Multimodal Music Generation with Explicit Bridges and Retrieval Augmentation













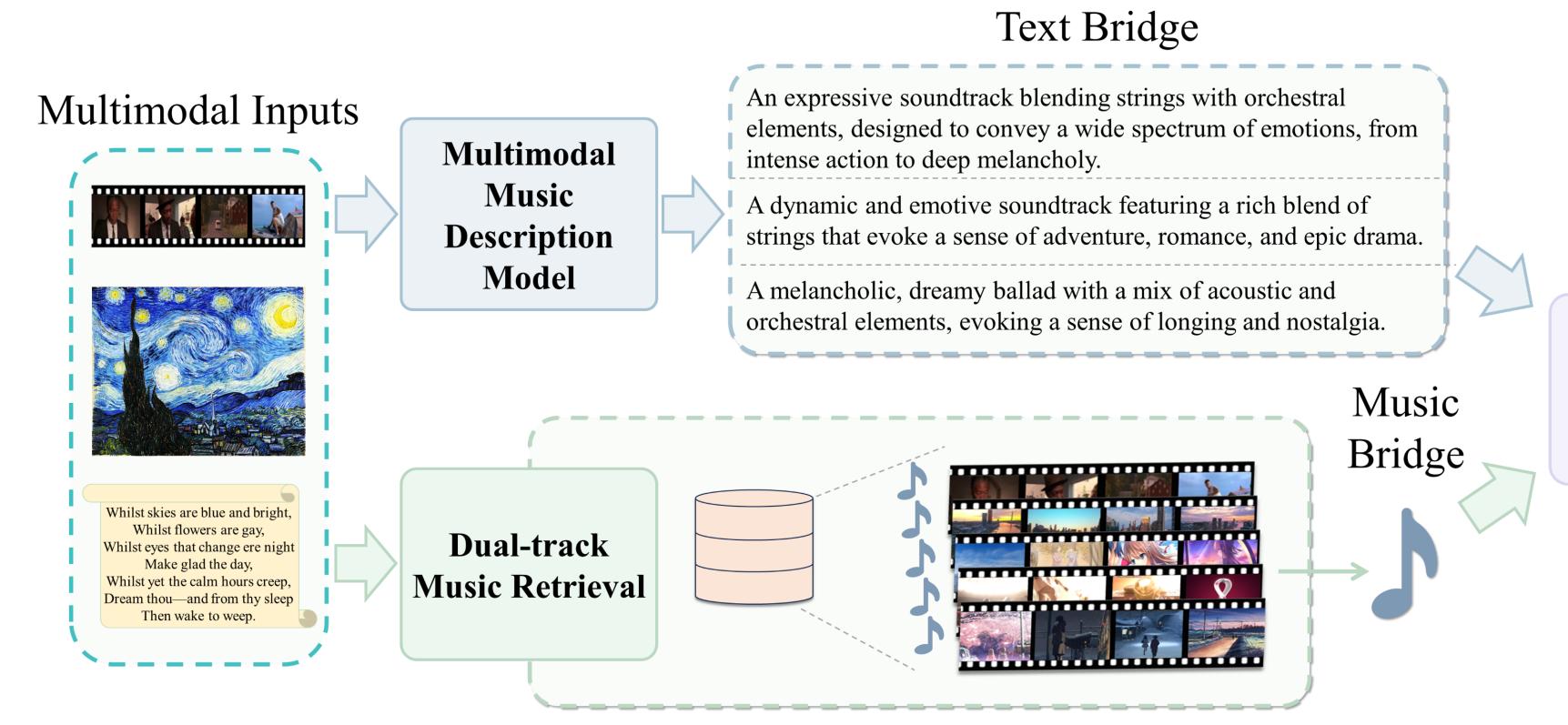
Baisen Wang^{1,2}, Le Zhuo³, Zhaokai Wang⁴, Chenxi Bao⁵, Chengjing Wu⁶ Xuecheng Nie⁶, Luoqi Liu⁶, Jiao Dai^{1,2}, Jizhong Han^{1,2}, Yue Liao⁷, Si Liu⁸

¹Institute of Information Engineering, Chinese Academy of Sciences ²School of Cyberspace Security, University of Chinese Academy of Sciences

³The Chinese University of Hong Kong ⁴Shanghai Jiao Tong University ⁵Music Tech Lab, DynamiX ⁶MT Lab, Meitu Inc. ⁷National University of Singapore ⁸Beihang University







Explicitly Music **Conditioned** Music Generation

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.

			Datase	et			
IV Collection	Au	to Tagging	Grou	nd Truth Ge	eneration		CoT Training
MV Crawling Metadata	Genre blues pop	Gonna Giv	Description Box: "Never we You Up" was a global . BPM: 131 Key: D major e're no strangers to love		sual Content	Captio	The video capto the man's dynamic and engaging Syn
		synthesizer fun pop	Paraph	raser	Generate	Supervisio	
Visual Content Audio Content	$e \in \mathbb{R}^N$			track, blending drums which	g and catchy poping bass, synthesizer, the is perfect for fun moments.	, ←	Music Description
CL.	→	Music Description	Source Separation	track, blending drums which	ng bass, synthesizer, ch is perfect for	, ←	
Audio Content Dataset	$e \in \mathbb{R}^N$ Genre		Source Separation	drums which nostalgic and	ng bass, synthesizer, ch is perfect for fun moments. Video Content	Size	Description
Audio Content Dataset HIMV-200K [2]	$e \in \mathbb{R}^N$ Genre		Source Separation	drums which nostalgic and Music Attr.	ng bass, synthesizer, ch is perfect for fun moments.	Size 200K	Description Length (Hours)
Audio Content Dataset HIMV-200K [2 AIST++ [33]	$e \in \mathbb{R}^N$ Genre		Source Separation X X X	drums which nostalgic and Music Attr.	ng bass, synthesizer, ch is perfect for fun moments. Video Content Music Video Dance Video	Size	Description
Audio Content Dataset HIMV-200K [2 AIST++ [33] TikTok [62]	$e \in \mathbb{R}^N$ Genre		Source Separation X X X X	drums which nostalgic and Music Attr.	ng bass, synthesizer, ch is perfect for fun moments. Video Content Music Video	Size 200K 1,408	Length (Hours) - 5.2
Audio Content Dataset HIMV-200K [2 AIST++ [33]	$e \in \mathbb{R}^{N}$ Genre 21] \checkmark \checkmark		Source Separation X X X X	drums which nostalgic and Music Attr.	ng bass, synthesizer, ch is perfect for fun moments. Video Content Music Video Dance Video Dance Video	Size 200K 1,408 445	Length (Hours) - 5.2 1.5
Audio Content Dataset HIMV-200K [2 AIST++ [33] TikTok [62] SymMV [63]	$e \in \mathbb{R}^{N}$ Genre 21] \checkmark \checkmark		Source Separation X X X X X	drums which nostalgic and Music Attr.	ng bass, synthesizer, ch is perfect for fun moments. Video Content Music Video Dance Video Dance Video Music Video	Size 200K 1,408 445 1,140	Length (Hours) - 5.2 1.5 76.5
Audio Content Dataset HIMV-200K [2 AIST++ [33] TikTok [62] SymMV [63] DISCO-MV [3	$e \in \mathbb{R}^{N}$ Genre X	Music Description X X X X X	Source Separation X X X X X X	drums which nostalgic and Music Attr.	ng bass, synthesizer, ch is perfect for fun moments. Video Content Music Video Dance Video Dance Video Music Video Music Video Music Video	Size 200K 1,408 445 1,140 2200K	Description Length (Hours) - 5.2 1.5 76.5 47K
Audio Content Dataset HIMV-200K [2 AIST++ [33] TikTok [62] SymMV [63] DISCO-MV [3 V2M [53]	$e \in \mathbb{R}^{N}$ Genre X		Source Separation X X X X X X X X	drums which nostalgic and Music Attr.	ng bass, synthesizer, ch is perfect for fun moments. Video Content Music Video Dance Video Dance Video Music Video Music Video General Video	Size 200K 1,408 445 1,140 2200K 360K	Description Length (Hours) - 5.2 1.5 76.5 47K 17.5K

- 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						
Dual-track Music Retrieval	Explicitly Conditioned Music Generation					
Embedding DB Cosine Sim Broad Search Targeted Search Tempo User-Specified Attributes Dataset Partition	Decoder DiT Block Control Former Block Control Former Block Control Former Block Copy Control Control Frompt \bigcirc m_{i^*} Prompt \bigcirc m_{k^*} Prompt \bigcirc m_{k^*} Control Signal					

- Text Bridge (MMDM): Translates visual inputs (image/video) into structured music descriptions, serving as the semantic bridge.
- Music Bridge (Dual-track Retrieval): Broad retrieval 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†			
1,100100		$\overline{\mathrm{KL}_{passt}} \downarrow$	$\text{FD}_{openl3} \downarrow$	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↑	
	$\overline{\mathrm{KL}_{passt}} \downarrow$	$\text{FD}_{openl3} \downarrow$	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	$\mathrm{KL}_{passt} \!\!\downarrow$	$\text{FD}_{openl3} \downarrow$	IB↑
CoDi [6]	216.48	251.52	9.60
MuMu-LLaMA [4] MTM (ours)	128.33 98.78	247.42 116.71	2.28 12.10

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

Image2Music in MUImage

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.

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

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.

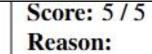
Conclusion: The description is highly appropriate for this image, requiring no further improvement. Score: 4.5/5



A gentle, melancholic melody, featuring soft piano and strings, to evoke the serene yet poignant atmosphere.

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.



Score: 5 / 5

Score: 5/5

Reason:

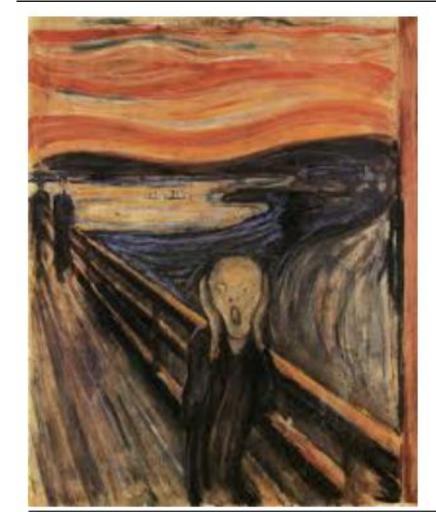
Reason:

Reason:

A slow, eerie, and melancholic melody,

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.



using a combination of dissonant chords and a haunting vocal line to evoke the sense of despair and isolation.