, Olivier Pietquin, Neil Zeghidour, et al. Singsong: Generating musical accompaniments from singing. *arXiv preprint arXiv:2301.12662*, 2023. 7

[15] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 8

[17] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022. 6, 8

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2, 8

[19] Seth Forsgren and Hayk Martiros. Riffusion-stable diffusion for real-time music generation, 2022. 6. 1, 2

[20] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 758–775. Springer, 2020. 1, 2

[21] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 2, 4

[22] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2, 4, 6

[23] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022. 2

[24] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich

Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the maestro dataset. *arXiv preprint arXiv:1810.12247*, 2018. 2

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[26] Yingqing He, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhai Wang, Jifeng Dai, Yong Zhang, Wei Xue, Qifeng Liu, Yike Guo, and Qifeng Chen. Llms meet multimodal generation and editing: A survey. *arXiv preprint arXiv:2405.19334*, 2024. 1

[27] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 2, 3

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[29] Sungeun Hong, Woobin Im, and Hyun S Yang. Content-based video-music retrieval using soft intra-modal structure constraint. *arXiv preprint arXiv:1704.06761*, 2017. 1, 2, 4, 6, 7

[30] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018. 1, 2

[31] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023. 1, 2

[32] Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. M^2UGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models. *arXiv preprint arXiv:2311.11255*, 2023. 1, 2, 4, 6, 7

[33] Jaeyong Kang, Soujanya Poria, and Dorien Herremans. Video2music: Suitable music generation from videos using an affective multimodal transformer model. *arXiv preprint arXiv:2311.00968*, 2023. 1, 2, 4, 6, 7

[34] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 4, 1, 3

[35] Gael Le Lan, Varun Nagaraja, Ernie Chang, David Kant, Zhaoheng Ni, Yangyang Shi, Forrest Iandola, and Vikas Chandra. Stack-and-delay: a new codebook pattern for music generation. *arXiv preprint arXiv:2309.08804*, 2023. 1

[36] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021. 2

[37] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1, 3, 4, 7

[38] Sizhe Li, Yiming Qin, Minghang Zheng, Xin Jin, and Yang Liu. Diff-bgm: A diffusion model for video background music generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27348–27357, 2024. 2, 4

[39] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2

[40] Jen-Yu Liu, Yu-Hua Chen, Yin-Cheng Yeh, and Yi-Hsuan Yang. Unconditional audio generation with generative adversarial networks and cycle regularization. *arXiv preprint arXiv:2005.08526*, 2020. 2

[41] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11669–11676, 2020. 2

[42] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13708–13718, 2021. 2

[43] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 2

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[45] Ang Lv, Xu Tan, Peiling Lu, Wei Ye, Shikun Zhang, Jiang Bian, and Rui Yan. Getmusic: Generating any music tracks with a unified representation and diffusion framework. *arXiv preprint arXiv:2305.10841*, 2023. 2

[46] Lin Ma. Research on the effect of different types of short music videos on viewers' psychological emotions. *Frontiers in Public Health*, 10:992200, 2022. 1

[47] Yue Ma, Xiaodong Cun, Yingqing He, Chenyang Qi, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. *arXiv preprint arXiv:2312.03047*, 2023. 2

[48] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 2

[49] Yinghao Ma, Anders Øland, Anton Ragni, Bleiz MacSen Del Sette, Charalampos Saitis, Chris Donahue, Chenghua Lin, Christos Plachouras, Emmanouil Benetos, Elona Shatri, et al. Foundation models for music: A survey. *arXiv preprint arXiv:2408.14340*, 2024. 2

[50] Kinyugo Maina. Msanii: High fidelity music synthesis on a shoestring budget. *arXiv preprint arXiv:2301.06468*, 2023. 2

[51] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 2

[52] Barbara Millet, Juan Chattah, and Soyeon Ahn. Soundtrack design: The impact of music on visual attention and affective responses. *Applied ergonomics*, 93:103301, 2021. 1

[53] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*, 2021. 2

[54] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 2

[55] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 6, 4

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4, 8

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[58] Joseph Rothstein. *MIDI: A comprehensive introduction*. AR Editions, Inc., 1995. 7

[59] Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 4, 1

[60] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 2

[61] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023. 1, 2

[62] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[63] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 2

[64] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021. 2

[65] Kun Su, Judith Yue Li, Qingqing Huang, Dima Kuzmin, Joonseok Lee, Chris Donahue, Fei Sha, Aren Jansen, Yu Wang, Mauro Verzetti, et al. V2meow: Meowing to the visual beat via music generation. *arXiv preprint arXiv:2305.06594*, 2023. 1, 2, 6, 7

[66] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer, 2020. 2

[67] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2, 4, 8

[68] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[69] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[71] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023. 1

[72] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4, 1

[73] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Xianbin Gu, and Gus Xia. Pop909: A pop-song dataset for music arrangement generation. *arXiv preprint arXiv:2008.07142*, 2020. 1, 2, 4

[74] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 2

[75] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and

keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 4

[76] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. *arXiv preprint arXiv:2402.17723*, 2024. 2

[77] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[78] Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023. 1

[79] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. 1, 2

[80] Jiashuo Yu, Yaohui Wang, Xinyuan Chen, Xiao Sun, and Yu Qiao. Long-term rhythmic video soundtracker. In *International Conference on Machine Learning*, pages 40339–40353. PMLR, 2023. 4, 6, 7

[81] Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, et al. Chatmusician: Understanding and generating music intrinsically with llm. *arXiv preprint arXiv:2402.16153*, 2024. 2, 7

[82] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020. 4

[83] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. 1

[84] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2

[85] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[86] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[87] Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov. Quantized gan for complex music generation from dance videos. In *European Conference on Computer Vision*, pages 182–199. Springer, 2022. 1, 3, 4, 6, 7

[88] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal music and image generation. *arXiv preprint arXiv:2206.07771*, 2022. 7

[89] Le Zhuo, Zhaokai Wang, Baisen Wang, Yue Liao, Chenxi Bao, Stanley Peng, Songhao Han, Aixi Zhang, Fei Fang, and Si Liu. Video background music generation: Dataset, method and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15637–15647, 2023. 1, 2, 4