

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

Supplementary Material

8. Additional Experiments

Additional experiments focusing on model inputs, codebook patterns, and finetuning effects are provided in the appendix. These parts provide insight into the decision-making process for selecting the experimental configurations within the VidMuse framework.

Exploration on model inputs. To explore the impact of different video sampling rates and the duration of video segments in the Short-Term module on performance, we conducted ablation studies on input FPS and short-term segment duration, detailed in Table A1. To intuitively assess the effectiveness of different settings, we employ an **Average Rank (AR)** metric. The AR metric ranks the results for a metric across all methods within the same table. The ranking result is from 1 to N (equals to the number of methods within the table), where 1 is the best and N is the worst. We eventually obtain AR results by averaging the ranking results for all metrics. Note that the AR results cannot be compared across different tables since this metric is designed to showcase the dominance of each method within one table clearly. From Table A1, we observe that increasing both FPS and duration tends to enhance model capabilities, suggesting that denser frame sampling yields a more detailed video representation, thereby improving music generation. Nevertheless, to balance computational costs and performance, we use a 30-second duration at 2 FPS as our optimal setting.

Codebook Pattern. The exploration of codebook interleaving patterns has attracted attention from researchers across several domains [10, 35, 71, 78, 83]. In our ablation study focusing on the patterns, we find that while the Parallel and Vall-E [71] patterns align with the findings for text-to-music generation in MusicGen [10], the flattened codebook pattern does not consistently exceed the performance of the delay pattern in tasks of generating music from video. The delay pattern, notable for its relatively low computational cost, is therefore selected for our implementation. The results of this study are presented in Tab. A2.

Finetuning Effect. Our ablation study on the effects of the data scale during finetuning, as detailed in Table A3, highlights a balance between data size and model performance. Despite not performing best in all the metrics, the model finetuned with 20k pair data emerges as our choice. The 20k data offers a compelling trade-off: it significantly improves performance across key metrics without requiring the extensive computational resources that larger datasets demand. The results also validate the effectiveness of our

ranking strategy based on ImageBind-AV scores (detailed in Appendix 9), showing that prioritizing videos with higher audio-visual alignment improves finetuning data quality and enhances model performance.

9. Details of Dataset Construction

Coarse Filtering. We design a rule-based filtering strategy for initial data screening. First, we perform illegal video and audio filters, which filter out the video without an audio track or a video track. Next, we apply a duration filter to filter out videos based on their duration, excluding those that are either too long (over 480 seconds) or too short (under 30 seconds). Additionally, we implement a domain filter to examine metadata and exclude specific categories such as *Interview*, *News*, and *Gaming*, which often have background music that lacks semantic alignment with the visual content. We also filter out videos containing inappropriate content, such as violence or explicit material.

Fine-grained Filtering. To further ensure the quality of our data, we conducted additional audio and visual analyses. For the audio analysis, raw videos may contain audio segments without music, such as speech, silence, *etc.* To ensure the final dataset consists of high-quality video-music pairs, we retain only those videos with a larger portion of music content. We utilize the sound event detection model PANNs [34], which provides frame-level event labels across the entire video to identify music events. Based on the observation from a subset of videos, we define two thresholds, *i.e.*, a confidence threshold and a duration threshold, for analyzing the music event. The confidence threshold is set at 0.5, indicating an audio frame is considered a music event if the PANNs model predicts the probability of the “Music” label to be over 0.5. The duration threshold of a music event requires that at least 50% of the audio’s frames are classified as music events for the video to be considered valid.

For the visual analysis, some videos only consisting of static images will be removed. Specifically, we uniformly sample multiple temporal windows without overlap from the video. Within each window, we use Structural Similarity Index Measure (SSIM) [72] between the first frame and the last frame. By aggregating average SSIM values from all temporal windows, we remove the videos with average SSIM values lower than a threshold of 0.8, empirically.

Music Source Separation. Since the irrelevant human speech in videos poses a negative impact on music generation, we apply music source separation to process the videos. We employ Demucs [59] as the music source separation

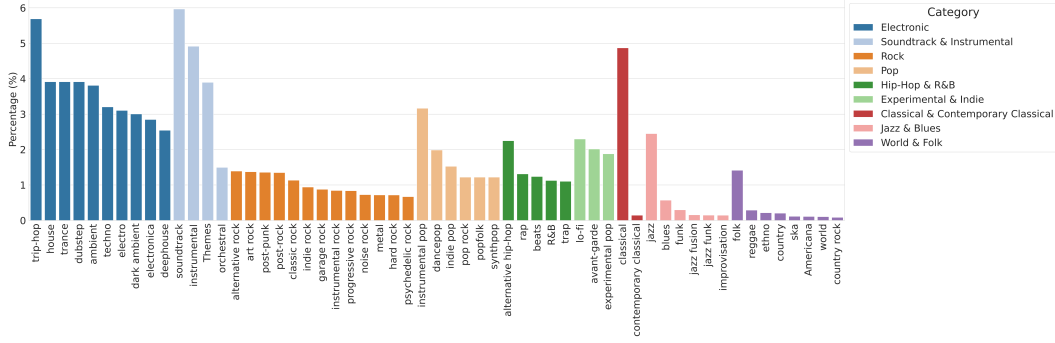


Figure A1. Distribution of music genres in the dataset, showcasing the diverse representation of genres such as electronic, classical, and jazz.

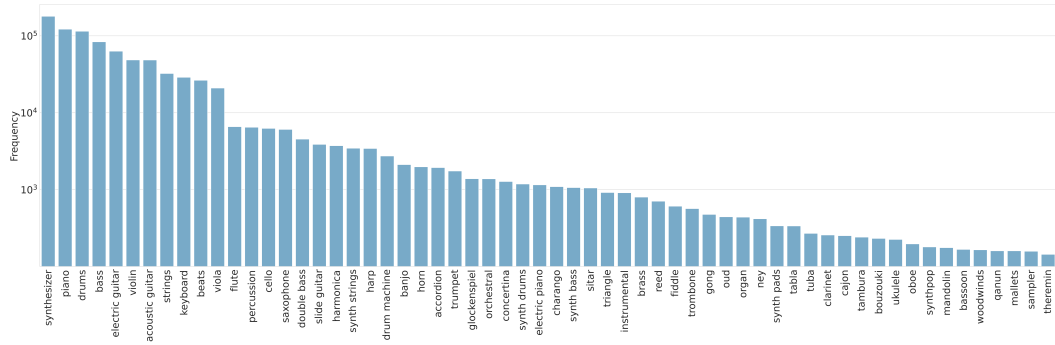


Figure A2. Distribution of instruments in the dataset, emphasizing the frequent usage of synthesizers, pianos, and drums, while also including diverse instruments such as violins and saxophones.

ration model to filter out the speech signals.

Audio-Video Alignment Ranking. ImageBind-AV [22] scores usually reflect the semantic correlation between the vision and audio modality. To construct a high-quality subset with better alignment, we compute the ImageBind-AV scores for all the data and rank them accordingly.

After filtering and ranking, we split the final videos into the training set, *V2M*, from all the paired data. The top 20K pairs are selected to form the finetuning subset, *V2M-20K*. In addition, we randomly sample 1,000 videos excluded from the training set. These 1,000 videos are then further evaluated by five human experts based on audio quality and the degree of audio-visual alignment. Ultimately, the top 300 high-quality videos are selected as a test set, termed as *V2M-bench*.

10. Additional Dataset Analysis

Music Genre Distribution. To better understand the diversity of our dataset, we analyze the distribution of music genres across all selected video-music pairs. The results are illustrated in Fig. A1. As shown, the dataset covers a wide range of genres, including but not limited to electronic, classical, pop, and rock. The diversity in genres ensures that the

dataset provides a comprehensive foundation for the task of video-to-music generation, enabling robust performance across various musical styles.

Instrument Usage Distribution. We also analyze the usage of different instruments within the dataset. The distribution is shown in Fig. A2. The frequent occurrences of synthesizers, pianos, and drums, along with a variety of other instruments, ensure the ability to capture diverse musical elements in the video-to-music generation task.

Mood Information. In addition to genres and instruments, we also explore the mood information present in the music data. A word cloud representation of the mood labels is shown in Fig. A3, where the font size corresponds to the frequency of each mood label. Commonly occurring moods include *inspiring*, *happy*, *dark*, *powerful*, and *sentimental*, showcasing the emotional diversity of the dataset. This emotional richness enhances the dataset’s capacity to generate music that aligns closely with the mood conveyed in videos.

All music-related metadata, including genre, instrument, and mood, is annotated using Qwen2-Audio, a state-of-the-art (SOTA) model for music understanding.

Figure A3. Word cloud of mood labels in the dataset, highlighting the diversity of emotions such as inspiring, happy, powerful, and dark.

Table A1. Ablation studies on video duration and FPS.

Duration(s)	FPS	Metrics						
		KL ↓	FD ↓	FAD ↓	density ↑	coverage ↑	Imagebind ↑	AR ↓
5	2	0.820	51.101	4.117	1.430	0.74	0.148	7.00
15	2	0.849	41.131	2.709	1.406	0.803	0.181	5.33
30	2	0.843	41.354	<u>2.413</u>	1.487	<u>0.840</u>	0.193	3.67
5	4	0.800	51.540	4.343	1.271	0.787	0.145	7.17
15	4	0.830	41.154	2.562	1.278	0.823	0.176	5.17
30	4	0.849	<u>40.032</u>	2.418	<u>1.538</u>	0.843	0.193	<u>2.84</u>
5	8	<u>0.819</u>	50.667	4.069	1.515	0.743	0.153	5.67
15	8	0.857	42.106	2.790	1.476	0.753	<u>0.187</u>	6.00
30	8	0.824	38.942	2.299	1.573	0.843	0.180	2.17

Table A2. Ablation studies on codebook pattern.

Patterns	Metrics					
	KL ↓	FD ↓	FAD ↓	density ↑	coverage ↑	Imagebind ↑
Parallel	0.921	68.603	18.243	0.562	0.183	0.166
Flatten	0.819	52.931	4.260	1.351	0.500	0.201
Delay	0.843	41.354	2.413	1.487	0.840	0.193
Vall-E	0.866	57.286	4.681	1.148	0.354	0.189

Table A3. Ablation studies on the ratio of finetuning data.

Finetuning Data	Metrics					
	KL ↓	FD ↓	FAD ↓	density ↑	coverage ↑	Imagebind ↑
0	0.712	38.184	3.956	1.125	0.583	0.181
10k	0.717	34.667	2.961	0.856	0.673	0.196
20k	0.734	29.946	2.459	1.250	0.730	0.202
40k	0.776	41.075	3.557	1.094	0.726	0.195
60k	0.828	40.160	2.844	0.977	0.660	0.192

11. Details of Evaluation Metrics

Fréchet Audio Distance (FAD) is a reference-free evaluation metric for assessing audio quality. Similar to Fréchet Image Distance (FID)[28], it compares the embedding statistics of the generated audio clip with ground truth audio clips. A shorter distance usually denotes better human-

perceived acoustic-level audio quality. However, this metric cannot reflect semantic-level information in audio. We report the FAD based on the VGGish[27] feature extractor.

Frchet Distance (FD) measures the similarity between generated samples and target samples in audio generation fields. It’s similar to FAD but uses a PANNs feature extractor instead. PANNs[34] have been pre-trained

on AudioSet[21], one of the largest audio understanding datasets, thus resulting in a more robust metric than FAD.

Kullback-Leibler Divergence (KL) reflects the acoustic similarity between the generated and reference samples to a certain extent. It is computed over PANNs’ multi-label class predictions.

Density and Coverage [55] measures the fidelity and diversity aspects of the generated samples. Fidelity measures how closely the generated samples match the real ones, while diversity assesses whether the generated samples capture the full range of variation found in real samples. We use CLAP[75] embeddings for manifold estimation.

Imagebind Score [22] assesses to what extent the generated music aligns with the videos. Despite the fact that Imagebind extends the CLIP to six modalities, we only use the branches of audio and vision. Since we use ImageBind to filter out video-audio pairs with a low matching score during dataset construction, the ImageBind score is naturally used in our evaluation. We acknowledge that ImageBind is not specifically trained on music data, which may limit its effectiveness in capturing the full complexity of video-music alignment. However, it remains the most suitable option available for this task at present.

12. Details of the Inference Process

When predicting music on videos of arbitrary length, maintaining music consistency and coherence is particularly important. However, it leads to a significant challenge on computational resources due to the quadratic dependency of transformers-based models on sequence length [5, 82].

To address this problem, we adopt a sliding window approach for inferring the whole video. During inference, given an input video with a length of L , we define L_s as the length of the sliding window and O as the overlap between consecutive windows. With the window start position t initially set to 0, the inference involves the following steps compactly while $t + L_s \leq L$: (1) using a visual encoder to extract feature representations \mathbf{X} and capture long-term dependencies \mathbf{X}_l ; (2) collecting embeddings within the window $[t, t + L_s]$ to obtain \mathbf{X}_s ; (3) predicting the music tokens $\bar{\mathbf{Y}}$ for the reduced window $[t, t + L_s - O]$ based on \mathbf{X}_l and \mathbf{X}_s ; (4) decoding $\bar{\mathbf{Y}}$ to the predicted audio $\bar{\mathbf{A}}$ using the audio decoder; (5) move the window forward by setting $t = t + L_s - O$, and repeating steps (2) to (5) until the end of the video.

After finishing the above steps, we can concatenate all musical segments to form a cohesive piece of music that aligns in duration with the video.

13. Qualitative Analysis

In Fig. A5, our qualitative analysis highlights specific limitations of CMT, Video2Music, and M²UGen. CMT and

Video2Music extract visual cues to generate symbolic music, *i.e.*, MIDI notes. However, CMT’s training strategy for symbolic music generation leads to discontinuities, particularly for slowly changing or static frames, where the model fails to predict symbolic music notes, resulting in periods of silence. Additionally, the approach of predicting MIDI notes and then rendering them into audio, as employed by both CMT and Video2Music, lacks high-frequency content, negatively affecting auditory perception. M²UGen utilizes LLMs to fuse multimodal representation and then project LLMs’ embeddings into music via a text-to-music generation model. However, this approach relies on text embeddings as intermediaries, which causes the loss of visual information and restricts the model’s ability to detect nuanced visual variations. As a result, the music generated by this method usually showcases repetitive musical themes and suffers from a lack of diversity, as evidenced in Fig. A5 and the supplementary videos. The last row of Fig. A5 demonstrates that our Long-Short-Term (LST) approach is capable of generating music that is rich in diversity and semantically consistent with the video.

14. User Study Interface

Fig. A4 illustrates the A/B test interface used during the user study. Participants evaluated the videos based on four criteria: Audio Quality, Video-Music Alignment, Musicality, and Overall Assessment. This interface shows participants comparing two videos side-by-side and selecting the better one for each criterion.

15. Supplementary Videos

For additional insights and demonstrations, we kindly refer readers to our supplementary video for a comprehensive showcase of our method’s performance.

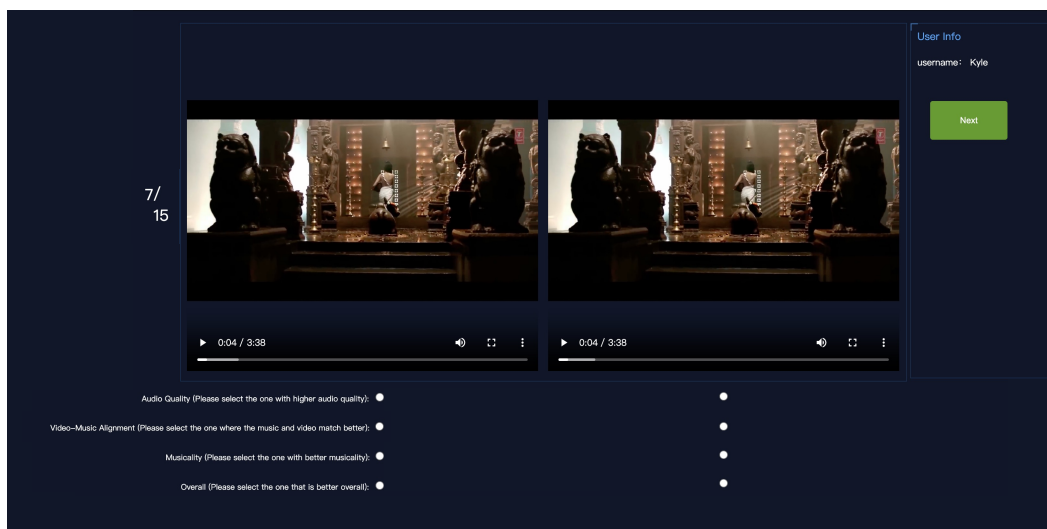


Figure A4. User study process. Participants evaluate the videos based on four criteria: Audio Quality, Video-Music Alignment, Musicality, and Overall Assessment.

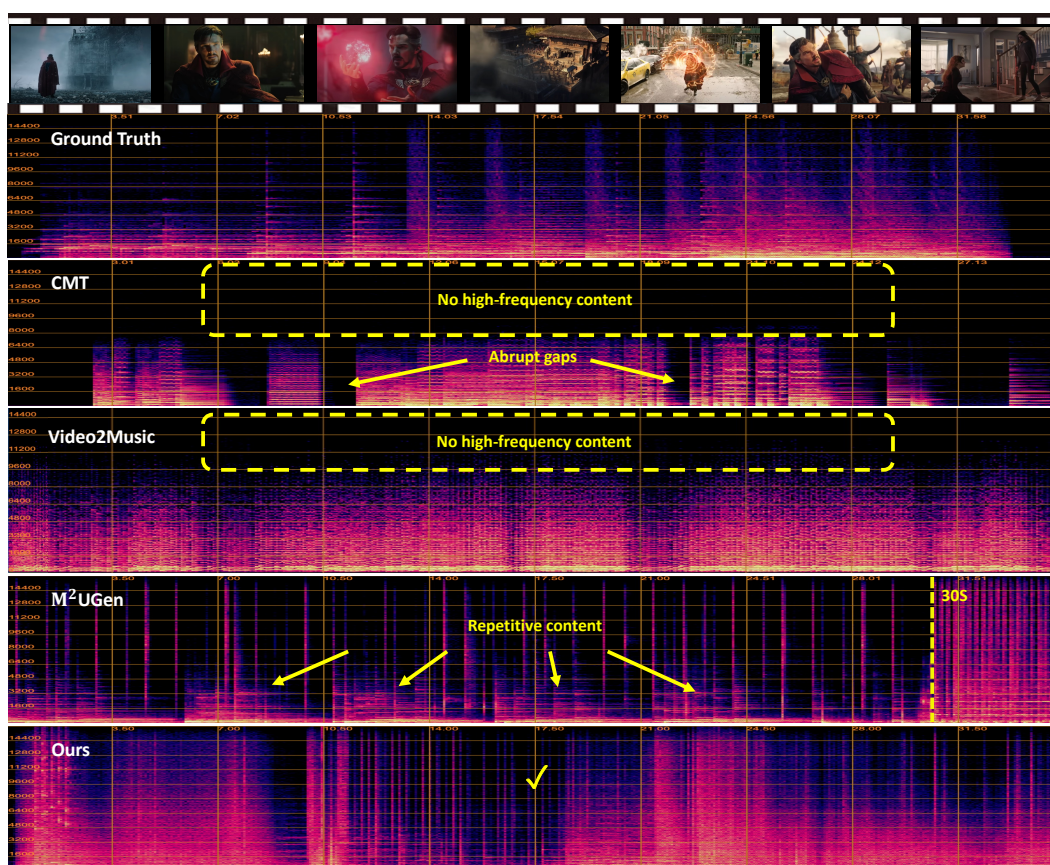


Figure A5. Qualitative Comparison results on sound spectrograms produced by different methods.