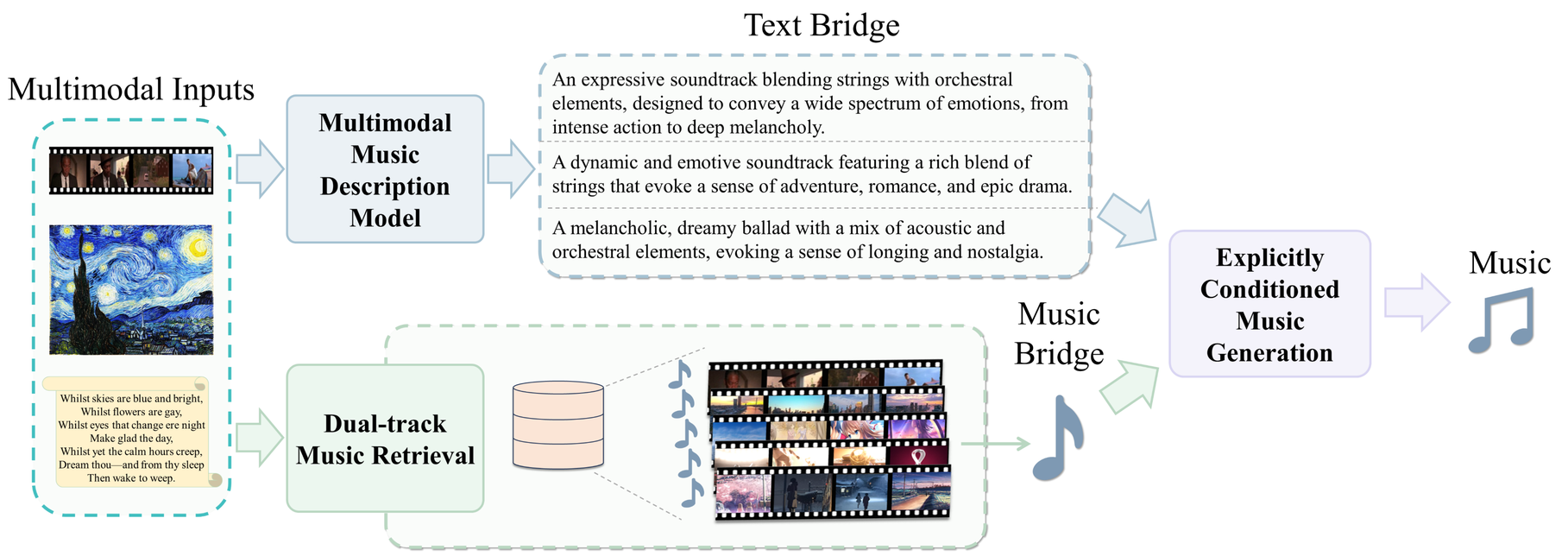
Multimodal Music Generation with Explicit Bridges and Retrieval Augmentation

Baisen Wang[1,2], Le Zhuo[3], Zhaokai Wang[4], Chenxi Bao[5], Chengjing Wu[6]
Xuecheng Nie[6], Luoqi Liu[6], Jiao Dai[1,2], Jizhong Han[1,2], Yue Liao[7], Si Liu[8]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyberspace Security, University of Chinese Academy of Sciences
[3]The Chinese University of Hong Kong [4]Shanghai Jiao Tong University [5]Music Tech Lab, DynamiX
[6]MT Lab, Meitu Inc. [7]National University of Singapore [8]Beihang University

ISMIR 2025



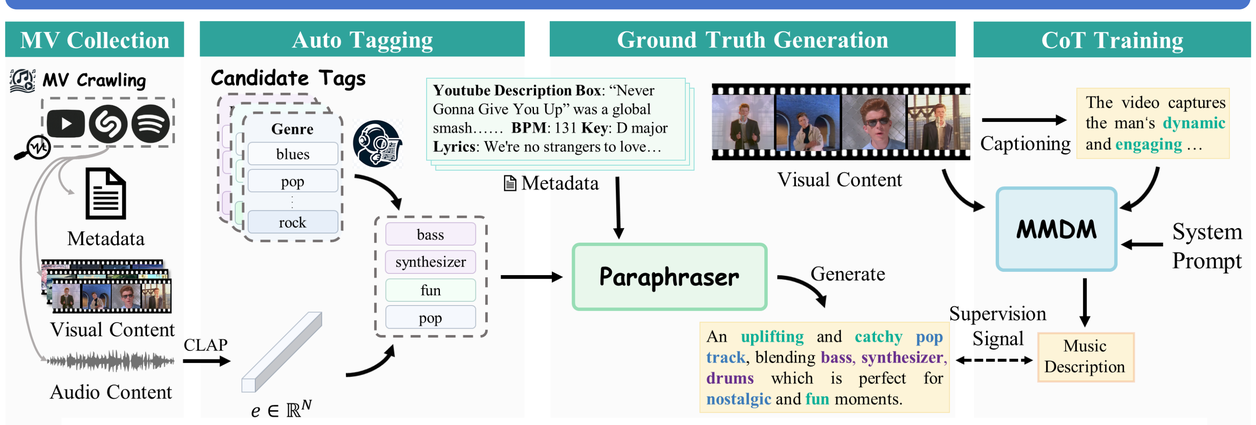
arXiv

Github

Introduction

- 🎶 Background: Multimodal music generation — creating music from text, images, or video, with applications in film, games, and XR.
- ⚠️ Problem: Existing methods suffer from limited data, weak cross-modal alignment, and lack of controllability.
- 💡 Motivation: We propose explicit cross-modal bridges (text bridge + music bridge) to improve alignment and enhance user control.

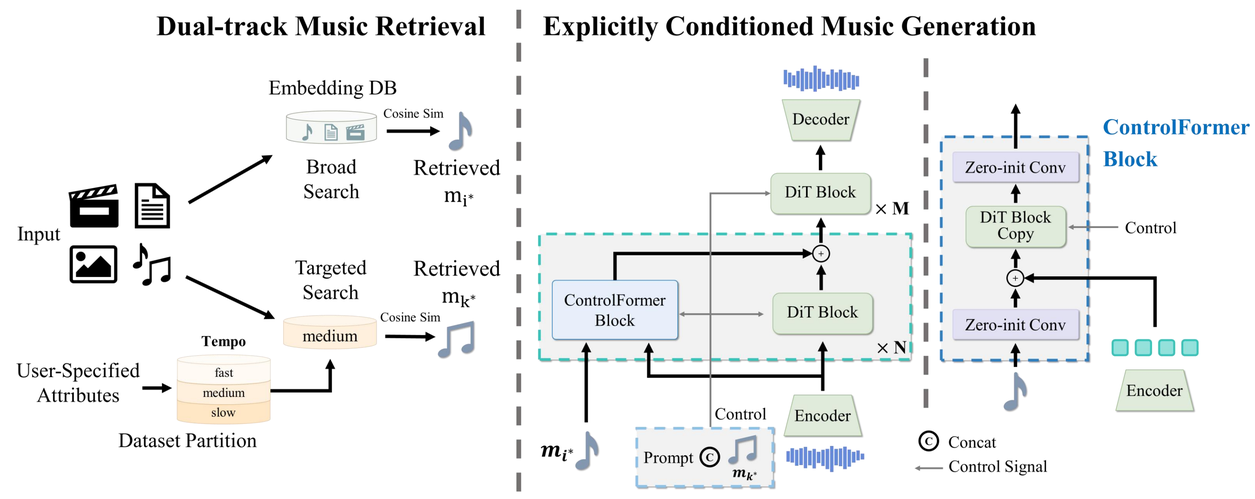
Dataset



Dataset	Genre	Music Description	Source Separation	Music Attr.	Video Content	Size	Length (Hours)
HIMV-200K [21]	✗	✗	✗	✗	Music Video	200K	-
AIST++ [33]	✓	✗	✗	✗	Dance Video	1,408	5.2
TikTok [62]	✗	✗	✗	✗	Dance Video	445	1.5
SymMV [63]	✓	✗	✗	✓	Music Video	1,140	76.5
DISCO-MV [31]	✗	✗	✗	✗	Music Video	2200K	47K
V2M [53]	✗	✗	✗	✗	General Video	360K	17.5K
MUVideo [37]	✗	coarse-grained	✗	✗	General Video	14.5K	40.3
BGM909 [34]	✓	fine-grained	✓	✓	Music Video	909	70.0
MTV-24K(Ours)	✓	fine-grained	✓	✓	Music Video	24K	1.7K

- MTV-24K: A curated video–music dataset with fine-grained alignment, used for training visual-to-music description.
- MT-512K: A large-scale text–music dataset (500K+ pairs) with rich annotations, forming the foundation for retrieval-augmented generation.

Method



- Text Bridge (MMDM): Translates visual inputs (image/video) into structured music descriptions, serving as the semantic bridge.
- Music Bridge (Dual-track Retrieval): Broad retrieval provides melody/rhythm reference; targeted retrieval offers controllable attributes like genre, mood, tempo.
- Explicitly Conditioned Generation (ECMG): Diffusion Transformer + ControlFormer, integrating both bridges for high-quality and controllable music generation.

Results

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**

Video2Music in 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**

Text2Music in SongDescriber

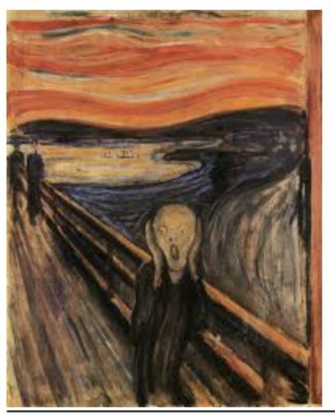
Method	KL_{passt}↓	FD_{openl3}↓	IB↑
CoDi [6]	216.48	251.52	9.60
MuMu-LLaMA [4]	128.33	247.42	2.28
MTM (ours)	**98.78**	**116.71**	**12.10**

Image2Music in MUImage

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

Music2Description

Qualitative Results

	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 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 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.