MULTIMODAL MUSIC GENERATION WITH EXPLICIT BRIDGES AND RETRIEVAL AUGMENTATION

Anonymous Authors
Anonymous Affiliations
anonymous@ismir.net

ABSTRACT

Multimodal music generation aims to produce music from diverse input modalities such as text, videos, and images. Existing methods typically rely on a shared embedding space, but face challenges such as insufficient cross-modal data, weak cross-modal alignment, and limited controllability. This paper addresses these issues by introducing explicit bridges of text and music for improved alignment. We propose a novel pipeline for constructing multimodal music datasets, yielding two new datasets, namely **MTV-24K** and **MT-512K**, all annotated with rich musical attributes. Additionally, we propose **MTM**, a Multimodal-to-Text-to-Music framework based on these bridges. Experiments on video-to-music, image-to-music, text-to-music, and controllable music generation tasks demonstrate that MTM significantly improves music quality, modality alignment, and user controllability compared to existing methods.

1. INTRODUCTION

Recent advances in generative models have made notable progress in text-to-music generation [1–3], but multimodal music generation (from text, images, or videos) remains challenging. Existing methods rely on common embedding spaces for alignment [4–6] but struggle to maintain accurate, high-quality cross-modal correlations.

We argue that key limitations of multimodal music generation include: (1) scarce large-scale multimodal music datasets, with existing ones lacking fine-grained musical annotations; (2) over-reliance on low-level visual cues without semantic understanding of mood or themes; (3) under-utilization of distinct, complementary modality contributions (e.g., text for semantics, videos for temporal dynamics); (4) poor interpretability and controllability over musical attributes in joint embedding approaches.

To address these, we propose **using text and music as explicit alignment bridges**. Text bridge converts visuals to detailed music descriptions via a multimodal model, guiding generation. Music bridge introduces retrieval-augmented generation (RAG) for thematic alignment and attribute control. These bridges complement each other and help to enhance controllability.

Based on this principle, we present MTM, a novel method generating music from text, video, or images, with four core components: (1) a 24K multimodal video-music-text dataset with rich musical annotations; (2) a multimodal music description model (based on InternVL2 [7]) translating visuals to structured musical descriptions; (3) dual-track retrieval (broad for fine-grained guidance, targeted for attribute control); (4) explicitly conditioned generation integrating both bridges via a diffusion transformer. The overview of our framework is shown in Fig. 1.

We conduct extensive experiments on video, text, and image-to-music, visual-to-description, and controllable music generation tasks. Results show that MTM enhances music quality, strengthens cross-modal alignment, and improves controllability, making it a robust and flexible framework with broad multimedia applications.

2. RELATED WORK

Multimodal Music Generation. Conditional music generation has advanced in both symbolic and audio domains, with representative text-to-music systems improving fidelity and diversity [1–3, 8–11]. For video-to-music (V2M), MIDI-based methods translate visual rhythm into symbolic patterns [12–15], while audio-based models exploit large corpora and strong encoders to learn richer correspondences [16, 17]. However, limited paired data leads to weak high-level semantic alignment and poor controllability. Any-to-music work such as NExT-GPT [5] and [6] treats music as generic audio, and MuMu-LLaMA [4] fuses modality features via an LLM, but these feature-level schemes under-use the structure of music and offer little user control.

Retrieval-augmented Generation. Retrieval-augmented generation (RAG) has become a common technique to enhance the fidelity and diversity of language models with an additional knowledge base in the field of natural language processing [18–21]. Due to its effectiveness, RAG is further generalized to a range of fields and tasks, such as visual recognition [22, 23], image generation [24, 25], 3D generation [26, 27], and drug discovery [28]. To the best of our knowledge, we propose the first RAG pipeline for multimodal music generation.

Attribution: F. Author, S. Author, and T. Author, “Multimodal Music Generation with Explicit Bridges and Retrieval Augmentation”, submitted to *ISMIR*, 2025.

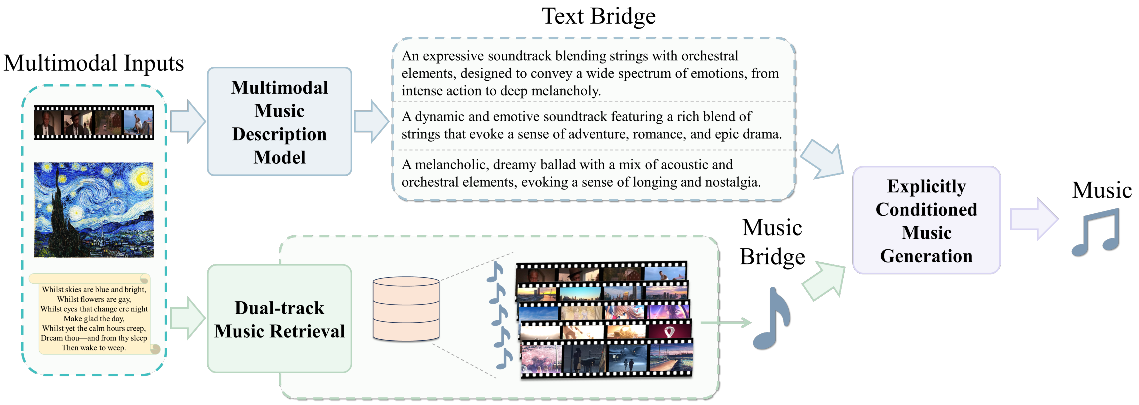

Multimodal Inputs
Whilst skies are blue and bright,
Whilst flowers are gay,
Whilst eyes that change ere night
Make glad the day,
Whilst yet the calm hours creep,
Dream thou—and from thy sleep
Then wake to weep.
Multimodal Music Description Model
Text Bridge
An expressive soundtrack blending strings with orchestral elements, designed to convey a wide spectrum of emotions, from intense action to deep melancholy.
A dynamic and emotive soundtrack featuring a rich blend of strings that evoke a sense of adventure, romance, and epic drama.
A melancholic, dreamy ballad with a mix of acoustic and orchestral elements, evoking a sense of longing and nostalgia.
Dual-track Music Retrieval
Music Bridge
Explicitly Conditioned Music Generation
Music

Figure 1: **Overview of the MTM framework.** We employ text and music as two explicit bridges for multimodal music generation. Text-form music description is obtained with the Multimodal Music Description model. Reference music is retrieved with the Dual-track Music Retrieval module. The two bridges are fed into the Explicitly Conditioned Music Generation module to generate output music.

3. METHODOLOGY

In this section, we present our multimodal music generation framework (Fig. 1.), including the following key components: (1) a **Multimodal Music Dataset** which effectively aligns music, video and text; (2) a **Multimodal Music Description Model (MMDM)**, which generates textual descriptions for emotional and thematic framing, providing the textual bridge; (3) a **Dual-track Music Retrieval (DMR)** module, which retrieves relevant music tracks based on input criteria, providing the musical bridge; and (4) an **Explicitly Conditioned Music Generation (ECMG)** module, which uses both bridges to generate the final music composition.

3.1 Multimodal Music Dataset

High-quality datasets that effectively align music, video, and text remain a critical bottleneck in multimodal music generation research. Existing datasets (Tab. 1) are small, weakly annotated, or single-modal, failing to capture the intricate relationships between audio, visual, and textual elements. To address this gap, we introduce **MTV-24K**, a dataset of **24K music-video-text triplets** with fine-grained musical attributes and descriptions. This dataset opens new possibilities for training and evaluating advanced generative models conditioned on multimodal inputs. With the same pipeline we further collect MT-512K, 512k music-text pairs that fine-tune the T2M model further demonstrating the high quality and reliability of our data construction process. For details on the construction of the datasets, please refer to the Appendix.

Dataset	Genre	Desc.	SrcSep.	Attr.	Size
SymMV [13]	✓	✗	✗	✓	1,140
DISCO-MV [29]	✗	✗	✗	✗	2200K
MUVideo [4]	✗	Coarse	✗	✗	14.5K
BGM909 [15]	✓	Fine	✓	✓	909
V2M [17]	✗	✗	✗	✗	360K
MTV-24K (Ours)	✓	Fine	✓	✓	24K

Table 1: **Comparison between different multimodal music datasets.** This table compares datasets based on their genre, music description (Desc.), source separation (SrcSep), music attributes (Attr.), and size.

3.2 Multimodal Music Description Model

To bridge the gap between multimodal information and the music modality, our approach employs textual descriptions as the first explicit bridge, effectively linking complex multimodal inputs with music generation. Leveraging the cross-modal understanding and generation ability of multimodal large language models (MLLMs) [30–32], we introduce the Multimodal Music Description Model (MMDM), which is built on InternVL2 [7] with MTV-24K to map multimodal signals into textual music description to guide the music generation process.

3.3 Dual-track Music Retrieval

Musicians composing for multimedia often start by referencing existing music—an approach AI can mirror via retrieval-augmented generation, especially when datasets are limited. To enhance multimodal music generation and address alignment challenges (e.g., text-to-music models struggling with chord progressions or rhythm), we introduce Dual-track Music Retrieval, using music as an explicit bridge to mitigate language-music gaps.

We use music embeddings to form an in-domain link between diverse inputs (text, images) and target audio. As shown in Fig. 2, our pipeline combines two retrieval strategies on the MTV-24K dataset:

Broad retrieval for Fine-grained Control. For each conditional modality, we first retrieve relevant music by computing similarities using CLIP [33] or CLAP [34] embeddings. Given a target music, we directly use CLAP to compute audio embeddings and compare them with embeddings in our music database. The retrieved music $\mathbf{m}_{i^*}$ provides fine-grained conditions (e.g., melody, rhythm). At inference, CLIP embeddings of text/visuals can replace

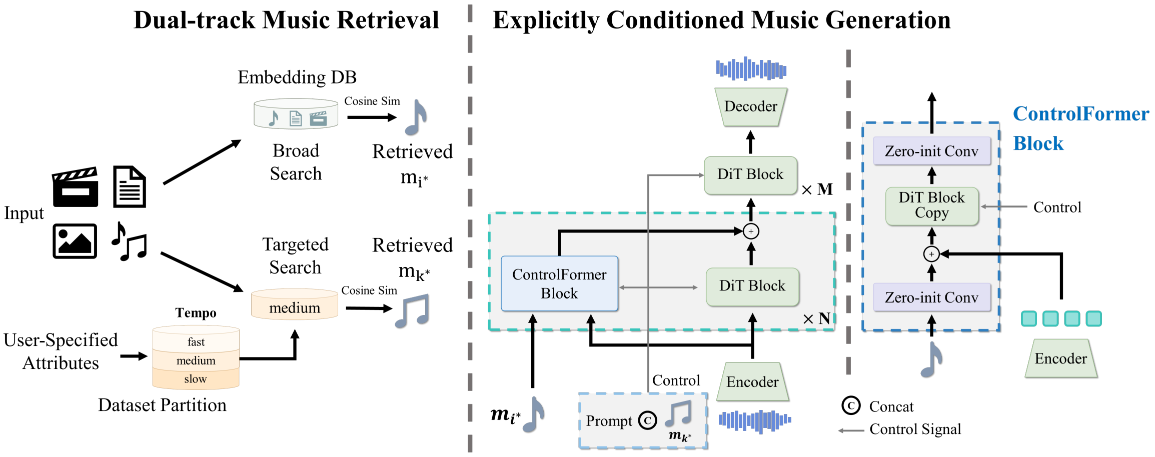


Figure 2: **Framework of Dual-track Music Retrieval and Explicitly Conditioned Music Generation.** The left part illustrates the Dual-track Music Retrieval process, which leverages our multimodal dataset to perform both broad and targeted retrieval. The right part shows the Explicitly Conditioned Music Generation pathway, where music is generated through a ControlFormer block integrating embeddings from selected music bridge, text bridge, and noisy inputs.

audio inputs to align with semantic context.

Targeted retrieval for Coarse-Grained Control. In parallel to the broad retrieval, we conduct a targeted retrieval based on musical attributes labeled in MTV-24K. For instance, the tempo partition is categorized into "fast", "medium", and "slow" subsets. This allows users to flexibly select songs that match specific attributes among genre, tempo, or mood partition. To retrieve a fast-paced song, we query directly within the "fast" subset in the tempo partition. In each subset, the most suitable music piece is determined by computing the cosine similarity between the CLAP embeddings of the desired textual attribute. The embedding of the retrieved music is fed into the subsequent modules.

3.4 Explicitly Conditioned Music Generation

Our base model uses a latent diffusion transformer (DiT) [35], built on Stable Audio Open [3]. Music is encoded into a latent space via a pre-trained VAE, with Gaussian noise added. Transformer blocks process the noisy input, incorporating T5-encoded text features and timestep via cross-attention, optimized with the diffusion v-prediction objective [36].

Music ControlFormer. To integrate fine-grained guidance from broad retrieval, we adapt ControlNet principles [37, 38] for diffusion transformers (Fig. 2). Instead of full model duplication, we replicate early transformer layers to form a ControlFormer branch, balancing efficiency and structural alignment. At each layer, the main branch produces a hidden state, and the ControlFormer produces another. We combine these hidden states by element-wise addition To ensure stable training and prevent disrupting the extracted representations from the pretrained model, we initialize the input and output convolution layers of the ControlFormer to zero.

Stylization Module. For coarse-grained attribute control from targeted retrieval, we fuse retrieved music characteristics into generation by: (1) appending attribute descriptions to the input prompt; (2) concatenating CLAP embeddings of retrieved music with text/timestep embeddings to form a unified conditional representation (adjusted to text embedding dimensions); (3) using cross-attention to integrate this into noisy music, aligning generated output with specified attributes.

4. EXPERIMENTS

In this section, we comprehensively evaluate MTM on video-to-music (V2M), text-to-music (T2M), and ablation studies. For details of our dataset and full experimental setup, and additional results on *image-to-music generation, visual-to-description, and controllable generation*, please refer to the appendix.

We use objective metrics including KL_{passt} [39], FD_{openl3} [40], CLAPScore [34], ImageBind score (IB) [41], and BeatMSE. KL_{passt} and FD_{openl3} evaluate the music quality from its statistical similarity with real music data and perceptual audio quality. CLAPScore and IB measure the cross-modal semantic alignment between text, videos or images and the corresponding generated music. BeatMSE calculates the Mean Squared Error (MSE) between the ground truth and generated music, specifically evaluating the alignment of rhythmic patterns.

4.1 Video-to-music Generation

We evaluate the proposed MTM model on the V2M task, where the goal is to generate background music for input videos with both high-quality audio and strong video-music alignment. We benchmark the performance of MTM on SymMV [13] dataset against several state-of-the-art methods, including symbolic methods (CMT [12], Video2music [14] and Diff-BGM [15]) and audio methods (VidMuse [17] and MuMu-LLaMA [4]). Apart from objective metrics, we also conduct subjective evaluation to ask participants to rate music tracks on Musical Pleasantness (MP), Emotional Correspondence (EC),

Method	Output	Objective Metrics				Subjective Metrics↑			
		KL$_{passt}$↓	FD$_{openl3}$↓	IB↑	BeatMSE ↓	MP	EC	TC	RC
CMT [12]	MIDI	52.76	269.63	8.54	1748.1	3.06	2.68	2.72	3.04
Video2music [14]	MIDI	103.56	533.46	5.26	**943.4**	2.93	2.53	2.59	2.53
Diff-BGM [15]	MIDI	104.28	472.53	10.29	1842.3	3.10	2.92	2.77	2.74
MuMu-LLaMA [4]	Audio	60.41	180.72	15.58	1388.1	2.98	2.44	2.44	2.71
VidMuse [17]	Audio	56.48	187.13	22.09	1427.2	3.21	2.98	3.06	3.16
MTM (ours)	Audio	**47.12**	**101.43**	**22.93**	1172.1	**3.85**	**3.40**	**3.40**	**3.64**

Table 2: Video-to-music generation performance on SymMV.

Method	Objective Metrics				Subjective Metrics↑	
	KL$_{passt}$↓	FD$_{openl3}$↓	CLAPScore↑	IB↑	MP	TMA
AudioLDM [11]	99.85	293.86	17.61	20.01	2.31	2.65
MusicGen [2]	46.89	181.59	33.95	22.46	3.12	3.33
MuMu-LLaMA [4]	49.03	188.84	28.76	16.70	3.21	3.19
Stable Audio Open [3]	42.89	183.09	40.92	24.67	3.42	3.51
MTM (ours)	**38.28**	**134.34**	**41.28**	**29.36**	**3.78**	**3.57**

Table 3: Text-to-music generation performance on SongDescriber dataset.

B	T	KL↓	FD↓	IB↑
✓	✓	**75.3**	**177.3**	**24.7**
✓	×	91.9	199.7	20.7
×	✓	91.1	387.1	20.5
×	×	96.4	360.3	14.7

(a)

Method	KL↓	FD↓	CLAPScore↑
Retrieval	56.6	163.0	37.0
Full	**38.3**	**134.3**	**41.3**

(b)

Table 4: (a) Effect of broad (B) and target (T) retrieval on video-to-music performance. (b) Comparison of retrieval-only and full models.

Thematic Correspondence (TC), and Rhythmic Correspondence (RC) using a 5-point Likert scale. A total of 80 valid responses were gathered. More details are in the appendix.

Results. As shown in Tab. 2, MTM achieves the best performance across all objective metrics excluding BeatMSE, indicating superior realism, perceptual quality, and strong semantic alignment with video content. Subjective evaluations highlight MTM's ability to enhance video narratives emotionally and thematically, thanks to the DMR module, which dynamically synchronizes music with visuals. While MTM does not explicitly optimize local rhythmic accuracy, it achieves better global rhythm coherence than baselines, avoiding overfitting to local beats observed in CMT and Diff-BGM. This balance ensures natural musical flow, enhancing both immersion and storytelling.

4.2 Text-to-music Generation

We evaluate the MTM model on the T2M task, where the goal is to generate music that aligns with a given text description. We use the SongDescriber [42], a curated collection specifically for this task, excluding samples with vocal parts to maintain instrumental coherence. We compare MTM with several state-of-the-art text-to-music models: AudioLDM [11], MusicGen [2], Stable Audio Open [3] and MuMu-LLaMA [4].

Results. As presented in Tab. 3, MTM outperforms baseline models across the objective metrics. It achieves the lowest KL$_{passt}$ and FD$_{openl3}$, indicating that the music generated by MTM is statistically closer to real-world music distributions and exhibits higher perceptual quality. MTM also achieves a highest CLAPScore and ImageBind-Score, further validating its robust performance in aligning text with music. Subjectively, MTM excels with the highest MP and TMA score, suggesting it generates musically coherent and contextually accurate compositions.

4.3 Ablation Studies

We assess the individual contributions of broad retrieval (BR) and targeted retrieval (TR) in the video-to-music task using the SymMV dataset. As shown in Tab. 4(a), combining BR and TR achieves the best performance across all metrics, demonstrating their complementary roles in music generation. Removing either component degrades performance, underscoring their complementary roles in music generation.

As shown in Tab. 4(b), the retrieval-only model directly uses retrieved music as output, leading to weak alignment with input descriptions and higher KL and FD scores. In contrast, the full model outperforms it across all metrics, showing the necessity of generation.

5. CONCLUSION

We presented **MTM**, a multimodal music generator that treats *text* and *retrieved music* as explicit bridges to align and control generation from text, images, and video. Built on the new *MTV-24K* dataset and a dual-track retrieval module, MTM delivers state-of-the-art audio quality and fine-grained controllability on V2M, I2M, and T2M benchmarks. These results point to MTM's potential as a scalable engine for personalised, context-aware soundtracks in entertainment, XR, and other interactive media.

6. REFERENCES

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

[2] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," *arXiv preprint arXiv:2407.14358*, 2024.

[4] S. Liu, A. S. Hussain, Q. Wu, C. Sun, and Y. Shan, "Mumu-llama: Multi-modal music understanding and generation via large language models," *arXiv preprint arXiv:2412.06660*, vol. 3, no. 5, p. 6, 2024.

[5] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," *arXiv: 2309.05519*, 2023.

[6] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal, "Any-to-any generation via composable diffusion," *NeurIPS*, vol. 36, 2024.

[7] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma *et al.*, "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites," *arXiv:2404.16821*, 2024.

[8] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *AAAI*, 2018.

[9] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, I. Simon, C. Hawthorne, N. Shazeer, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer: Generating music with long-term structure," in *ICLR*, 2019.

[10] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *MM*, 2020.

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

[12] S. Di, Z. Jiang, S. Liu, Z. Wang, L. Zhu, Z. He, H. Liu, and S. Yan, "Video background music generation with controllable music transformer," in *MM*, 2021.

[13] L. Zhuo, Z. Wang, B. Wang, Y. Liao, C. Bao, S. Peng, S. Han, A. Zhang, F. Fang, and S. Liu, "Video background music generation: Dataset, method and evaluation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 637–15 647.

[14] J. Kang, S. Poria, and D. Herremans, "Video2music: Suitable music generation from videos using an affective multimodal transformer model," *Expert Systems with Applications*, p. 123640, 2024.

[15] S. Li, Y. Qin, M. Zheng, X. Jin, and Y. Liu, "Diff-bgm: A diffusion model for video background music generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 348–27 357.

[16] K. Su, J. Y. Li, Q. Huang, D. Kuzmin, J. Lee, C. Donahue, F. Sha, A. Jansen, Y. Wang, M. Verzetti *et al.*, "V2meow: Meowing to the visual beat via music generation," *arXiv preprint arXiv:2305.06594*, 2023.

[17] Z. Tian, Z. Liu, R. Yuan, J. Pan, Q. Liu, X. Tan, Q. Chen, W. Xue, and Y. Guo, "Vidmuse: A simple video-to-music generation framework with long-short-term modeling," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18 782–18 793.

[18] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[19] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.

[20] P. Zhang, S. Xiao, Z. Liu, Z. Dou, and J.-Y. Nie, "Retrieve anything to augment large language models," *arXiv preprint arXiv:2310.07554*, 2023.

[21] D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1818–1826.

[22] A. Long, W. Yin, T. Ajanthan, V. Nguyen, P. Purkait, R. Garg, A. Blair, C. Shen, and A. van den Hengel, "Retrieval augmented classification for long-tail visual recognition," in *CVPR*, 2022, pp. 6959–6969.

[23] Z. Hu, A. Iscen, C. Sun, Z. Wang, K.-W. Chang, Y. Sun, C. Schmid, D. A. Ross, and A. Fathi, "Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 23 369–23 379.

[24] S. Sheynin, O. Ashual, A. Polyak, U. Singer, O. Gafni, E. Nachmani, and Y. Taigman, "Knn-diffusion: Image generation via large-scale retrieval," *arXiv preprint arXiv:2204.02849*, 2022.

[25] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," *arXiv preprint arXiv:2209.14491*, 2022.

[26] R. Wu and C. Zheng, "Learning to generate 3d shapes from a single example," *arXiv preprint arXiv:2208.02946*, 2022.

[27] Z. Wang, T. Wang, G. Hancke, Z. Liu, and R. W. Lau, "Themestation: Generating theme-aware 3d assets from few exemplars," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.

[28] M. Aguilar Rangel, A. Bedwell, E. Costanzi, R. J. Taylor, R. Russo, G. J. Bernardes, S. Ricagno, J. Frydman, M. Vendruscolo, and P. Sormanni, "Fragment-based computational design of antibodies targeting structured epitopes," *Science Advances*, vol. 8, no. 45, p. eabp9540, 2022.

[29] L. Lanzendörfer, F. Grötschla, E. Funke, and R. Wattenhofer, "Disco-10m: A large-scale music dataset," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv: 2309.17421*, vol. 9, 2023.

[31] G. Luo, X. Yang, W. Dou, Z. Wang, J. Dai, Y. Qiao, and X. Zhu, "Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training," *arXiv preprint arXiv:2410.08202*, 2024.

[32] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," *arXiv: 2310.03744*, 2023.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[34] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[35] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *CVPR*, 2023, pp. 4195–4205.

[36] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[37] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023.

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

[39] K. Koutini, J. Schlüter, H. Eghbal-Zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," *arXiv preprint arXiv:2110.05069*, 2021.

[40] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP.* IEEE, 2019, pp. 3852–3856.

[41] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind one embedding space to bind them all," in *CVPR*, 2023, pp. 15 180–15 190.

[42] I. Manco, B. Weck, S. Doh, M. Won, Y. Zhang, D. Bogdanov, Y. Wu, K. Chen, P. Tovstogan, E. Benetos *et al.*, "The song describer dataset: a corpus of audio captions for music-and-language evaluation," *arXiv preprint arXiv:2311.10057*, 2023.

[43] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[44] S. Deshmukh, D. Alharthi, B. Elizalde, H. Gamper, M. A. Ismail, R. Singh, B. Raj, and H. Wang, "Pam: Prompting audio-language models for audio quality assessment," *arXiv preprint arXiv:2402.00282*, 2024.

[45] OpenAI, "Gpt-4v(ision) system card," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263218031

[46] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[47] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[48] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," *arXiv preprint arXiv:2307.08691*, 2023.

[49] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," *arXiv: 2312.14238*, 2023.

[50] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *ICLR*, 2022.

[51] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, vol. 21, pp. 140:1–140:67, 2020.

[52] I. Jang, Z. Yang, Z. Zhang, X. Jin, and M. Chowdhury, "Oobleck: Resilient distributed training of large models using pipeline templates," in *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023, pp. 382–395.

[53] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[54] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "Zero: Memory optimizations toward training trillion parameter models," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2020, pp. 1–16.

[55] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," *arXiv preprint arXiv:2211.01095*, 2022.

[56] H. Fang, P. Xiong, L. Xu, and Y. Chen, "Clip2video: Mastering video-text retrieval via image clip," *arXiv preprint arXiv:2106.11097*, 2021.

A. CONSTRUCTION OF THE MTV-24K DATASET

A.1 Data Collection and Annotation

We design a pipeline to collect diverse music-video pairs and produce high-quality, context-rich textual annotations. The pipeline consists of three primary stages: MV collection, auto-tagging, and ground truth generation, as summarized in Fig. 3. This multi-stage approach was designed to ensure comprehensive coverage of music modalities and robust alignment between audiovisual content and associated descriptions.

Music Video Data Collection. Data collection phase targets music videos (MVs) since they exhibit strong correspondence between visual content and musical expression. Each entry in the dataset contains:

- **Video frames** capturing key scenes and transitions.
- **Music tracks** processed with hybrid Demucs [43] for source separation to isolate instrumentals.
- **Metadata** such as titles, artist information, and video descriptions from platforms like YouTube and Shazam, providing supplementary context for each music-video pair.

Automatic Music Description Generation. After constructing the initial MV dataset, we proceed to generate music descriptions from the raw data. This process begins with an auto-tagging method that annotates each audio file across dimensions such as emotional tone and music theory elements. Subsequently, large language models are employed to formulate unstructured metadata and tags into detailed music descriptions in natural language.

The auto-tagging process is grounded in a comprehensive label set developed by music experts, covering essential categories such as instruments, genres, and emotions for standardized and interpretable tagging. CLAP [34] embeddings of candidate tags are extracted and align candidate tags with audio CLAP embeddings through cosine similarity. To ensure high precision in labeling, we use carefully calibrated similarity thresholds to filter out low-alignment tags.

To generate detailed music descriptions, we combine auto-generated tags and metadata with an LLM used as a paraphraser. We design detailed, structured prompts to constrain generation scope, ensuring outputs remain semantically faithful to the original data. This strategy produces grounded descriptions that mitigate hallucinations and improve text-audio alignment.

A.2 Quality Validation

Quality assurance combines automatic filtering and expert review. CLAP similarity removes low-alignment tags; PAMScore [44] and outlier detection drop noisy pairs; a manual audit of 10% of samples tunes thresholds. This two-tier protocol yields finer multimodal consistency than prior MV datasets that rely solely on raw metadata. More details are provided in the supplementary.

Method	$\mathrm{KL}_{passt}\downarrow$	$\mathrm{FD}_{openl3}\downarrow$	IB↑
CoDi [6]	216.48	251.52	9.60
MuMu-LLaMA [4]	128.33	247.42	2.28
MTM (ours)	**98.78**	**116.71**	**12.10**

Table 5: Image-to-music generation performance.

Model	CLAPScore↑
GPT-4V [45]	44.41
InternVL [7]	44.21
MuMu-LLaMA [4]	41.91
MMDM	50.88

Table 6: Video-to-description generation performance.

B. ADDITIONAL EXPERIMENTS

B.1 Image-to-music Generation

To demonstrate the versatility of our proposed framework, we evaluate the MTM model on the image-to-music (I2M) generation task. In this setting, models are required to generate music that aligns semantically and emotionally with a given image. Evaluations are conducted using the MUImage [4] dataset, a high-quality collection of image-music pairs curated for nuanced cross-modal generation. We use 1,500 image-music pairs as the test set.

Baselines. We benchmark MTM against state-of-the-art models, including CoDi [6] and MuMu-LLaMA [4]. CoDi, an any-to-any generation model, supports image-to-music, making it a strong baseline for comparison.

Results.Tab. 5 summarizes the performance of MTM, CoDi, and MuMu-LLaMA on the I2M task. MTM demonstrates a substantially lower KL_{passt} and FD_{openl3} compared to the baselines, alongside a higher ImageBind Score (IB), indicating its ability to generate music that aligns closely with real-world distributions. Notably, although MTM is not explicitly trained on images, it effectively captures both the semantic and emotional content of visual inputs, underscoring its strong generalizability and its capacity to produce perceptually high-quality music from images.

B.2 Visual-to-Description Generation

We evaluate our MMDM model on a subset of 8,042 dynamic videos filtered from the DISCO-200K-high-quality dataset [29]. To assess the quality of the generated descriptions, we use the CLAPScore to measure the similarity between the original and generated textual descriptions in terms of their alignment with the corresponding music. Since MuMu-LLaMA [4] focuses on user interaction, the output typically begins with 'Here is a piece of music that is [music description].' We then parse the output and filter out any abnormal results.

In Tab. 6, MMDM achieves highest CLAPScore. While both excel in multimodal tasks, they are not specialized for music-text generation, leading to lower performance.

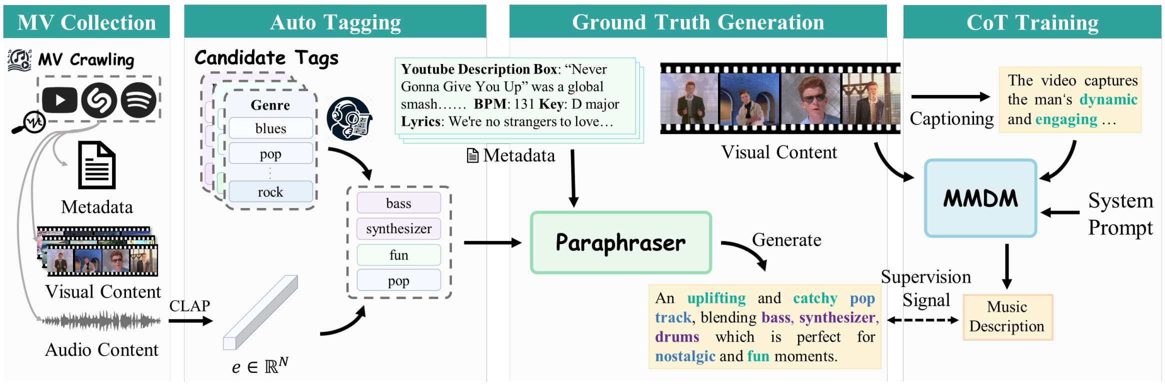

MV Collection
Auto Tagging
Ground Truth Generation
CoT Training
MV Crawling
Metadata
Visual Content
Audio Content
CLAP
Candidate Tags
Genre
blues
pop
rock
bass
synthesizer
fun
pop
$e \in \mathbb{R}^N$
Youtube Description Box: "Never Gonna Give You Up" was a global smash...... BPM: 131 Key: D major Lyrics: We're no strangers to love...
Metadata
Visual Content
Paraphraser
Generate
An uplifting and catchy pop track, blending bass, synthesizer, drums which is perfect for nostalgic and fun moments.
Captioning
The video captures the man's dynamic and engaging ...
MMDM
System Prompt
Supervision Signal
Music Description

Figure 3: **Pipeline of constructing MTV-24K and the Multimodal Music Description Model (MMDM).** This process starts with the collection of music videos, followed by automated tagging to refine audio annotations using CLAP embedding similarities. Metadata and thematic descriptions are paraphrased by LLM to create training targets. The training utilizes LoRA fine-tuning in the MMDM to transform multimodal inputs into targeted music descriptions that align with the visual content's themes.

Attribute	Δ
Instrument	+11.46
Genre	+3.03
Mood	+4.14

Table 7: Average change in CLAPScore when controlling different attributes.

Metric	**Mood**	**Genre**	**Instrument**	**Beat**
Agreement (%)	95.2	91.0	94.3	92.8
Likert Score (1–5)	4.2	3.8	4.2	4.1

Table 8: User study of controllability.

MuMu-LLaMA struggles with multimodal fusion and reasoning, further limiting its effectiveness. These results highlight MMDM's superior ability to generate accurate and contextually relevant music descriptions.

B.3 Controllable Generation

To evaluate MTM's controllability, we examine its ability to modify music attributes. For each attribute we define a binary contrast, for example "happy" versus "non-happy" for mood, and randomly select 20 songs that do not fall clearly on either side. From every seed we generate 10 variants conditioned only on the target attribute while keeping the rest of the prompt unchanged. We then compute the mean change in CLAPScore, Δ, between each controlled output and its baseline, using the target-attribute text as the CLAP query. The results, listed in Tab. 7, show that a larger positive Δ corresponds to closer alignment with the requested attribute.

To further evaluate the controllability of our model, we test its ability to adjust the tempo of the generated music. Similar to other music attributes such as genre, instrument, and mood, we report the average change in tempo on the generated variations to quantify the model's performance.

For this experiment, we categorize the dataset into distinct beats per minute (BPM) groups: "Fast," "Medium," and "Slow." We randomly sample 20 songs for each BPM group and generate 10 variations for each song, conditioned on the sampled song itself. This setup allows us to assess the model's capability to independently control tempo while maintaining the overall coherence of the generated music.

We calculate the average BPM across the 200 generated songs per group (20 songs × 10 variations) to measure how well the generated music aligns with the expected tempo adjustments. The results are summarized in Tab. 9.

Model	Average BPM
Fast	143.55
Medium	122.64
Slow	93.88

Table 9: Average BPM of music generated under varying tempo conditions.

B.4 Additional Ablation Studies

For efficiency, we trained models for this ablation study using only half the standard training epochs. However, we observed that performance trends remained consistent with full-epoch training, ensuring the validity of our conclusions.

We also remove the text modification when exclude target retrieval, and results are shown in Tab. 10. The results are similar since user attributes are randomly sampled during whole experiment.

We further conduct experiments on the scale of retrieving set. As shown in Tab. 11, performance consistently improves with larger retrieval sets. Compared to 1% and 0.1% subsets, using the full retrieval pool achieves the best results across all metrics—lower KL and FD, and higher IB. This confirms that broader retrieval coverage enhances both semantic alignment and audio quality, highlighting the importance of retrieval scale in the video-to-music task. Notably, even under extremely low-resource conditions (e.g., 0.1%), our model still maintains reasonable performance, demonstrating its robustness in limited retrieval scenarios.

Text Modification	KL↓	FD↓	IB↑
×	96.4	360.3	14.7
✓	97.1	364.3	14.5

Table 10: Ablation of text modification.

Dataset	KL_{passt}↓	FD_{openl3}↓	IB↑
1%(247 pieces)	47.43	132.16	21.72
0.1%(25 pieces)	50.28	150.82	19.09
Full	47.12	101.43	22.93

Table 11: Low-resource setting on video-to-music task.

C. ANALYSIS OF GENERATED SAMPLES

C.1 Samples of V2M Generation

In a comparative evaluation with existing models, our MTM framework displayed notable advancements in the stability and thematic alignment of music generation across various video contexts. Unlike CMT, which excels at matching music to minute changes in video content, our model prioritizes broader narrative coherence. This approach ensures a stronger, long-term narrative connection with the video, even if it occasionally sacrifices the precision of minor video-to-music correspondences that CMT might capture more explicitly.

When compared to VidMuse, another leading model in the field, the MTM framework stands out significantly in terms of musical clarity and completeness. The music generated by our model not only demonstrated superior clarity but also showed a better correspondence with the overarching themes of the videos. For example, in an “Anime” video clip where the challenge lay in capturing the complex emotion of sorrow masked by cheerful sounds, *i.e.,* a nuance rooted in the original score, MTM was uniquely successful. Our model adeptly generated a sorrowful major key composition that maintained the cheerful impression while subtly conveying the underlying sadness, mirroring the complex emotional layering of the original music.

Furthermore, in a scene depicting a forest at the start of a film, MTM distinguished itself by producing a piece with a long melodic line that naturally began to fade in, enhancing the scene’s mysterious and vague ambiance. This capability to align musical elements with the intended atmospheric qualities of a scene underscored our model’s advanced understanding of contextual and thematic elements, setting it apart from other models which failed to capture this nuanced musical requirement.

These examples underscore the MTM framework’s capacity to generate music that not only complements but enhances the storytelling of visual media, offering a more immersive and contextually appropriate auditory experience than its contemporaries.

C.2 Samples of Visual-to-Description Generation

To evaluate our model’s performance, we sampled 10 frames from each video and provided the following prompt to the models: `Based on the emotional tone, pacing, and visuals of this video, how would you compose the music? Provide a one-sentence summary of the key musical elements that you use.` This method ensures that the generated music aligns well with the emotional and visual characteristics of the video. Samples of these music descriptions are displayed in Tab. 12. Additionally, the content shown in Figure 1 of the main paper, which illustrates how a poem by P.B. Shelley, “The Flower that Smiles Today,” can also be transformed into a music description, is also generated using our model.

C.3 Conclusion

Compared to other models, our model captures musical details, including style and emotion curves more effectively.

D. DATASET ANALYSIS

We utilize four self-curated datasets for the following purposes: 1) training the ECMG module (MT-512K); 2) constructing the DMR retrieval dataset (MTV-24K); 3) conducting subjective and qualitative evaluations; and 4) assessing the performance of the MMDM module. Additionally, we document modifications made to existing datasets, namely SymMV [13] and SongDescriber [42].

D.1 MT-512K

We collect a total of 512K music tracks. Tracks containing vocal components are identified using PANN [46] and CLAP [34], based on pre-defined tags such as `vocal`, `choir`, and `human voice`. Scores are computed between each track and the target tags, and a threshold of 0.1 is selected to identify tracks with vocal components. This threshold is determined through a manual review process, where we listen to a subset of music tracks to ensure the balance between accurately identifying vocal parts and minimizing false positives. The decision to remove tracks with human voices follows previous works [17], and is guided by three key considerations: 1) to lower the complexity of modeling voice components; 2) to enable the model to focus more effectively on instrumental music; and 3) to align with the primary goal of video background music generation [12, 13], which typically excludes vocal elements. Additionally, we filter out low-quality tracks by applying the PAM [44] score with a threshold of 0.95. After filtering, we obtain 54,112 high-quality tracks. The dis-

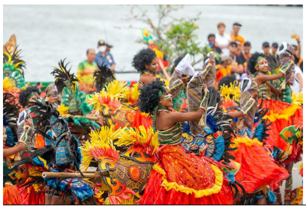
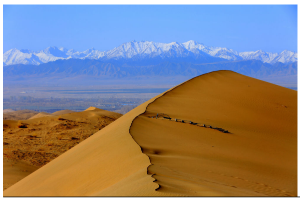
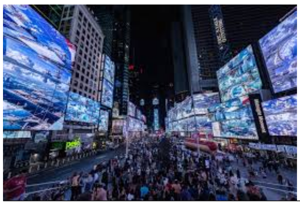
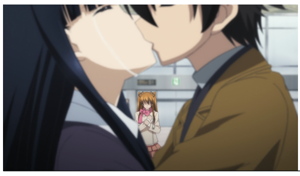
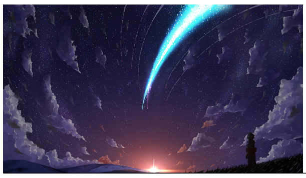
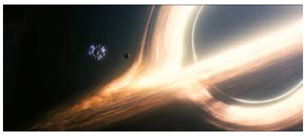
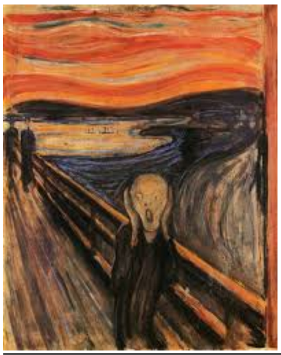
Image	Description	GPT-4 Evaluation
	A vibrant, energetic, and epic soundtrack featuring a dynamic blend of strings, brass, and orchestral elements, perfectly capturing a sense of adventure and excitement.	**Score:** 5 / 5 **Reason:** **Emotion Match:** The energy and vibrancy of the description match perfectly with the lively festival scene. **Scene Association:** The use of "strings, brass, and orchestral elements" effectively aligns with the celebratory and grand setting. **Conclusion:** The description is highly appropriate for this image, requiring no further improvement.
	A dynamic and emotive soundtrack featuring a mix of guitar, bass, strings, and orchestra, perfectly capturing a range of moods from action-packed and adventurous to melancholic and romantic.	**Score:** 4.5 / 5 **Reason:** **Emotion Match:** The description's adventurous tone works well for the desert landscape. However, the "melancholic and romantic" portion slightly detracts from the overall alignment. **Scene Association:** The instrumentation (guitar, bass, and orchestra) complements the sense of vastness and exploration in the desert. **Conclusion:** A more focused description on adventure and solitude could enhance the fit.
	A fast-paced, dynamic electronic soundtrack featuring pulsating rhythms, vibrant synthwave beats.	**Score:** 5 / 5 **Reason:** **Emotion Match:** The fast-paced, electronic, and vibrant tone perfectly reflects the energy, dynamism, and modernity of the urban night scene. **Scene Association:** The use of "synthwave beats" aligns exceptionally well with the futuristic, neon-lit visuals. **Conclusion:** The updated description captures the essence of the image flawlessly, making it an excellent match.
	A gentle piano melody, accompanied by soft strings, to evoke a sense of tenderness.	**Score:** 5 / 5 **Reason:** **Emotion Match:** The gentle piano melody perfectly evoke the tenderness and nostalgia expressed in the characters' emotional moment. **Scene Association:** The use of "soft strings" aligns with the intimate and heartfelt nature of the scene, enhancing the emotional depth. **Conclusion:** The description is highly appropriate for this image, requiring no further improvement.
	A gentle, melancholic melody, featuring soft piano and strings, to evoke the serene yet poignant atmosphere.	**Score:** 4.5 / 5 **Reason:** **Emotion Match:** The melancholic melody matches the serene and poignant atmosphere of the comet-lit sky. **Scene Association:** The inclusion of "soft piano" and "strings" reflects the calmness and wonder of the scene but does not fully emphasize the awe-inspiring grandeur of the comet. **Conclusion:** While the description aligns well with the scene, adding a sense of scale and majesty could enhance the match.
	A sense of awe and wonder, using a combination of orchestral strings, and electronic soundscapes.	**Score:** 5 / 5 **Reason:** **Emotion Match:** The energy and vibrancy of the description match perfectly with the lively festival scene. **Scene Association:** The use of "strings, brass, and orchestral elements" effectively aligns with the celebratory and grand setting. **Conclusion:** The description is highly appropriate for this image, requiring no further improvement.
	A slow, eerie, and melancholic melody, using a combination of dissonant chords and a haunting vocal line to evoke the sense of despair and isolation.	**Score:** 5 / 5 **Reason:** **Emotion Match:** The slow, eerie melody and dissonant chords align seamlessly with the despair and isolation depicted in *The Scream*. **Scene Association:** The "dissonant chords" effectively complements the painting's unsettling and surreal nature. **Conclusion:** The description accurately reflects the psychological intensity of the image, requiring no further improvement.

Table 12: Samples of visual-to-description generation.

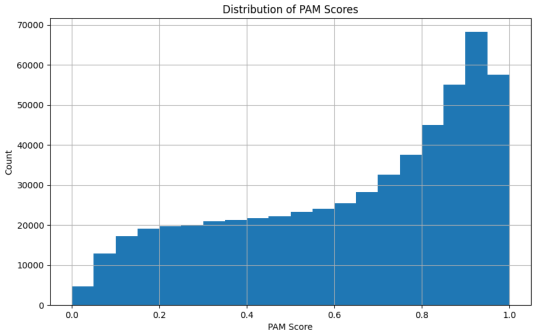


Figure 4: Distribution of PAM Scores across the raw training dataset.

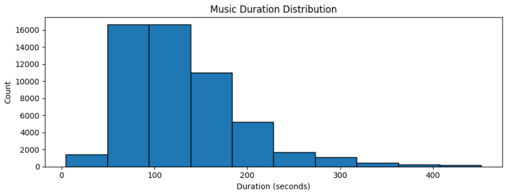


Figure 5: Histogram of music duration in the training dataset.

tribution of PAM score in the whole dataset is illustrated in Fig. 4.

The selected tracks are annotated using the methodology described in MMDM. Candidate tags are pre-defined by musicians who serve as domain experts to align with commonly used labels in music generation while minimizing the effect of long-tail distribution. The detailed annotations are provided in the file `label.json`.

We use **instrument** tagging as an example to illustrate the experts' taxonomy. Instruments are first categorized into five primary groups: strings, keyboards, wind instruments, percussion instruments, and others. To address the issue of long-tail distributions, rare tags are combined into broader categories. For instance, instruments like the double bass, which are seldom played independently, are treated as indivisible units. Therefore, we use strings to replace double bass. This approach ensures robust categorization while accurately reflecting the practical usage of instruments in compositions. Other tagging principles, which align with the general design philosophy of instrument categorization, are omitted here for brevity.

To generate natural descriptions, we employ the Llama-3.1-8B-Instruct [47] model to generate one-sentence descriptions for each music track. The system prompt is: *You are a professional music expert. Here are tags about this music. You need to generate a one-sentence description for this music. Only generate the description without any other information.* Metadata tags were supplied via the user prompt.

Fig. 5 and Fig. 6 illustrate the distribution of music durations and the corresponding descriptions, providing insight into the characteristics of the dataset.

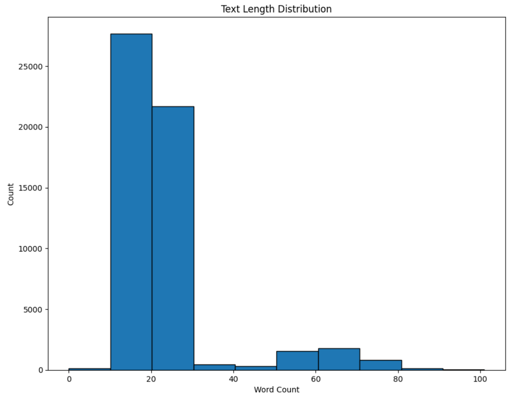


Figure 6: Histogram of text word counts in the training dataset.

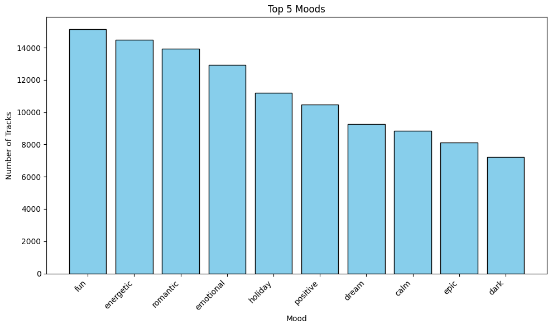


Figure 7: Distribution of mood tags across the retrieval dataset. This histogram shows the frequency of various mood categories, illustrating the emotional diversity captured in our data.

D.2 MTV-24K

We have introduced the methodology of collecting the MMDM training and retrieval dataset. Here, we provide other details.

First, we scrape the hottest singers and their songs from Spotify. We use YoutubeAPI[1] to search for their original official music video with keywords in the format of `{singer} + {song name} + Official Music Video`. We scrape all the music videos with their Youtube description boxes and top-10 comments. We also use Shazam[2] to get other metadata. Finally we obtain the primary dataset.

We use methodology in MMDM to label each music track. Eventually our dataset contains 24,719 video-text-music pairs.

In targeted retrieval, we partition our whole datasets. We follow three standards and get three different partitions, *i.e.*, genre, tempo, and mood partition. We use the labeled tags to partition the whole dataset. Distribution of each attribute is shown in Fig. 7, 8, 9, and 10.

Fig. 11 shows the distribution of audio durations in the

[1] https://developers.google.com/youtube/v3
[2] https://www.shazam.com/

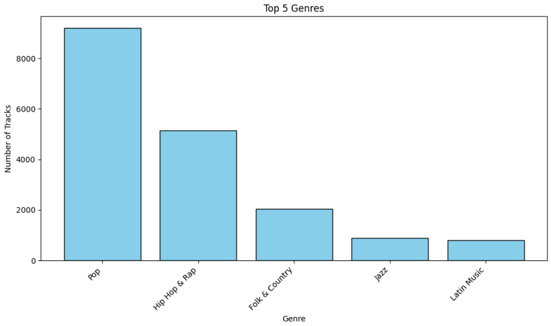


Figure 8: Genre distribution within the retrieval dataset. This bar graph reflects the variety of music genres represented, indicating the dataset's broad applicability for genre-specific retrieval tasks.

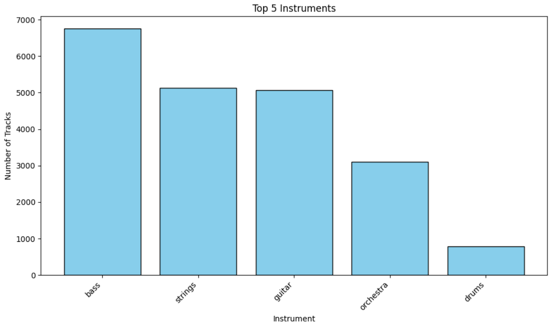


Figure 9: Histogram of instrument tags in our retrieval dataset. This figure shows the range of musical instruments represented, underscoring the dataset's comprehensive coverage of instrumental music.

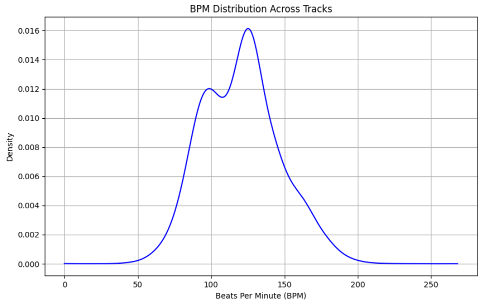


Figure 10: Density curve of Beats Per Minute (BPM) across the retrieval dataset. This plot illustrates the distribution of Beats Per Minute, showcasing the tempo range covered in our collection.

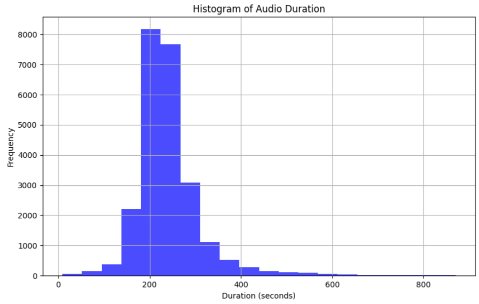


Figure 11: Histogram of audio durations in retrieval dataset. This shows the distribution of song lengths in the dataset.

dataset. The majority of the audio files have durations concentrated within a specific range, suggesting consistent lengths across the dataset. The associated text data is characterized by word count distribution in Fig. 12, showing a variety of text lengths, with most falling in the mid-range. Finally, the lexical diversity, representing vocabulary usage variation, is displayed in Fig. 13. Most texts demonstrate high lexical diversity, indicating rich vocabulary usage across the dataset.

D.3 Subjective Evaluation Dataset

To evaluate our model across diverse video contexts, we construct a dataset encompassing seven distinct video categories: Scene/Vlog, Documentary, Advertisement, Movie, Game, Anime, and Sports. These categories are selected to ensure a comprehensive benchmark that reflects a wide range of video types and their corresponding music needs. Each video is manually reviewed to ensure quality, resulting in a final dataset of 35 high-quality videos.

We apply the following criteria for video selection:

- Duration: Videos are limited to a maximum length of three minutes to ensure efficient evaluation.

- Content Quality: Only classic or high-quality videos are included, reflecting diverse and impactful visual themes.

Due to limitations on providing external URLs in the supplementary material, we describe each category as follows:

Scene/Vlog. This category focuses on travel and scenic content, featuring dynamic landscapes, urban exploration, and lifestyle moments. Examples include aerial footage of Amoy, a travel vlog in Tokyo, and a beachside promotional video capturing the beauty of coastal environments.

Documentary. Highlighting impactful narratives, this category includes trailers for iconic documentaries such as *Planet Earth II*, which captures breathtaking natural phenomena; *The Last Dance*, chronicling Michael Jordan's legendary career; and *Free Solo*, an intense portrayal of extreme rock climbing.

Advertisement. Promotional videos from globally recognized brands are featured here, such as Honda's engineering marvel in its Classic Honda Commercial, Coca-Cola's emotionally resonant advertisements, and Samsung's futuristic product showcases.

Movie. Iconic film moments and trailers dominate this category, with examples including the montage from *The Shawshank Redemption*, the epic trailer for *Interstellar*, and the thrilling scenes from *Jurassic Park* and *Star Wars*.

Game. This category showcases the artistry of game trailers, featuring visually stunning examples like the official

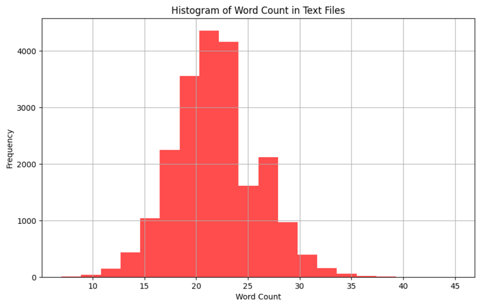


Figure 12: Histogram of text word counts in retrieval dataset. This represents the distribution of word counts in the associated text data.

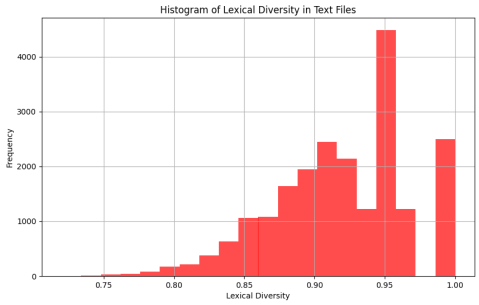
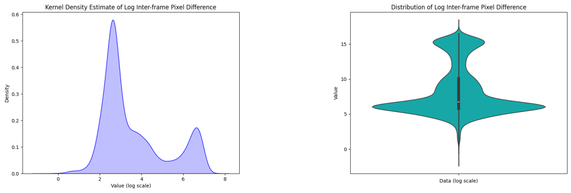


Figure 13: Histogram of lexical diversity scores in retrieval dataset. This shows the variation in vocabulary usage across text samples.



(a) KDE of log inter-frame difference

(b) Violin plot of log inter-frame distribution

Figure 14: Distribution of log-transformed inter-frame motion across all videos in Disco-200K-high-quality.

trailer for *Genshin Impact*, the adventurous world of *Zelda: Breath of the Wild*, and the nostalgic appeal of *Final Fantasy VII*.

Anime. This category includes promotional materials and memorable moments from classic anime series. Examples include the heartwarming trailer for *My Neighbor Totoro*, the action-packed sequences from *Attack on Titan*, the emotional depth of *Evangelion*, and the visually stunning storytelling of *Your Name*.

Sport. Dramatic and inspiring moments from the world of sports are showcased in this category. Examples include the excitement of the Qatar World Cup trailer, Usain Bolt's electrifying highlights, the high-speed intensity of Formula 1 racing, and the Paris 2024 Olympics promotional film.

D.4 Visual-to-Description Evaluation Dataset

We downloaded all available videos from the Disco-200K-high-quality dataset [29] as of July 30, 2023. Upon review, we noted that many videos consisted of static images or were lyrics-based music videos. To filter for dynamic content, we calculated the average pixel difference between consecutive frames to identify static videos. Videos close to the threshold are manually reviewed to confirm their dynamic nature. Following this rigorous selection process, we curated a subset of 8,042 videos that demonstrated sufficient motion suitable for our analysis.

To visualize the motion characteristics of the Disco-200K-high-quality dataset, we computed the average pixel difference between consecutive frames for each video and applied a logarithmic transformation to reduce the influence of extreme outliers (e.g., flashes, scene cuts). As shown in Fig. 14, the distribution is highly right-skewed: most videos exhibit minimal motion, while a minority contain substantial dynamic content. We used this distribution to guide our filtering strategy. Videos near the boundary were manually inspected to ensure quality. This filtering ensures that the retained 8,042 videos are visually dynamic and better suited for downstream multimodal alignment tasks.

D.5 SymMV, SongDescriber, and MUImage

We utilize all available videos from the SymMV [13] dataset as of November 1, 2024.

While Stable Audio Open [3] claimed to have filtered the SongDescriber dataset to exclude tracks with vocal components, we conduct an additional manual review to ensure thorough filtering. After this process, we retain a total of 279 text-music pairs.

For the MUImage dataset [4], as the authors do not specify a valid test set but only provided the training set, we sort the entire dataset alphabetically by file name and use the first 1,500 image-music pairs as the test set. Consequently, the evaluation of MuMu-LLaMA on MUImage is actually conducted on its training set.

The specific video-music / text-music pairs used in our study are documented in the files `Dataset/SymMV.csv` and `Dataset/SongDescriber.csv`.

E. METHOD DETAILS

E.1 Chain-of-Thought Training in MTV-24K

Pretrained multimodal large language models (MLLMs) exhibit strong generalization. In our setting, we use music videos (MVs) as training data because they offer high-quality music–visual alignment and professionally produced audio. However, the genre coverage of these videos is limited, which makes direct fine-tuning prone to overfitting.

To address this, we propose a chain-of-thought (CoT) training scheme that explicitly links visual evidence to musical attributes. The goal is to preserve the relevance between the generated music description and the video content, thereby improving generalization—especially for the music-description generation task.

Given an input video V, we first use InternVL-2 [7] to produce a video caption C. We then fine-tune InternVL-2 with the following instruction-style prompt, conditioned on V and C: *[V] You are a film music supervisor. The following describes a film clip: [C] Based on the emotional tone and visuals of this clip, how would you compose the music? Provide a one-sentence summary of the key musical elements you would use.* This prompt encourages the model to articulate a concise rationale that ties visual cues (e.g., mood, pacing, scene dynamics) to musical decisions (e.g., tempo, instrumentation, harmony), reinforcing music–visual relevance during training.

E.2 Details of Broad Search

For each conditional modality, we first retrieve music that closely aligns with the input by computing similarities using CLIP [33] or CLAP [34] embeddings. Given a target music, we directly use CLAP to compute audio embeddings $\mathbf{a}_{\text{input}}$ and compare them with embeddings $\{\mathbf{e}_i^{\text{audio}}\}$ in our music database:

$$i^* = \arg\max_i \cos\left(\mathbf{e}_{\text{input}}^{\text{audio}}, \mathbf{e}_i^{\text{audio}}\right). \quad (1)$$

The music pieces retrieved from the broad retrieval $\mathbf{m}_{i^*}$ serve as fine-grained conditions. Notably, although we only apply music-centric retrieval during training, it is feasible to use CLIP embeddings of the input text $\mathbf{t}_{\text{input}}$ or visual signals $\mathbf{v}_{\text{input}}$ and compare them with text-music or visual-music pairs in the database as inference-time retrieval. The music associated with the closest matches is retrieved as:

$$i^* = \arg\max_i \cos\left(\mathbf{e}_{\text{input}}^{\text{text/visual}}, \mathbf{e}_i^{\text{text/visual}}\right), \quad (2)$$

where $\mathbf{e}_{\text{input}}^{\text{text/visual}}$ and $\mathbf{e}_i^{\text{text/visual}}$ are the CLIP embeddings of the input text or visual content and each dataset entry. This ensures that the retrieved music complements the semantic and emotional context of the visual or textual input, allowing the model to effectively capture complex musical features, such as melody and rhythm.

E.3 Details of Target Search

In parallel to the broad retrieval, we conduct a targeted retrieval within partitions of our dataset, specifically organized based on musical attributes. For instance, the tempo partition is categorized into “fast”, “medium”, and “slow” subsets. This allows users to flexibly select songs that match specific attributes among genre, tempo, or mood partition. To retrieve a fast-paced song, we query directly within the “fast” subset in the tempo partition. In each subset, the most suitable music piece is determined by computing the cosine similarity between the CLAP embeddings of the desired textual attribute, encoded as $\mathbf{e}_{\text{desired}}^{\text{attr}}$ and $\{\mathbf{e}_k^{\text{audio}}\}$:

$$k^* = \arg\max_k \cos\left(\mathbf{e}_{\text{desired}}^{\text{attr}}, \mathbf{e}_k^{\text{audio}}\right). \quad (3)$$

The embedding of the retrieved music $\mathbf{e}_{k^*}^{\text{audio}}$ is used as input in a subsequent module, where it is integrated into the generation process.

F. EXPERIMENT DETAILS

In this section, we provide details of the experiments in the main paper.

F.1 Implementation Details

For MMDM module, We utilize the InternVL-2 [7] architecture with mixed precision training (bfloat16) and the AdamW optimizer (learning rate: 1e-6, cosine decay). Top-k sampling (k=50) and nucleus sampling (p=1.0) are used for generation, with a temperature of 1.0. The model features FlashAttention v2 [48], 32 layers, and 16 attention heads, paired with a vision backbone (Intern-ViT-6B) [49] at 448x448 resolution. Training is performed using a batch size of 2 with gradient accumulation over 8 steps. Fine-tuning is conducted with LoRA [50] at rank 16, utilizing 8 NVIDIA A800-SXM4-80GB GPUs over 3 days for a total of 10 epochs.

In ECMG module, We leverage a DiT [35] model with T5 [51] conditioning and an audio autoencoder for fine-tuning on audio-text alignment tasks. The architecture includes a 24-layer continuous transformer with 1536 embedding dimensions and 24 attention heads. Audio inputs are preprocessed via an autoencoder featuring Oobleck [52] encoders/decoders with multi-scale channel configurations and a latent dimension of 64.

Training uses AdamW [53] (learning rate: 1e-6, weight decay: 0.001) with an InverseLR scheduler. The model operates at a 44.1kHz sample rate with a batch size of 1 and gradient accumulation over 8 steps. Mixed precision (16-bit) training is employed, using DeepSpeed [54] for optimization with 4 NVIDIA A100-PCIE-40GB GPUs.

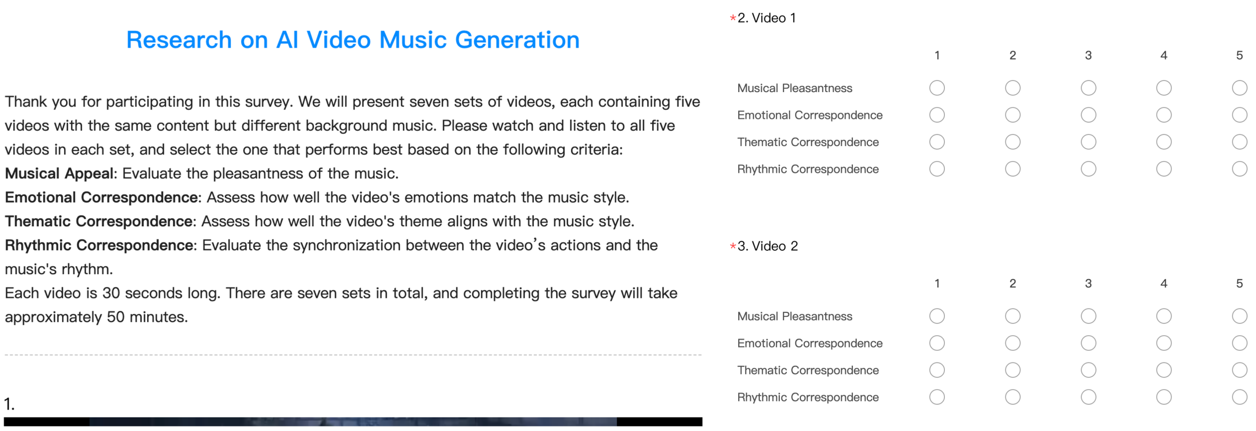
Conditioning involves audio features, time attributes, and textual prompts (via T5-base, max length: 128).

F.2 Video-to-Music Generation

We evaluate video-to-music generation using several state-of-the-art models, adhering to their default settings for fair comparison. Due to the sequence length limitation of MuMu-LLaMA [4], all samples are truncated to 30 seconds. Additionally, to accommodate the resolution constraint of CMT [12], all videos are resized to 360p.

For all MIDI based model, we use SGM-v2.01-Sal-Guit-Bass-V1.3 as the sound font. We also follow the default setting in each generation model. For MTM, we set a standard sample rate of 44,100 Hz and adjust the CFG scale to 7 for balancing conditioning influence, with 100 steps for detailed noise reduction. Our sampler, “DPM++ 3M SDE” [55], manages noise sampling across a sigma range from 0.03 to 1000. For Diff-BGM [15], we notice that, contrary to their claim in the paper [15], the visual encoder is actually sourced from CLIP2Video [3] rather than

[3] https://github.com/CryhanFang/CLIP2Video



Research on AI Video Music Generation

Thank you for participating in this survey. We will present seven sets of videos, each containing five videos with the same content but different background music. Please watch and listen to all five videos in each set, and select the one that performs best based on the following criteria:
Musical Appeal: Evaluate the pleasantness of the music.
Emotional Correspondence: Assess how well the video's emotions match the music style.
Thematic Correspondence: Assess how well the video's theme aligns with the music style.
Rhythmic Correspondence: Evaluate the synchronization between the video's actions and the music's rhythm.
Each video is 30 seconds long. There are seven sets in total, and completing the survey will take approximately 50 minutes.

1.

*2. Video 1

	1	2	3	4	5
Musical Pleasantness	○	○	○	○	○
Emotional Correspondence	○	○	○	○	○
Thematic Correspondence	○	○	○	○	○
Rhythmic Correspondence	○	○	○	○	○

*3. Video 2

	1	2	3	4	5
Musical Pleasantness	○	○	○	○	○
Emotional Correspondence	○	○	○	○	○
Thematic Correspondence	○	○	○	○	○
Rhythmic Correspondence	○	○	○	○	○

Figure 15: Screenshot of the user study questionnaire in subjective evaluation.

VideoCLIP. This is evident from their implementation [4] and the fact that the embedding dimensions do not match those of VideoCLIP but align with those in CLIP2Video. Additionally, the generated music quality remains strong. Based on this, we preprocess using CLIP2Video [56].

In our experiments, we use a server with dual Intel Xeon Processors (Icelake) with 64 cores each and four NVIDIA A100-PCIE-40GB GPUs, under a KVM virtualized Linux environment. These configurations ensure the reproducibility of our inference time measurements.

F.3 Text-to-Music Generation

For text-to-music models, we also adhere to their default settings for a fair comparison. Due to the maximum sequence length limitation of MuMu-LLaMA [4], all generated music samples are capped at 30 seconds.

MTM follows the parameters outlined in [3]. It uses a 44,100 Hz sample rate, a CFG scale of 7, 100 steps for diffusion, and a 'DPM++ 3M SDE' sampler with a sigma range from 0.03 to 1000.

F.4 Subjective Evaluation

For the subjective evaluation, we provide each participant with two sets of evaluations: one for video-to-music generation and another for text-to-music generation. Each set consists of seven groups, corresponding to the seven categories in our subjective evaluation dataset. For each category, a video is randomly sampled. Each group includes five music pieces, one generated by each of the compared models.

For video-to-music (V2M) generation, participants are briefed on the purpose of the study and instructed to evaluate how well the generated music matched the video in terms of mood, thematic alignment, and overall enhancement of the viewing experience. Each video is presented seven times, each time paired with a background music track generated from a different model. The order of music presentations is randomized to eliminate order effects. After watching all versions of a video, participants are asked to rate each version using a Likert scale, *i.e.* rate from 1 to 5, considering factors such as emotional impact, thematic coherence, and suitability for the video's content. Additionally, participants are encouraged to provide qualitative feedback explaining their preferences, highlighting specific emotional or thematic factors that influenced their choices.

For the V2M task, we utilize the following metrics to guide participant evaluations:

- **Musical Pleasantness (MP)**: The aesthetic quality of the music, independent of context.
- **Emotional Correspondence (EC)**: How well the music conveys the intended emotions of the video.
- **Thematic Correspondence (TC)**: The alignment between the video's theme and the generated music.
- **Rhythmic Correspondence (RC)**: The synchronization between the video's motion and the music's rhythm.

The evaluation process for text-to-music (T2M) generation follows a similar structure. Participants are provided with textual prompts and asked to evaluate the generated music based on how well it reflects the mood, style, and thematic elements described in the text. Each textual prompt is paired with five music tracks generated from five models, and participants rate the tracks using the same Likert scoring system. Feedback is collected to identify specific strengths and weaknesses of the generated music in conveying textual meaning.

For the T2M task, we select the following metrics:

- **Musical Pleasantness (MP)**: The aesthetic quality of the music, independent of context.
- **Text-Music Alignment (TMA)**: How effectively the music captures the mood, style, and themes described in the text.

Each participant rated 40 tracks for the V2M task and 35 tracks for the T2M task, completing the questionnaire in

[4] https://github.com/sizhelee/Diff-BGM

approximately 50 minutes. In addition to quantitative ratings, participants also provided qualitative feedback, explaining their preferences and highlighting the emotional and thematic factors influencing their decisions. Each participant received a $10 reward, and a total of 80 valid responses were collected via social media recruitment, which is the largest scale compare to 20-55 participants in existing works [4, 12–15, 17].

A notable result is observed in the RC metric. Despite not explicitly introducing rhythm-focused features, our model achieves superior performance compared to baseline methods. This suggests that an alternative approach, which emphasizes a broader perspective rather than heavily focusing on local rhythmic features, can also be effective. Overemphasis on local rhythms may sometimes challenge the overall coherence of the music and its alignment with the broader narrative. By adopting a balanced approach, our model maintains rhythmic flow while aligning closely with the intended emotional context, leading to an enhanced audiovisual experience. We also provide a screenshot of the questionnaire with full text of instructions in Fig. 15.

We also conduct subjective evaluation with 42 participants to evaluate perceptual controllability. For each attribute, listeners are presented with a pair of outputs (with and without the target control) and asked which better matched the condition. They also rated alignment on a 1–5 Likert scale. Results in Tab. 8 show high agreement and strong subjective ratings, confirming the effectiveness of MTM's control mechanisms.

G. BROADER IMPACTS, LIMITATIONS, AND FUTURE WORKS

Broader Impacts. The MTM framework introduces several significant impacts across various domains. Primarily, MTM enhances accessibility in entertainment technologies such as gaming and virtual reality by enabling the automatic generation of emotionally resonant background music, which could improve user engagement and accessibility for people with diverse abilities. However, the technology also presents potential negative impacts. The automation of music generation might reduce the need for human composers in certain contexts, potentially affecting their livelihoods. Moreover, biases in the training data could lead to outputs that do not adequately represent the diversity of global musical expressions, potentially introducing unfairness in musical representation.

Limitations. The MTM framework represents a significant advancement in multimodal music generation; however, it is not devoid of limitations. The model's effectiveness heavily relies on the diversity and quality of the dataset it is trained on. Presently, available datasets may not adequately represent the vast array of musical styles and cultural expressions, thus limiting the system's ability to produce a broad spectrum of musical outputs. Moreover, the challenge of accurately translating complex emotional and thematic nuances across different modalities persists, as the system occasionally fails to capture the depth and subtlety inherent in human compositions.

Future Works. Future enhancements to the MTM framework should focus on several key areas. Broadening the diversity of the dataset to encompass a greater variety of musical styles and cultural expressions would considerably enhance the model's generative capabilities. Improving the system's understanding of and ability to accurately translate intricate emotional and thematic nuances between modalities would make the technology more effective. Integrating music theory into the music generation process could provide a more robust framework for generating musically coherent outputs.