From Frames to Beats: Progress and Challenges in Video-to-Music Generation

Zhaokai Wang

Shanghai Jiao Tong University

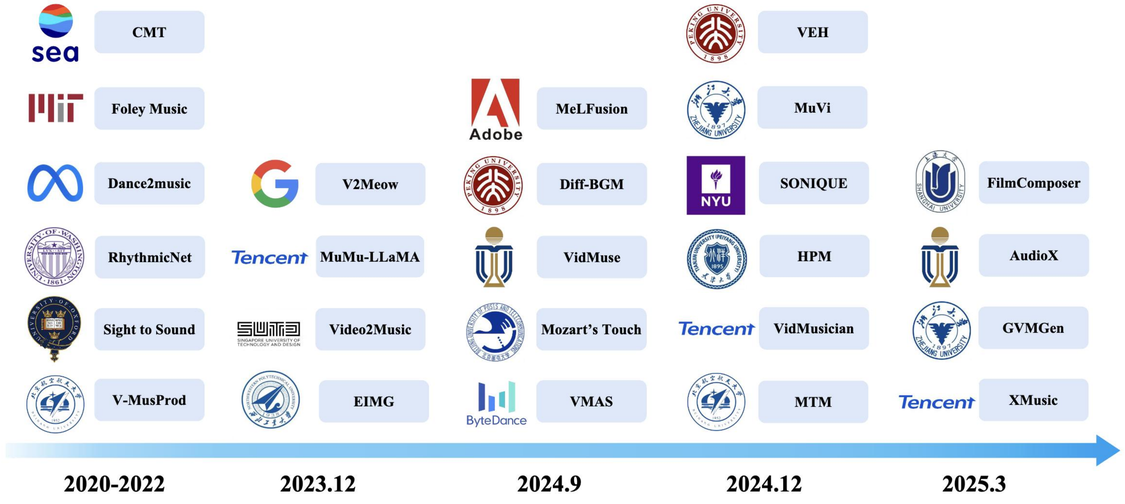
wangzhaokai@sjtu.edu.cn

Introduction

- Video-to-Music Generation

Create music that is semantically, rhythmically, and emotionally aligned with a given video

- Applications: film scoring, games, short video creation, dance music synthesis, VR, ...
- Current status: underdeveloped due to complex vision-music relationships; few industrial deployment



Timeline of Representative Works

Overview of Video-to-Music

Initial Attempt: CMT (2021)

Advanced Method: MusProd (2023)

Recent Work: VMB (2025)

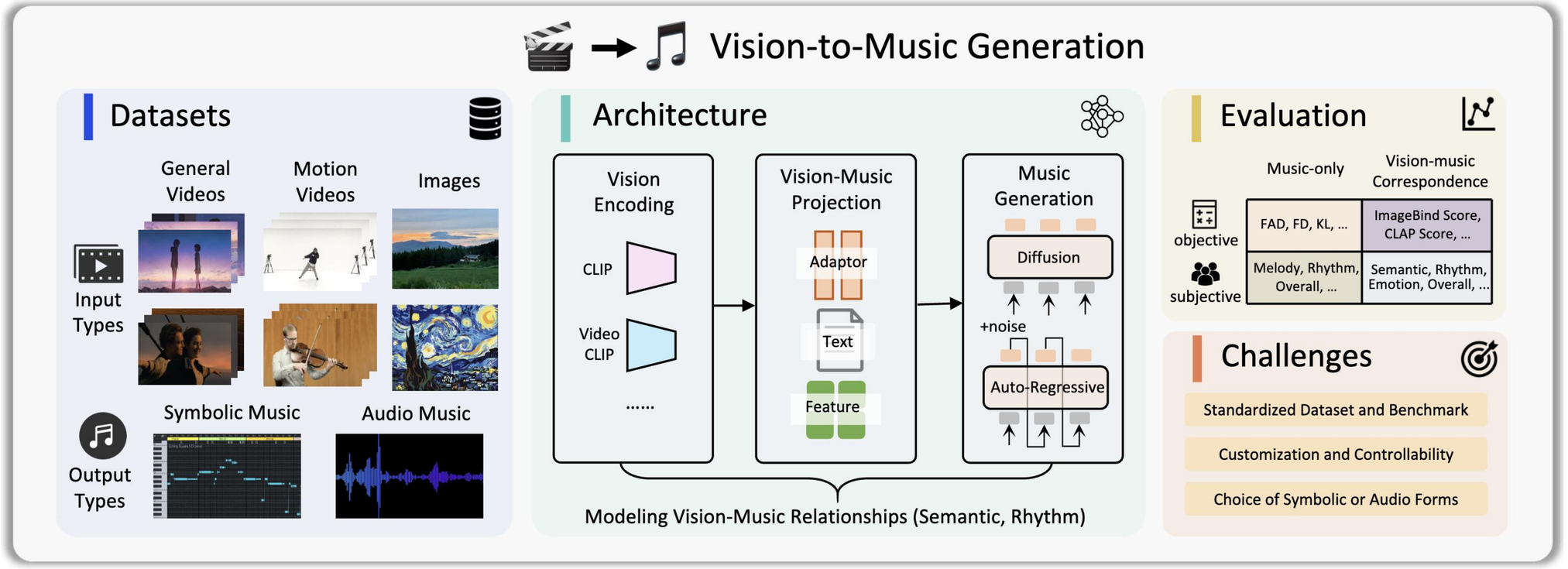
Discussion on Social Impact

A Survey on Vision-to-Music Generation: Methods, Datasets, Evaluation, and Challenges

Zhaokai Wang[1], Chenxi Bao[2], Le Zhuo[3], Jingrui Han[4],
Yang Yue[5], Yihong Tang[6], Victor Shea-Jay Huang[3], Yue Liao[7]

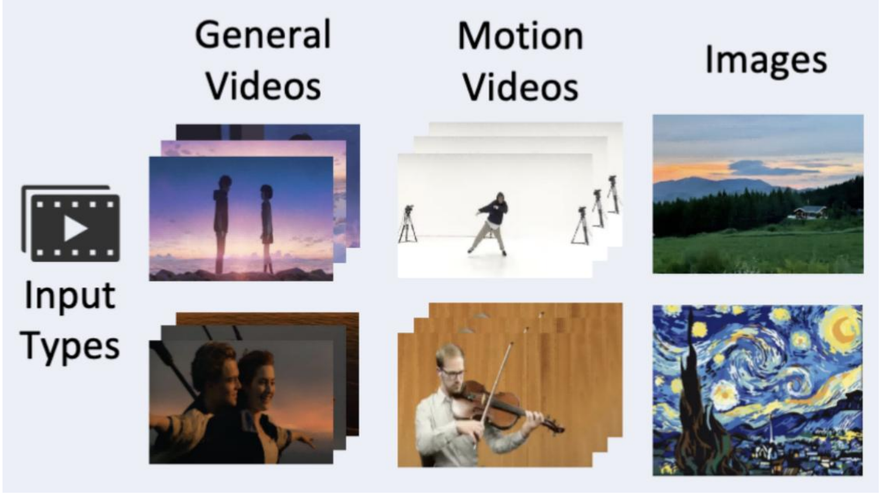
[1]Shanghai Jiao Tong University [2]Music Tech Lab, DynamiX [3]The Chinese University of Hong Kong
[4]Beijing Film Academy [5]Tsinghua University [6]McGill University [7]National University of Singapore

Overview of Vision-to-Music Generation



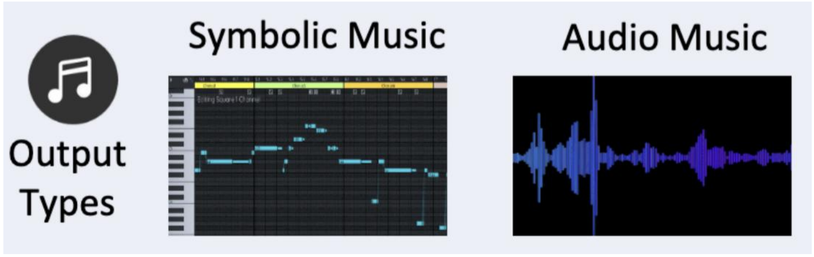
Input

- General Videos
  - E.g. natural landscapes, films, sports, animations
  - Focus on extracting features like **motion, color, or visual semantics**
- Human Movement Videos
  - Instrument performance: music is determined; more like **reconstruction**
  - Dance, sports and other human movements: emphasize local **rhythmic alignment,** semantic constraints are weaker. 2D/3D keypoints of human motion are directly used
- Images
  - Focus on overall style **without local constraints**, as images lack temporal dimension
  - Less application scenarios (iPhone photo album)



Output

- Symbolic Music
  - Discrete elements like notes, chords, or musical symbols, e.g. MIDI events
  - Early methods are mainly symbolic
  - Pros: Integration of music theory; Better controllability; Longer music pieces.
  - Cons: Limited data scalability; Weaker expressiveness
- Audio Music
  - Synthesize wave-form realistic sound in audio format
  - Pros: Large-scale datasets; Rich expressiveness
  - Cons: Lacks controllability and editing ability; Shorter music pieces



Methods

Method	Date	Input Type	Modality	Music Length	Vision-Music Relationships	Vision Encoding	Vision-Music Projection	Music Generation
▼ *General Videos and Images:*								
CMT [17]	2021/11	General Video	Symbolic	3min	Rhy	-	Elements	AR (CP [34])
V-MusProd [133]	2022/11	General Video	Symbolic	6min	Sem, Rhy	CLIP2Video [23], Histogan [3]	Feature	AR (CP [34])
V2Meow [84]	2023/05	General Video	Audio	10sec	Sem, Rhy	CLIP [76], I3D Flow [9], ViT-VQGAN [118]	Feature	AR
MuMu-LLaMA [61] (M^2UGen [62])	2023/11	General Video, Image	Audio	30sec	Sem	ViT [18], ViViT [7]	Adapter	AR (LLaMA2 [98])
Video2Music [41]	2023/11	General Video	Symbolic	5min	Sem, Rhy	CLIP [76]	Feature	AR
EIMG [106]	2023/12	Image	Symbolic	15sec	Sem	ALAE [73], β-VAE [30], VQ-VAE [100]	Adapter	VAE (FNT [88], LSR [70])
Diff-BGM [56]	2024/05	General Video	Symbolic	5min	Sem	VideoCLIP [115]	Feature	Diff. (Polyffusion [68])
Mozart's Touch [48]	2024/05	General Video, Image	Audio	10sec	Sem	BLIP [49]	Text	AR (MusicGen [16])
MeLFusion [14]	2024/06	Image	Audio	10sec	Sem	DDIM [83] + T2I LDM [78]	Feature	Diff.
VidMuse [96]	2024/06	General Video	Audio	20sec	Sem	CLIP [76]	Adapter	AR (MusicGen [16])
S2L2-V2M [39]	2024/08	General Video	Audio	10sec	Sem	Enhanced Video Mamba	Adapter	AR (LLaMA2 [98])
VMAS [59]	2024/09	General Video	Audio	10sec	Sem, Rhy	Hiera [80]	Feature	AR
MuVi [53]	2024/10	General Video	Audio	20sec	Sem, Rhy	VideoMAE V2 [104]	Adapter	Diff. (DiT [71])
SONIQUE [124]	2024/10	General Video	Audio	20sec	Sem, Rhy	Video-LLaMA [123], CLAP [19]	Text	Diff. (Stable Audio [21])
VEH [97]	2024/10	General Video	Symbolic	30sec	Sem	VideoChat [50]	Text	AR (T5 [77])
M2M-Gen [82]	2024/10	Image (Manga)	Audio	1min	Sem	CLIP [76], GPT-4 [2]	Text	AR (MusicLM [6])
HPM [75]	2024/11	General Video	Audio	10sec	Sem	CLIP [76], TAVAR [51], WECL [126]	Feature	Diff. (AudioLDM [60])
VidMusician [55]	2024/12	General Video	Audio	30sec	Sem, Rhy	CLIP [76], T5 [77]	Adapter	AR (MusicGen [16])
MTM [103]	2024/12	General Video, Image	Audio	30sec	Sem	InternVL2 [12]	Text	Diff. (Stable Audio Open [22])
XMusic [94]	2025/01	General Video, Image	Symbolic	20sec	Sem, Rhy	ResNet [28], CLIP [76]	Elements	AR (CP [34])
GVMGen [134]	2025/01	General Video	Audio	15sec	Sem	CLIP [76]	Adapter	AR (MusicGen [16])
AudioX [95]	2025/03	General Video	Audio	10sec	Sem	CLIP [76]	Feature	Diff. (Stable Audio Open [22])
FilmComposer [112]	2025/03	General Video	Audio	15sec	Sem, Rhy	Controllable Rhythm Transformer, GPT-4v [2], Motion Detector	Text	AR (MusicGen [16])
▼ *Human Movement Videos:*								
Audeo [85]	2020/06	Performance Video	Symbolic	30sec	Rhy	ResNet [28]	Feature	GAN
Foley Music [24]	2020/07	Performance Video	Symbolic	10sec	Rhy	2D Body Keypoints	Feature	AR
Multi-Instrucment Net [86]	2020/12	Performance Video	Audio	10sec	Rhy	2D Body Keypoints	Feature	VAE
RhythmicNet [87]	2021/06	Dance Video	Symbolic	10sec	Rhy	2D Body Keypoints	Feature	AR (REMI [36])
Dance2Music [4]	2021/07	Dance Video	Symbolic	12sec	Rhy	2D Body Keypoints	Feature	AR
D2M-GAN [129]	2022/04	Dance Video	Audio	2sec	Rhy	2D Body Keypoints, I3D [10]	Feature	GAN
CDCD [130]	2022/06	Dance Video	Audio	2sec	Rhy	2D Body Keypoints, I3D [10]	Feature	Diff.
LORIS [119]	2023/05	Movement Video	Audio	50sec	Rhy	2D Body Keypoints, I3D [10]	Feature	Diff.
VisBeatNet [63]	2024/01	Dance Video	Symbolic	Realtime	Rhy	2D Body Keypoints	Feature	AR
UniMuMo [116]	2024/10	Dance Video	Audio	10sec	Rhy	2D Body Keypoints	Feature	Diff.

Datasets

Dataset	Access	Date	Source	Modality	Size	Total Length (hour)	Avg. Length (second)	Annotations
▼ *General Videos:*								
HIMV-200K [32]	Link	2017/04	Music Video (Youtube-8M [1])	Audio	200K	-	-	-
MVED [69]	Link	2020/09	Music Video	Audio	1.9K	16.5	30	Emotion
SymMV [133]	Link	2022/11	Music Video	MIDI, Audio	1.1K	76.5	241	Lyrics, Genre, Chord, Melody, Tonality, Beat
MV100K [84]	-	2023/05	Music Video (Youtube-8M [1])	Audio	110K	5000	163	Genre
MusicCaps [6]	Link	2023/01	Diverse Videos (AudioSet [25])	Audio	5.5K	15.3	10	Genre, Caption, Emotion, Tempo, Instrument, Rhythm, ...
EmoMV [93]	Link	2023/03	Music Video (MVED [69], AudioSet [25])	Audio	6K	44.3	27	Emotion
MUVideo [62]	Link	2023/11	Diverse Videos (Balanced-AudioSet [25])	Audio	14.5K	40.3	10	Instructions
MuVi-Sync [41]	Link	2023/11	Music Video	MIDI, Audio	784	-	-	Scene Offset, Emotion, Motion, Semantic, Chord, Key, Loudness, Density, ...
BGM909 [56]	Link	2024/05	Music Video	MIDI	909	-	-	Caption, Style, Chord, Melody, Beat, Shot
V2M [96]	-	2024/06	Diverse Videos	Audio	360K	18000	180	Genre
DISCO-MV [59]	-	2024/09	Music Video (DISCO-10M [45])	Audio	2200K	47000	77	Genre
FilmScoreDB [75]	-	2024/11	Film Video	Audio	32K	90.3	10	Movie Title
DVMSet [55]	-	2024/12	Diverse Videos	Audio	3.8K	-	-	-
HarmonySet [128]	Link	2025/03	Diverse Videos	Audio	48K	458.8	32	Description
MusicPro-7k [112]	Link	2025/03	Film Video	Audio	7K	-	-	Description, Melody, Rhythm Spots
▼ *Human Movement Videos*								
URMP [47]	Link	2016/12	Performance Video	MIDI, Audio	44	1.3	106	Instruments
MUSIC [127]	Link	2018/04	Performance Video	Audio	685	45.7	239	Instruments
AIST++ [52]	Link	2021/01	Dance Video (AIST [99])	Audio	1.4K	5.2	13	3D Motion
TikTok Dance-Music [129]	Link	2022/04	Dance Video	Audio	445	1.5	12	-
LORIS [119]	Link	2023/05	Dance Video, Sports Video (AIST [99], FisV [114], FS1000 [111])	Audio	16K	86.43	19	2D Pose
▼ *Images*								
Music-Image [110]	Link	2016/07	Image (Music Video)	Audio	22.6K	377	60	Lyrics
Shuttersong [57]	Link	2017/08	Image (Shuttersong App)	Audio	586	-	-	Lyrics
IMAC [102]	Link	2019/04	Image (FI [117])	Audio	3.8K	63.3	60	Emotion
MUImage [62]	Link	2023/11	Image (Balanced-AudioSet [25])	Audio	14.5k	40.3	10	Instructions
EIMG [106]	Link	2023/12	Image (IAPS [44], NAPS [67])	MIDI	3K	12.5	15	VA Value
MeLBench [14]	Link	2024/06	Image (Diverse Videos)	Audio	11.2K	31.2	10	Genre, Caption

Evaluation Metrics

Metric	Used in Paper	Input	Type
▼ *Music-only:*			
Scale Consistency	[133, 39]	M	Pit
Pit Entropy	[133, 39, 106]	M	Pit
Pit Class Histogram Entropy	[133, 94, 39, 17, 56]	M	Pit
Empty Beat Rate	[133, 94, 39]	M	Rhy
Average Inter-Onset Interval	[133, 39]	M	Rhy
Grooving Pattern Similarity	[94, 17, 56]	M	Rhy
Structure Indicator	[17, 56]	M	Rhy
Frechet Audio Distance (FAD)	[96, 62, 14, 84, 53, 134, 112] [48, 39, 75, 97, 95, 59]	A	Fid
Frechet Distance (FD)	[103, 55, 124, 53, 96, 14, 84, 95]	A	Fid
Kullback-Leibler Divergence (KL)	[59, 96, 48, 62, 134, 103, 55, 84, 53, 112, 75] [124, 14, 97, 95, 39]	A	Fid
Beats Coverage Score (BCS)	[75, 53]	A	Rhy
Beats Hit Score (BHS)	[75, 53]	A	Rhy
Inception Score (IS)	[75, 53, 95]	A	Fid
▼ *Vision-music Correspondence:*			
ImageBind Score/Rank	[103, 55, 39, 96, 48, 62, 112, 95]	A,V/I	Sem
CLAP Score	[55, 124, 97]	A,A/T	Sem
Video-Music CLIP Precision	[133, 39]	A,V	Sem
Video-Music Correspondence	[56]	A,V	Sem
Cross-modal Relevance	[134]	A,V	Sem
Temporal Alignment	[134]	A,V	Rhy
Rhythm Alignment	[55]	A,V	Rhy

Objective Metrics

Metric	Used in Paper
▼ *Music-only:*	
Music Melody	[133, 56]
Music Rhythm	[133, 56]
Music Richness	[134, 94]
Audio Quality	[96, 53]
Overall Music Quality	[134, 82, 59, 41, 48, 14, 84, 97, 103, 96, 17, 112]
▼ *Vision-music Correspondence:*	
Semantic Consistency	[133, 103, 55, 53, 56, 112]
Rhythm Consistency	[133, 94, 103, 97, 56, 41, 112]
Emotion Consistency	[94, 103, 97]
Overall Correspondence	[133, 124, 82, 59, 39, 96, 17, 41, 48, 62, 14, 84, 134, 112]

Subjective Metrics

Challenges

- Lack of Standardized Datasets and Benchmarks
- Limited Customization and Controllability
- Trade-off Between Symbolic and Audio Forms
- Cross-Domain Generalization
- Limited Human-in-the-Loop Feedback for Iterative Improvement
- Under-utilization of Large Model Capabilities
- Industry Application Challenges
- ...

Overview of Video-to-Music

Initial Attempt: CMT (2021)

Advanced Method: MusProd (2023)

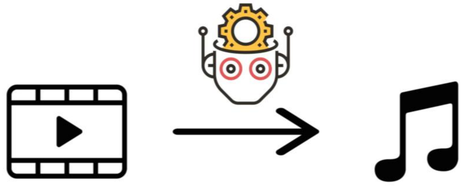
Recent Work: VMB (2025)

Discussion on Social Impact

Video Background Music Generation with Controllable Music Transformer

Shangzhe Di[1] Zeren Jiang[1] Si Liu[1] Zhaokai Wang[1]
Leyan Zhu[1] Zexin He[1] Hongming Liu[2] Shuicheng Yan[3]
[1]Beihang University, China [2]Charterhouse School, UK [3]Sea AI Lab, Singapore

ACM Multimedia 2021 Best Paper Award



Motivation

4 Million Active Video Producers[1]

500 Hours of Videos Uploaded Every Minute[1]

[1] Statistics in 2021, https://influencermarketinghub.com/youtube-stats/

Motivation

Background Music is Important for Video Production[2]

Conveys messages effectively

Makes content memorable

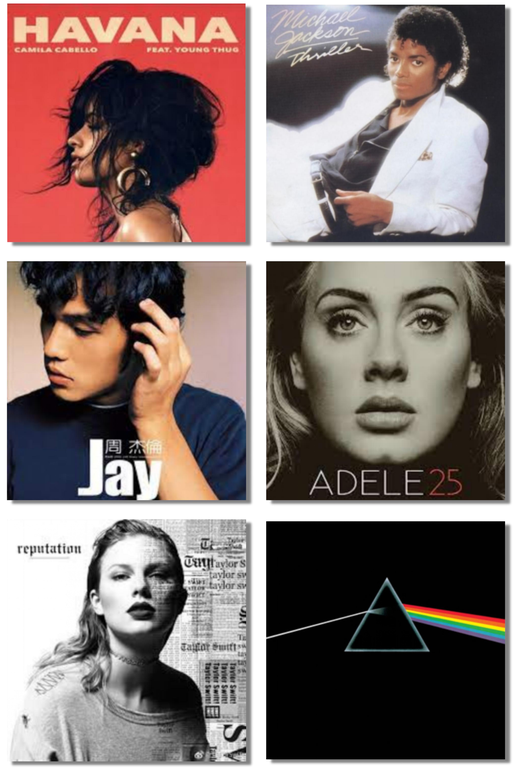
Depicts emotions

Attracts attention

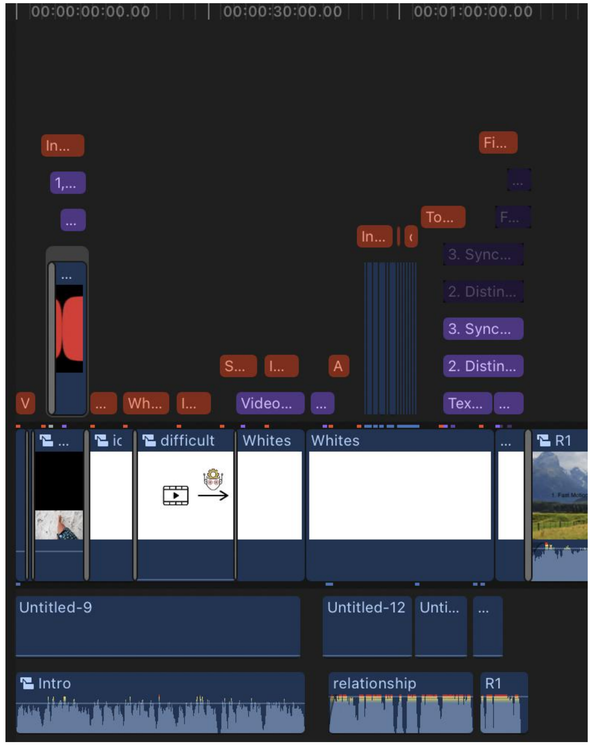
[2] https://musicforproductions.com/importance-of-background-music-in-your-videos-2/

Motivation

Prepare video clips



Pick background music



Edit video to fit music

Motivation

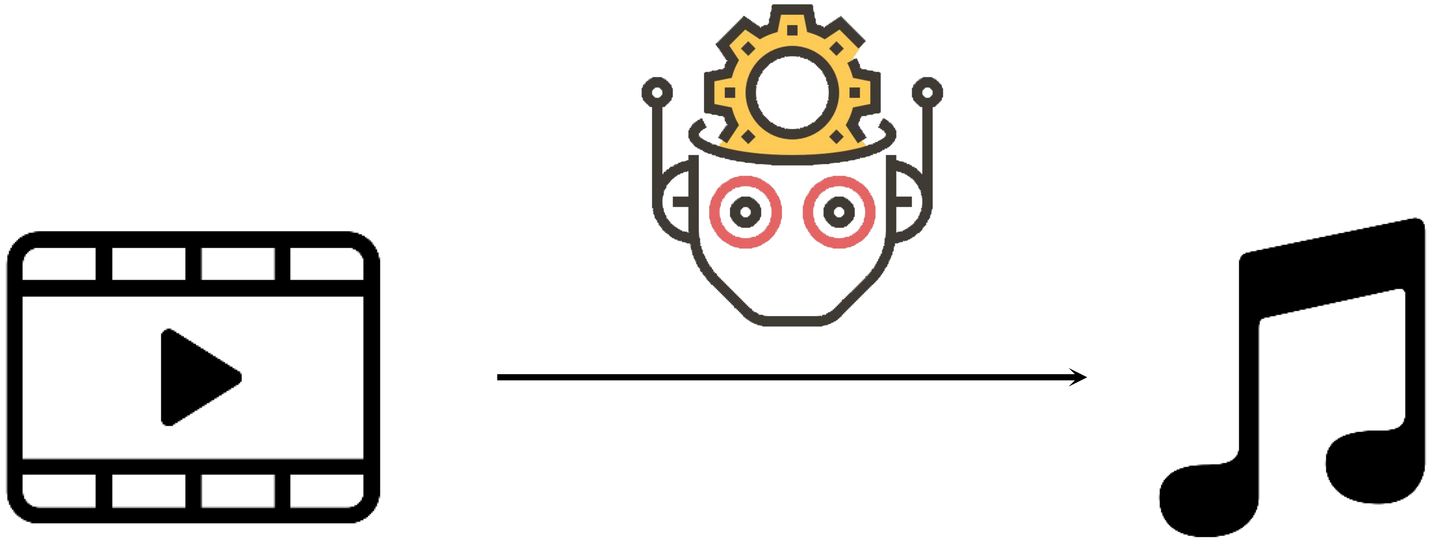
Make background music fit frame by frame

Time Consuming!

Copyright Issues

Prepare video clips **Pick background music** **Edit video to fit music**

Motivation



Motivation

Retrieval?

Video Background Music Generation

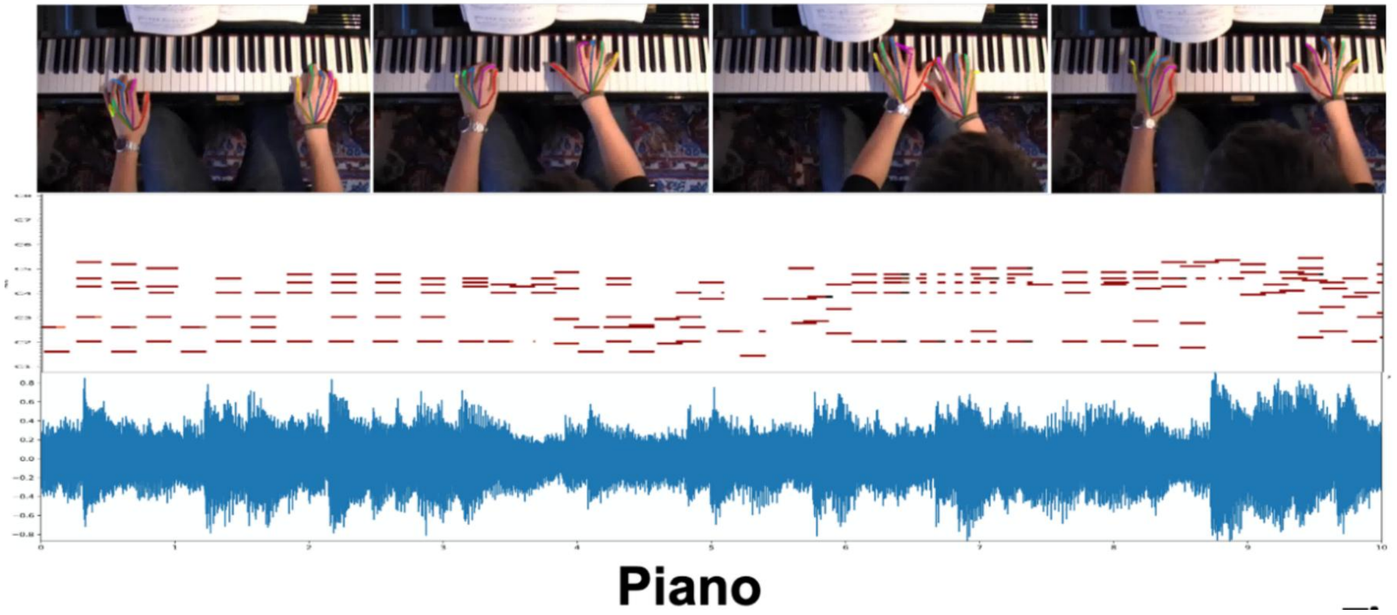
Dramatically shorten video production time

Save a lot of copyright fees

Meet people's individual needs

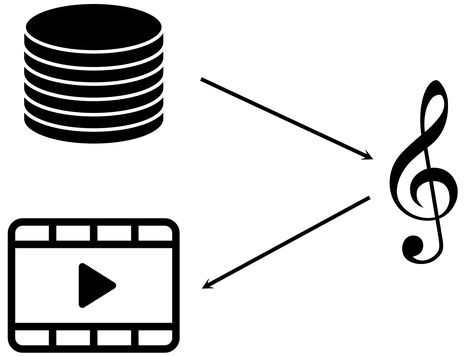
Related Works

- **Music Reconstruction from Silent Performance Video**
  - Foley Music (Gan et al., in ECCV'20), Audeo (Su et al., in NeurIPS'20)



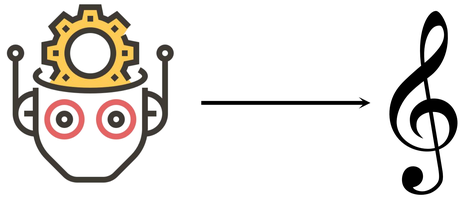
Related Works

- **Music Reconstruction from Silent Performance Video**
  - Foley Music (Gan et al., in ECCV’20), Audeo (Su et al., in NeurIPS’20)
- **Video Background Music Retrieval**
  - MRCMV (Li et al., in SIGIR’21)



Related Works

- **Music Reconstruction from Silent Performance Video**
  - Foley Music (Gan et al., in ECCV'20), Audeo (Su et al., in NeurIPS'20)
- **Video Background Music Retrieval**
  - MRCMV (Li et al., in SIGIR'21)
- **Music Generation**
  - MuseNet (OpenAI, in 2019), Jukebox (OpenAI, in 2020)

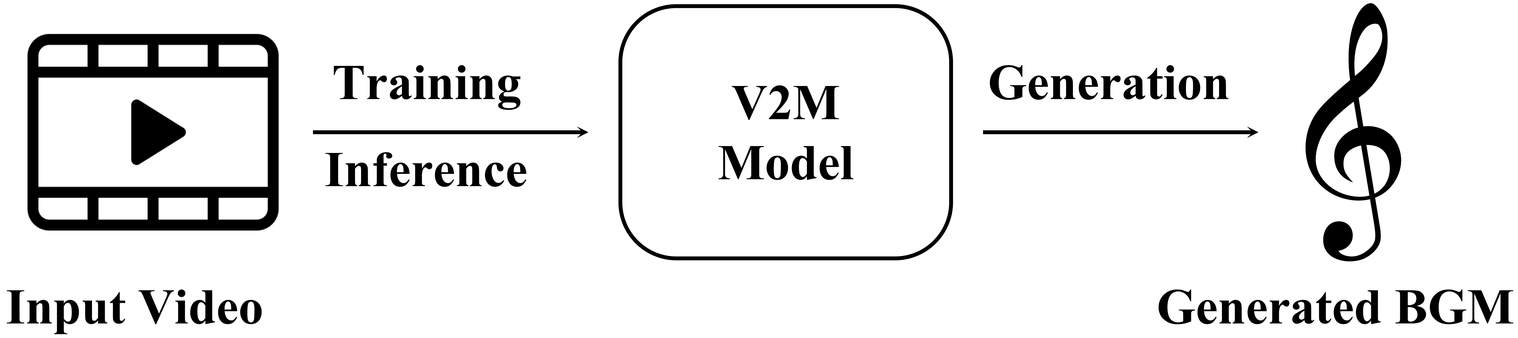


Related Works

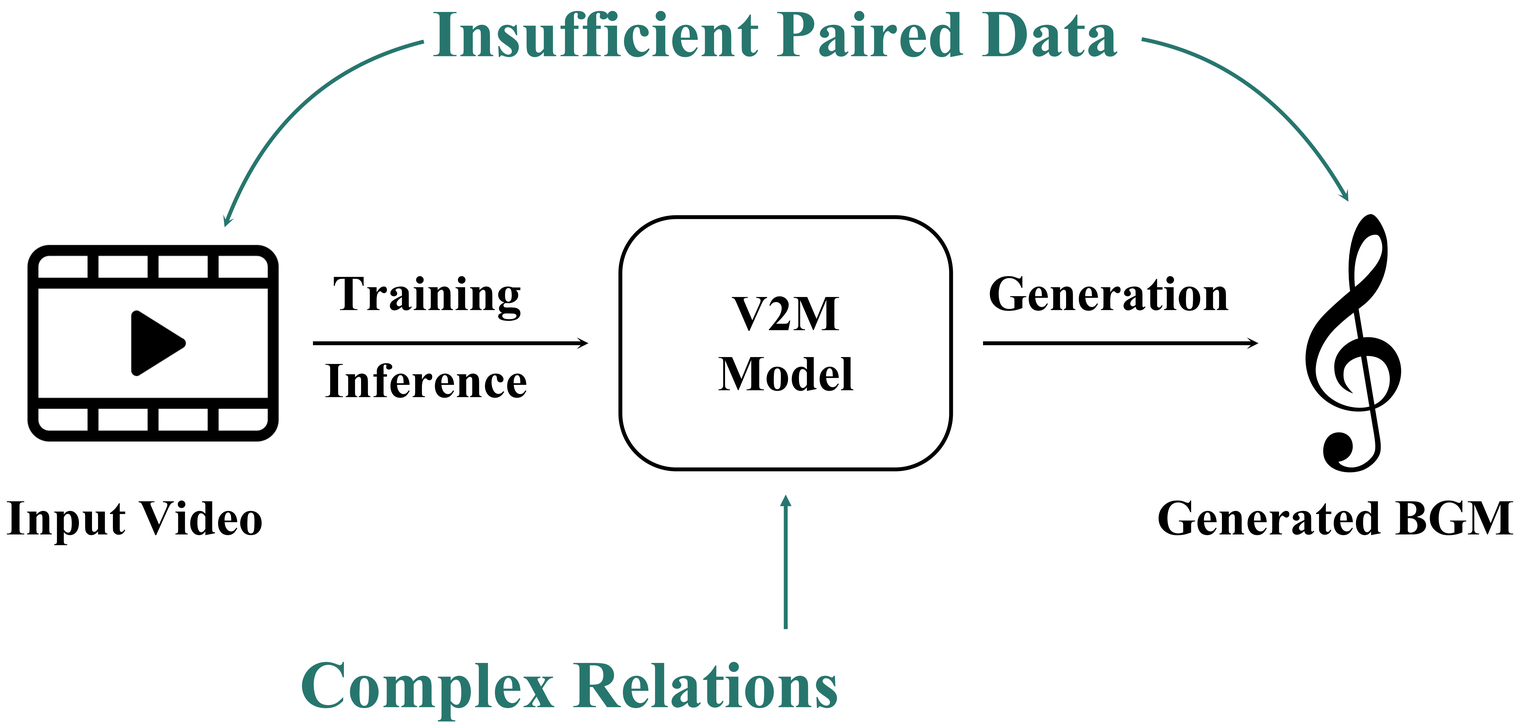
- **Music Reconstruction from Silent Performance Video**
  - Foley Music (Gan et al., in ECCV'20), Audeo (Su et al., in NeurIPS'20)
- **Video Background Music Retrieval**
  - MRCMV (Li et al., in SIGIR'21)
- **Music Generation**
  - MuseNet (OpenAI, in 2019), Jukebox (OpenAI, in 2020)

✕ **Cannot generate BGM tailored to a particular video.**

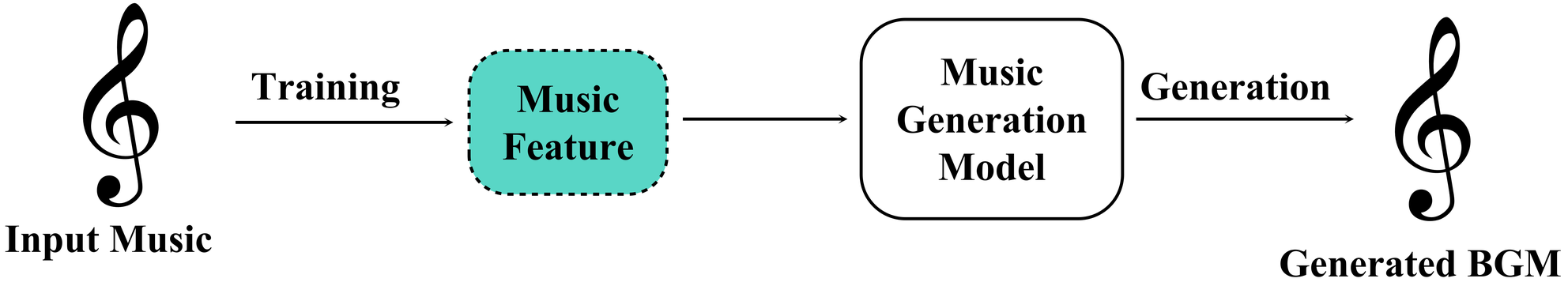
Video Background Music Generation



Video Background Music Generation

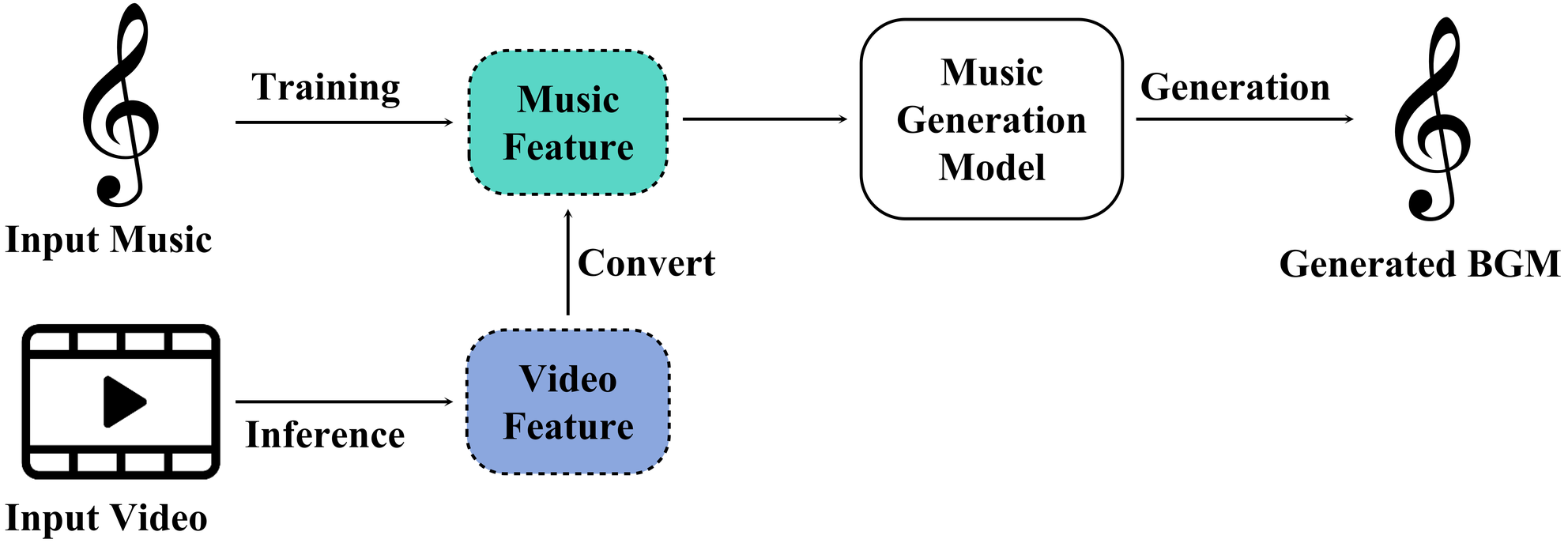


Video Background Music Generation



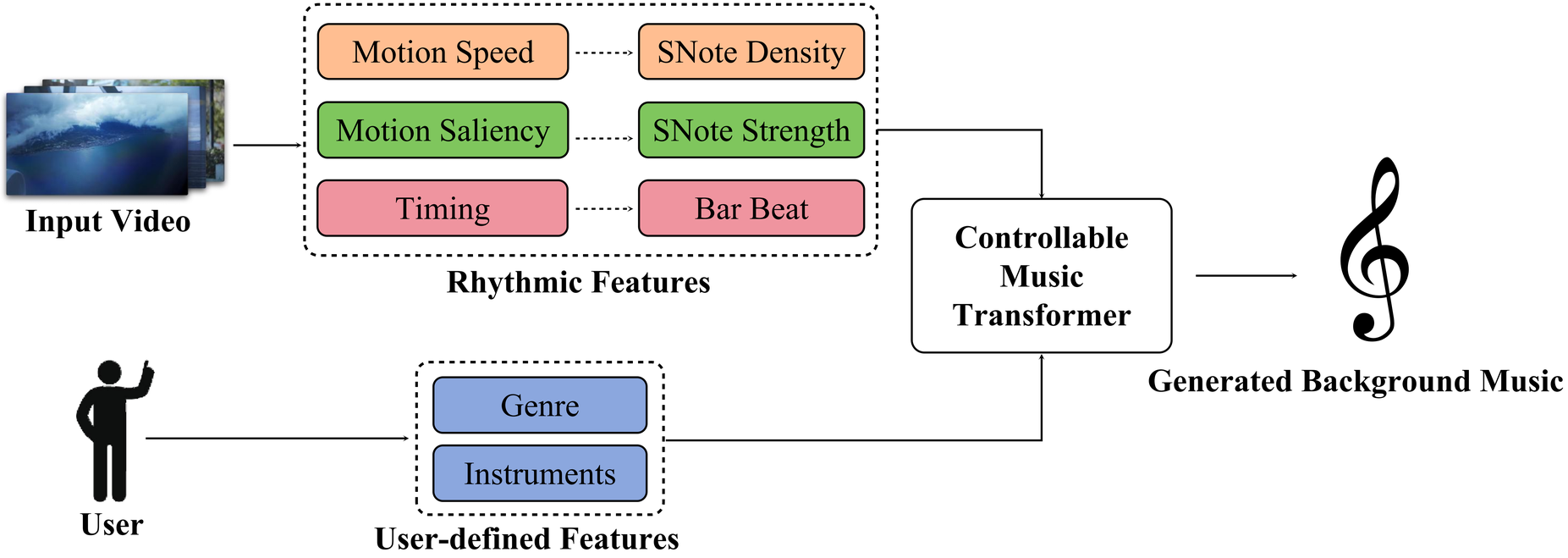
Train with only music data!

Video Background Music Generation

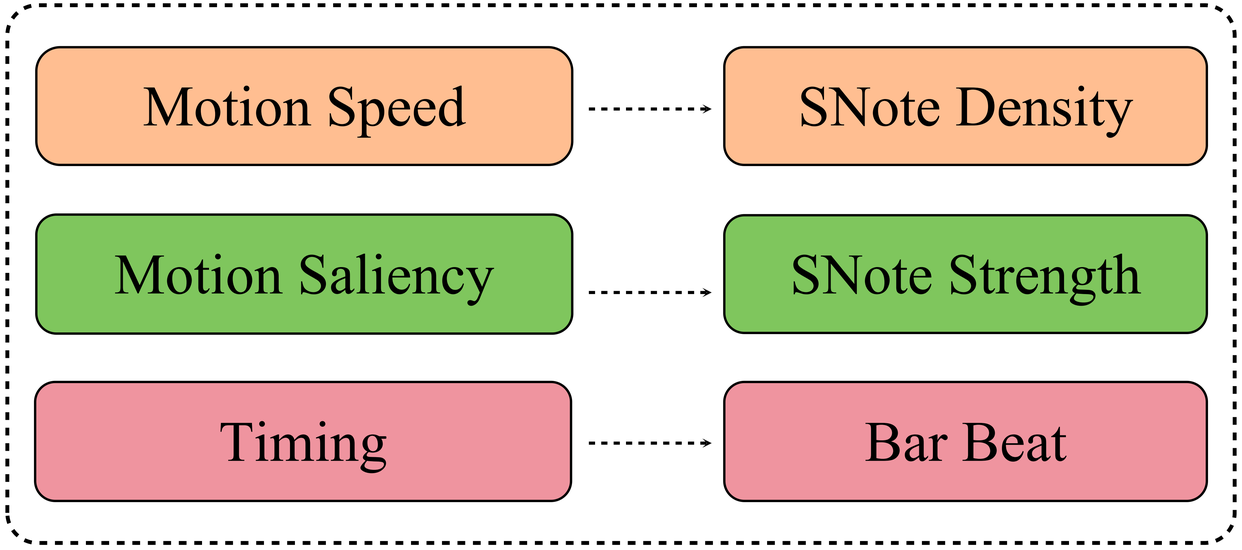


Train with only music data!

Our Method


Motion Speed
SNote Density
Motion Saliency
SNote Strength
Timing
Bar Beat
Rhythmic Features
Input Video
Controllable Music Transformer
Generated Background Music
Genre
Instruments
User-defined Features
User

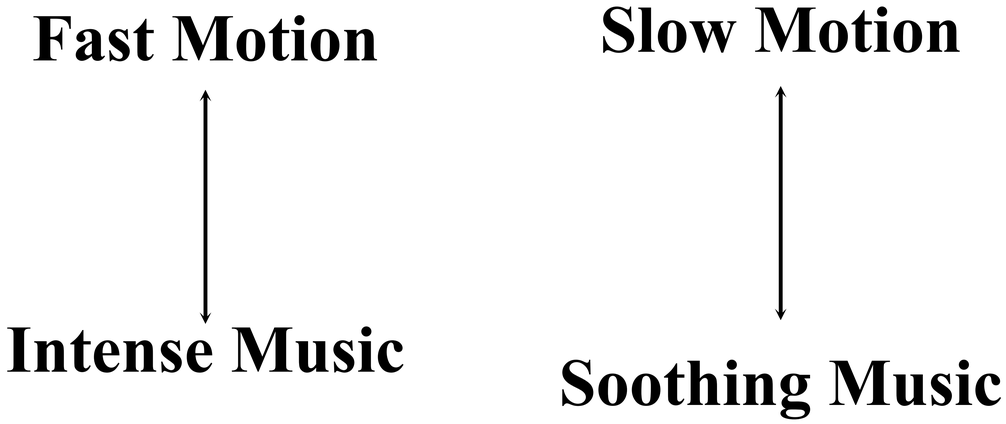
Rhythmic Relations



Rhythmic Relations

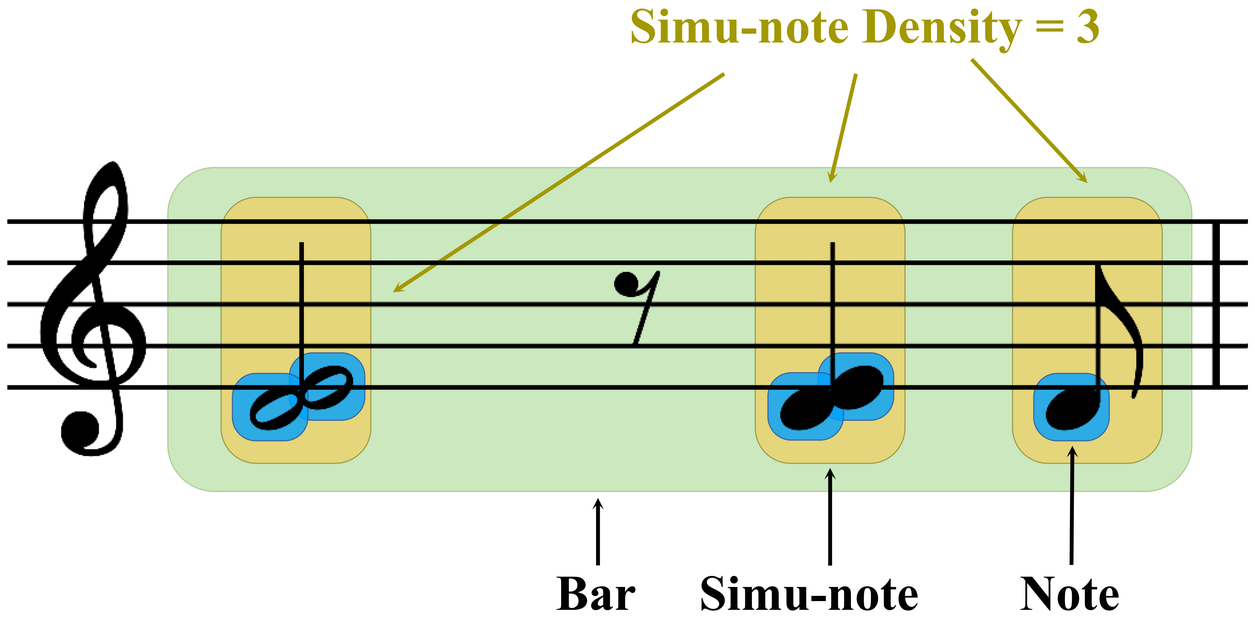
Rhythmic Relations

Motion Speed ⇌ Simu-note Density



Rhythmic Relations

Motion Speed ⇌ Simu-note Density



The higher the simu-note density

The faster the local music rhythm.

Rhythmic Relations

Motion Speed ⇌ Simu-note Density

Optical Flow: $f_t(x, y) \in \mathbb{R}^{H \times W \times 2}$

Optical Flow Magnitude: $F_t = \frac{1}{HW} \sum_{x,y} |f_t(x, y)|$

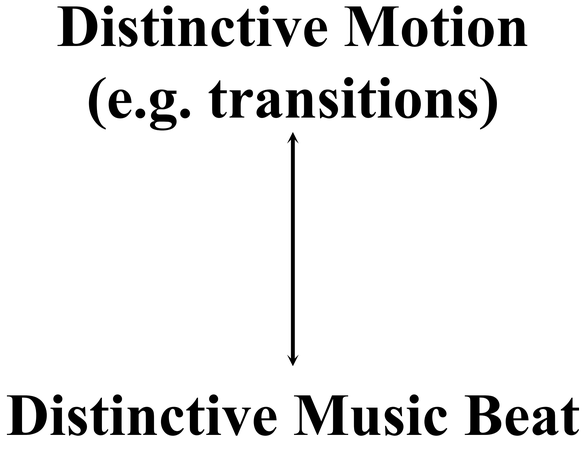
average of the absolute optical flow over pixels of a flow map

Motion Speed: $\frac{1}{t_1 - t_0} \sum_{t=t_0}^{t_1} F_t$

average of the optical flow magnitude over frames of a clip

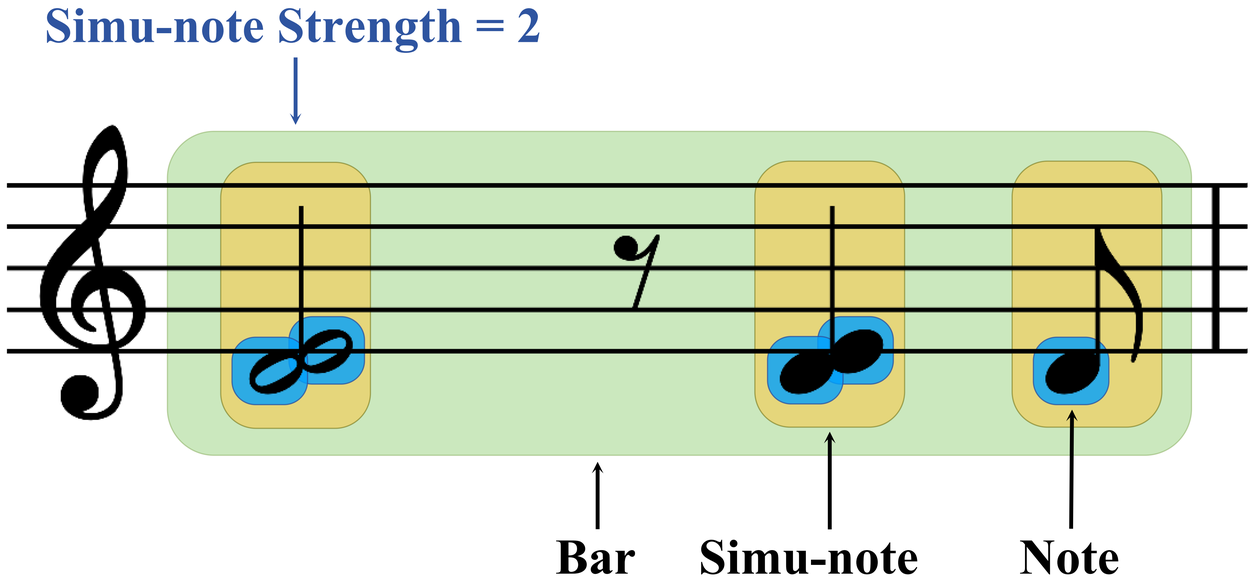
Rhythmic Relations

Motion Saliency ⇌ Simu-note Strength



Rhythmic Relations

Motion Saliency ⇌ Simu-note Strength



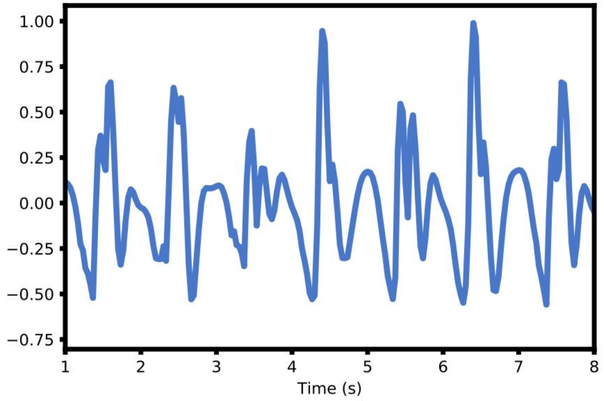
The larger the strength

The richer and more distinct the sound

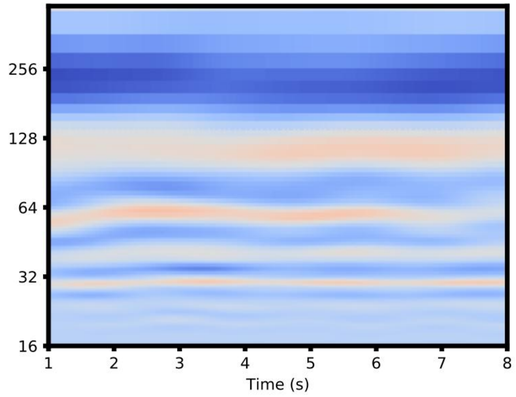
Rhythmic Relations

Motion Saliency ⇌ Simu-note Strength

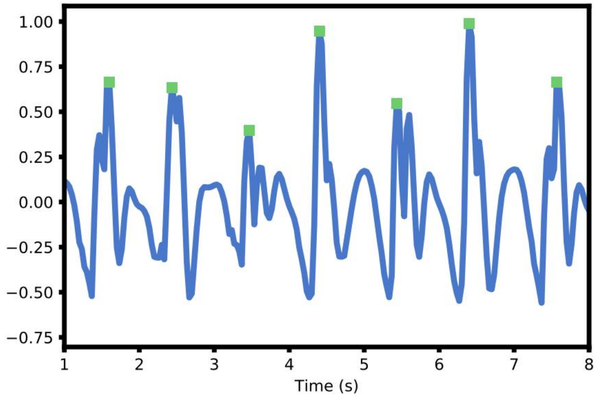
Visual beats are sudden decelerations. [1]



Motion Saliency: the summation of per-direction decelerations



Dominant tempo: the largest spike in the autocorrelation of motion saliency



Visual beats: local maxima above a threshold and follow the dominant tempo

[1] Abe Davis and Maneesh Agrawala, **Visual rhythm and beat**, *SIGGRAPH 2018*.

Rhythmic Relations

Synchronize the Opening and Ending of Video and Music

Video Start/Ending

Music Appear/Disappear

Rhythmic Relations

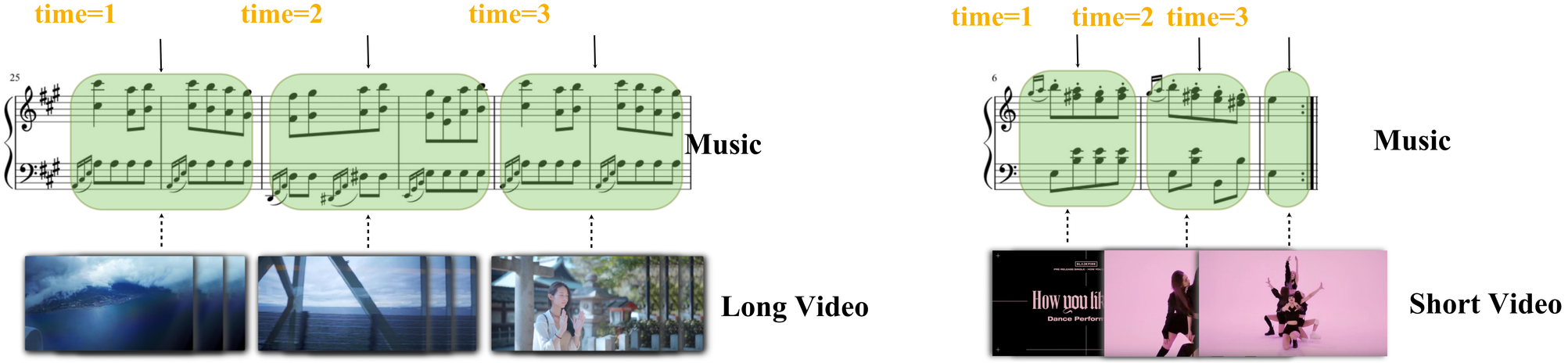
Synchronize the opening and ending of Video and Music

Time encoding

Add different positional encoding to tokens of different pieces

$$t_i = \text{Embedding}_t\left(round\left(M\frac{beat_i}{N_{beat}}\right)\right),$$

$$\vec{x_i} = x_i + BPE + t_i,$$

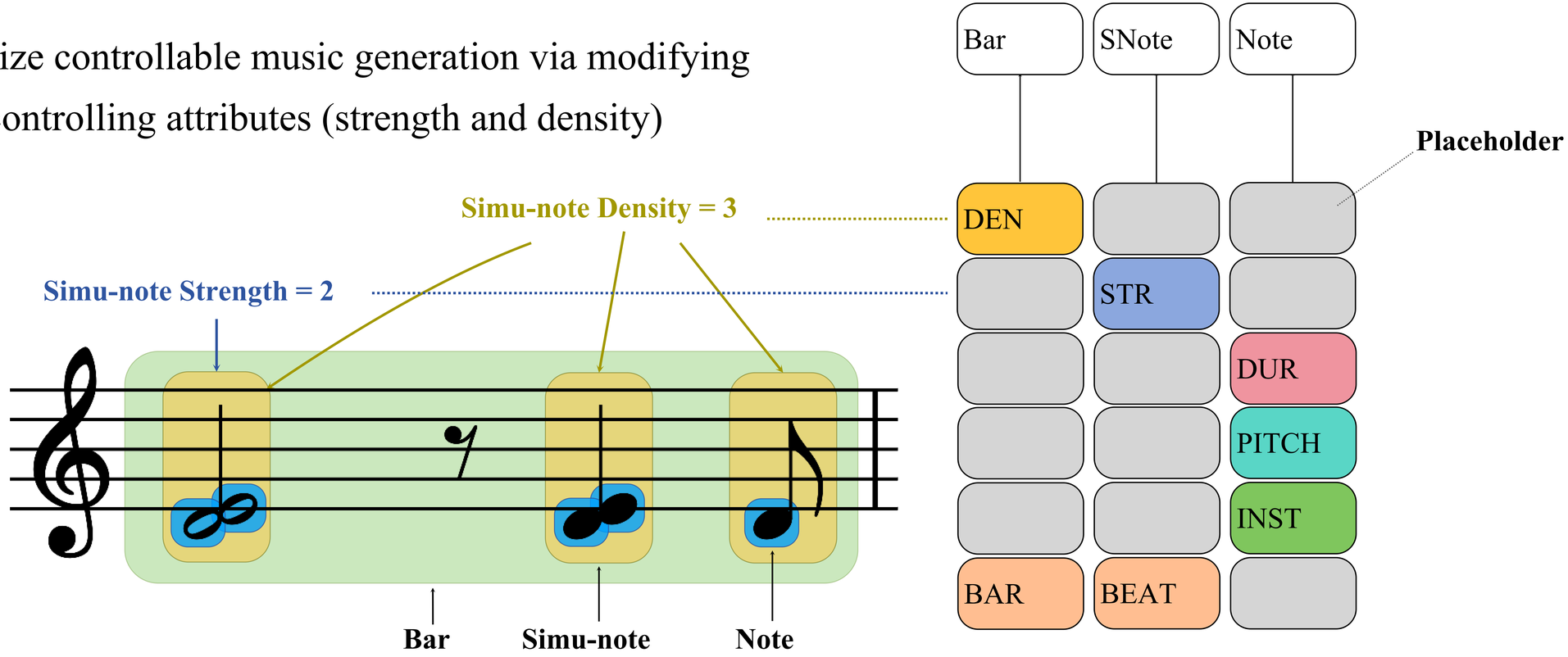


Controllable Music Transformer

Multi-track Symbolic Music Representation

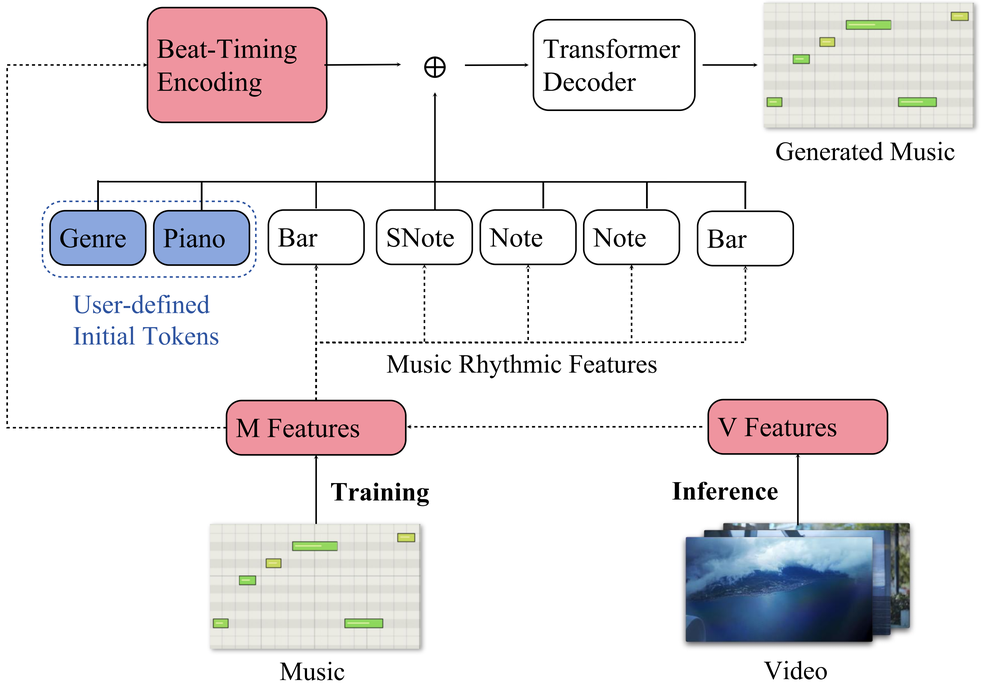
2 types of tokens, 6 attributes

Realize controllable music generation via modifying the controlling attributes (strength and density)



Controllable Music Transformer

Framework



- **Training:** reconstructing music given the extracted music features
- **Inference:**
  - Convert visual features into music features
  - Generate background music

Experiments

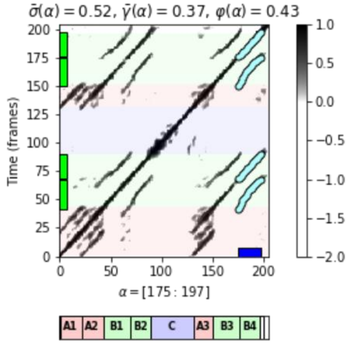
Dataset

- Lakh Pianoroll Dataset (LPD) [1]
- LPD-5-cleansed version of LPD
- 5 instruments (Drums, Piano, Guitar, Bass and Strings)
- 6 types (Country, Dance, Electronic, Metal, Pop, Rock)
- 3,038 MIDI music pieces

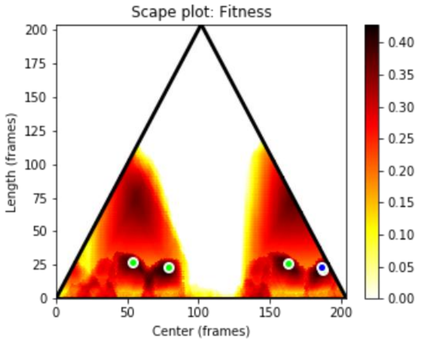
[1] Hao-Wen Dong, et al, **MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment**, *AAAI 2018.*

Experiments

Objective evaluation metrics



Self-similarity matrix



Fitness Scape Plot

- **Pitch Histogram Entropy**
  - Assessing the music's quality in tonality.
  - If a piece's tonality is clear, several pitch classes should dominate the pitch histogram, resulting in a low entropy.
- **Grooving Pattern Similarity**
  - Measuring the music's rhythmicity.
  - If a piece possesses a clear sense of rhythm, the grooving patterns between pairs of bars should be similar, thereby producing high scores .
- **Structure Indicators**
  - Measuring the music's repetitive structure.

[1] Shih-Lun Wu and Yi-Hsuan Yang, **The Jazz Transformer on the Front Line: Exploring the Shortcomings of AI-composed Music through Quantitative Measures**, *ISMIR 2021*.

Experiments

Objective evaluation

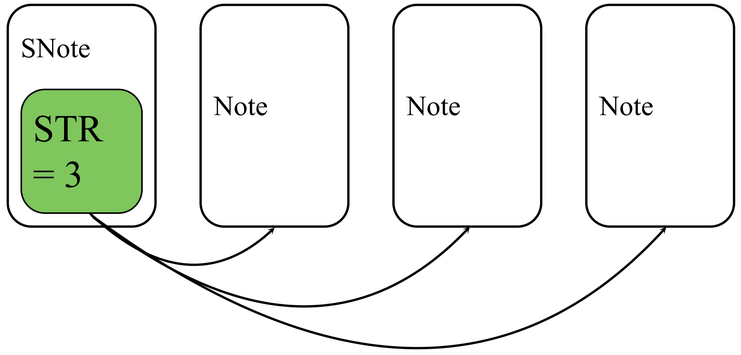
Model	Data	Without Control				
No.	-	1	2	3	4	5
Density	-	-	○	-	-	○
Strength	-	-	-	○	-	○
Beat-timing encoding	-	-	-	-	√	√
Pitch Histogram Entropy	4.452	3.634	2.998	3.667	3.573	3.617
Grooving Pattern Similarity	0.968	0.677	0.714	0.647	0.778	**0.810**
Structureness Indicator	0.488	0.219	0.227	0.215	0.223	0.241
Overall Rank ↓	-	5.000	5.000	5.333	4.000	**2.667**

Simu-note Density

Simu-note Strength

Beat-timing Encoding

Music Generation



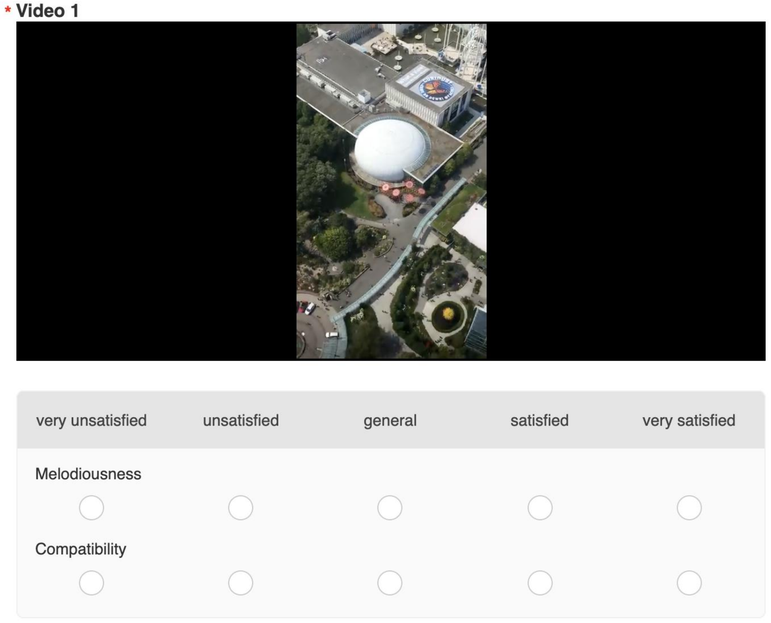
Plan-before-generation Strategy

Make the model easier to learn

Experiments

Subjective evaluation

- Questionnaire
  - 36 participants
  - Videos from 3 categories (edited, unedited, and animation)
  - Rate BGM's melodiousness and compatibility w/ video
  - Rank the three models from an overall perspective



- Result

Music selected that best matches the input video

Music generated w/o control

Model	Baseline	Matched	Ours
Melodiousness ↑	3.4	**4.0**	3.8
Compatibility ↑	3.4	3.7	**3.9**
Overall Rank ↓	2.3	1.9	**1.8**

↑: the higher the better, ↓: the lower the better.

Bolds: best performance

1. Music generation model is comparable to human composers in terms of melodiousness.

2. Our method has the best performance for compatibility and overall ranking

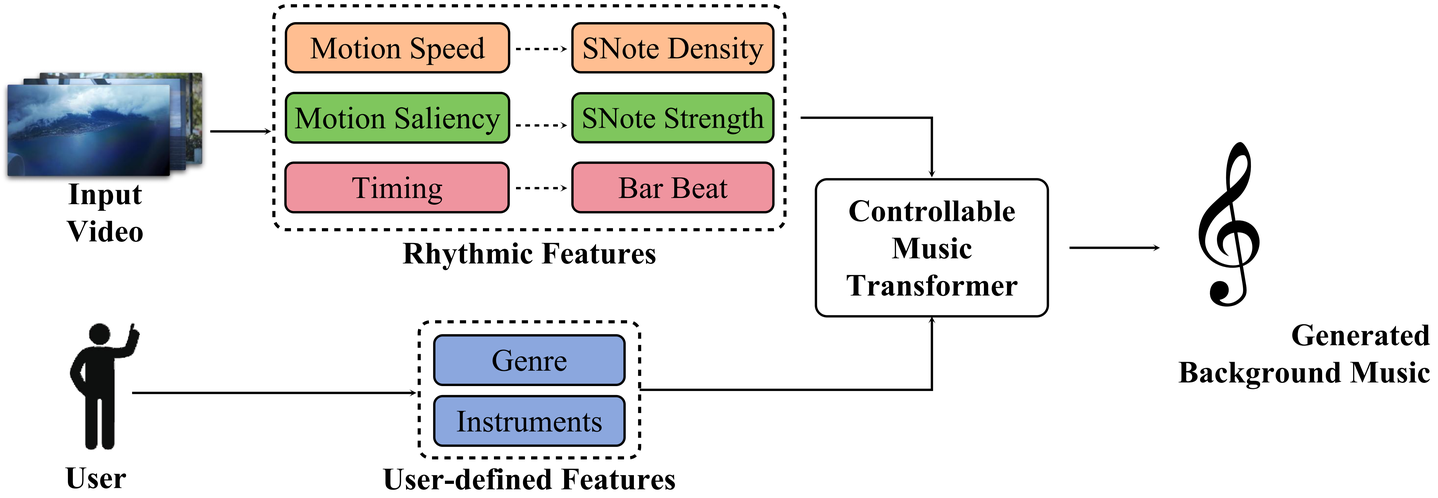
Experiments

Demo: https://wzk1015.github.io/cmt/

Conclusion

◆The first method for video background music generation

◆Established rhythmic relations between music and video



◆Designed a controllable music generation model that can be trained without paired dataset

◆Can generate impressive results

Overview of Video-to-Music

Initial Attempt: CMT (2021)

Advanced Method: MusProd (2023)

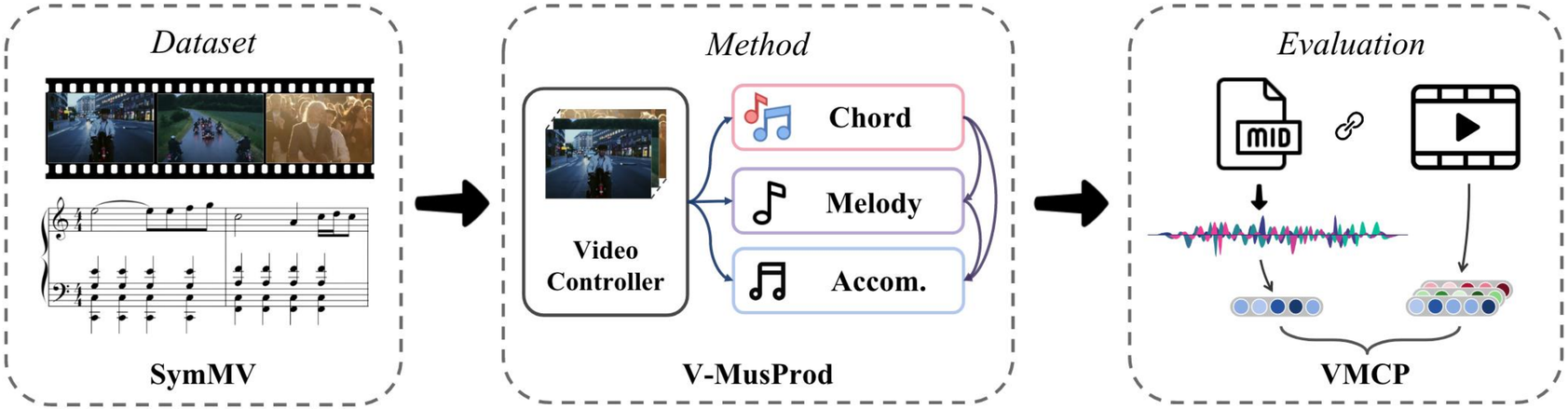
Recent Work: VMB (2025)

Discussion on Social Impact

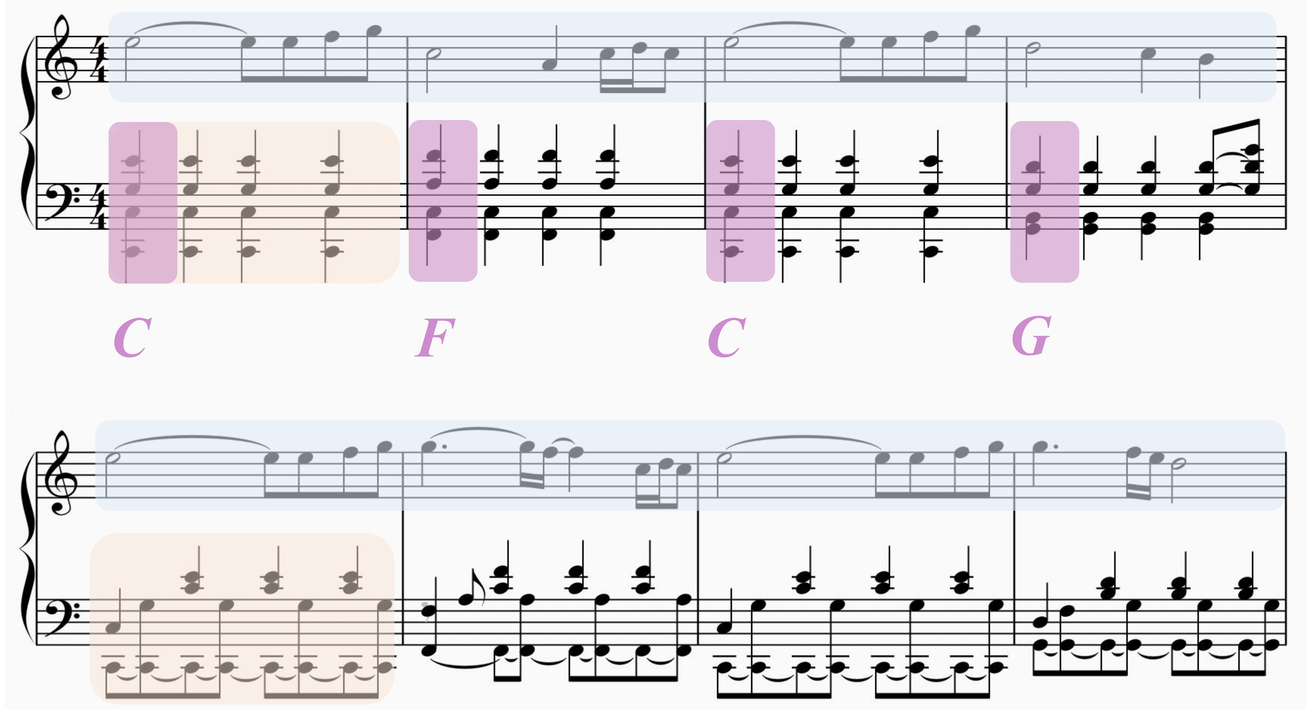
Video Background Music Generation: Dataset, Method and Evaluation

Le Zhuo[1*] Zhaokai Wang[1*] Baisen Wang[1*] Yue Liao[1†] Chenxi Bao[1,2]
Stanley Peng[1] Songhao Han[1] Aixi Zhang[3] Fei Fang[3] Si Liu[1]
[1]Beihang University [2]Edinburgh College of Art, University of Edinburgh [3]Alibaba Group

ICCV 2023

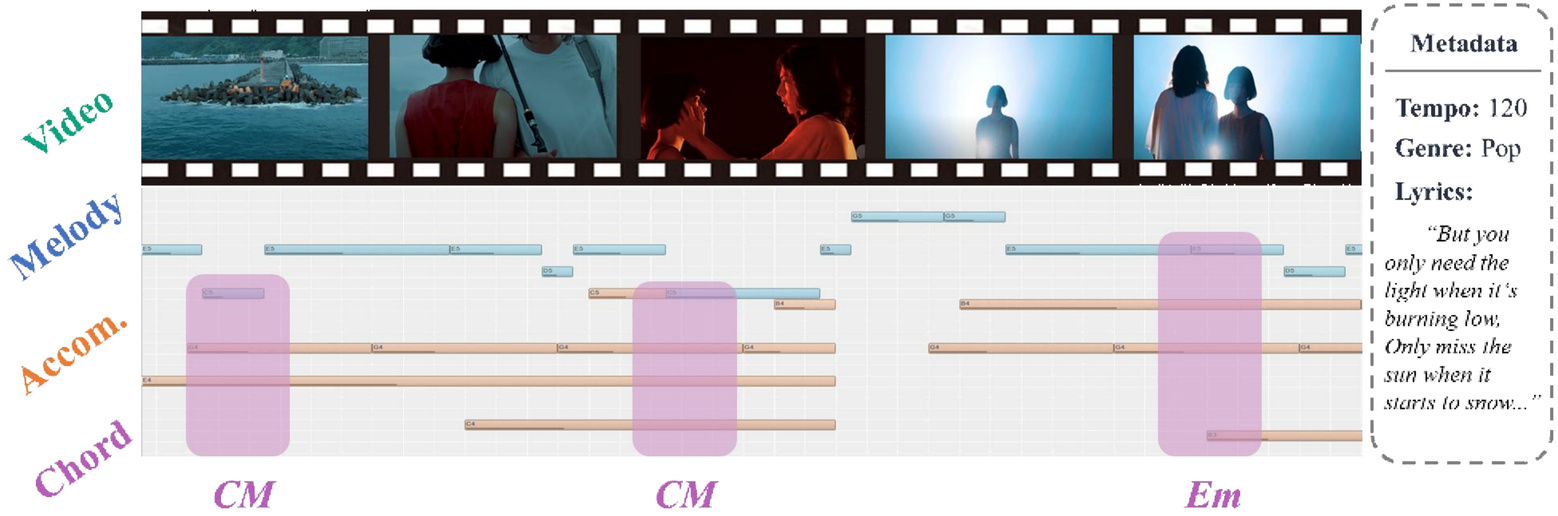


Video Background Music Generation:
Dataset, Method and Evaluation



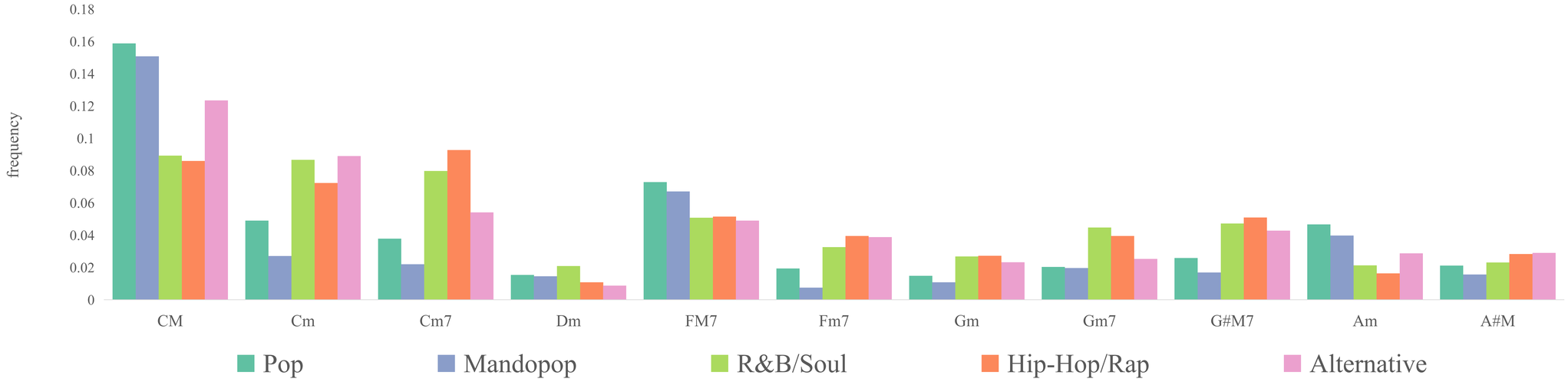
Video Background Music Generation:
Dataset, Method and Evaluation

- Crawl piano performance audio from music videos, convert the piano audio into MIDI files using a music transcription model to construct video-MIDI pairs
- Extract music features, crawl other metadata, and further incorporate music theory knowledge.



Video Background Music Generation:
Dataset, Method and Evaluation

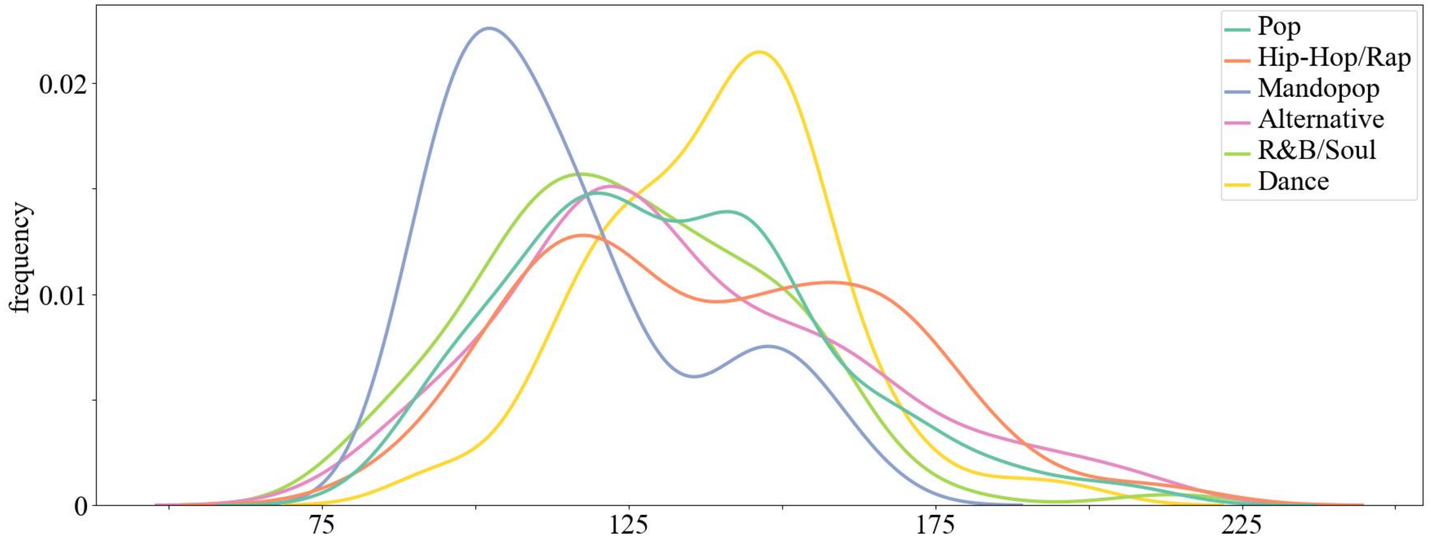
- Use genre as a bridge to explore the connection between visual feautres and music attributes



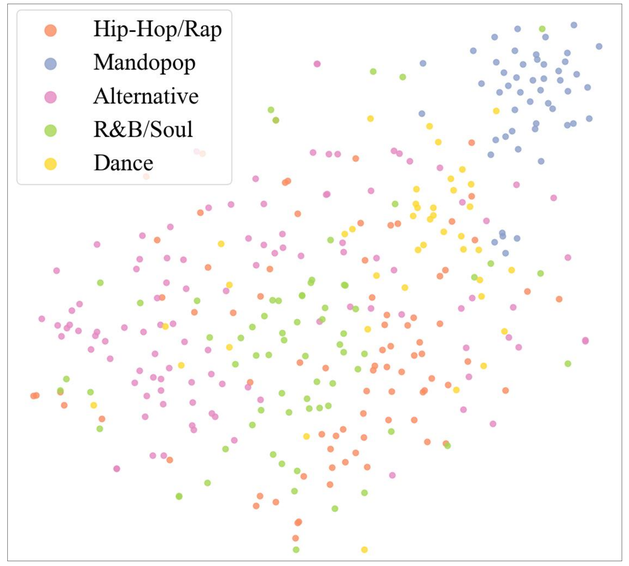
Genre & Chord Distribution

Video Background Music Generation:
Dataset, Method and Evaluation

- Use genre as a bridge to explore the connection between visual feautres and music attributes



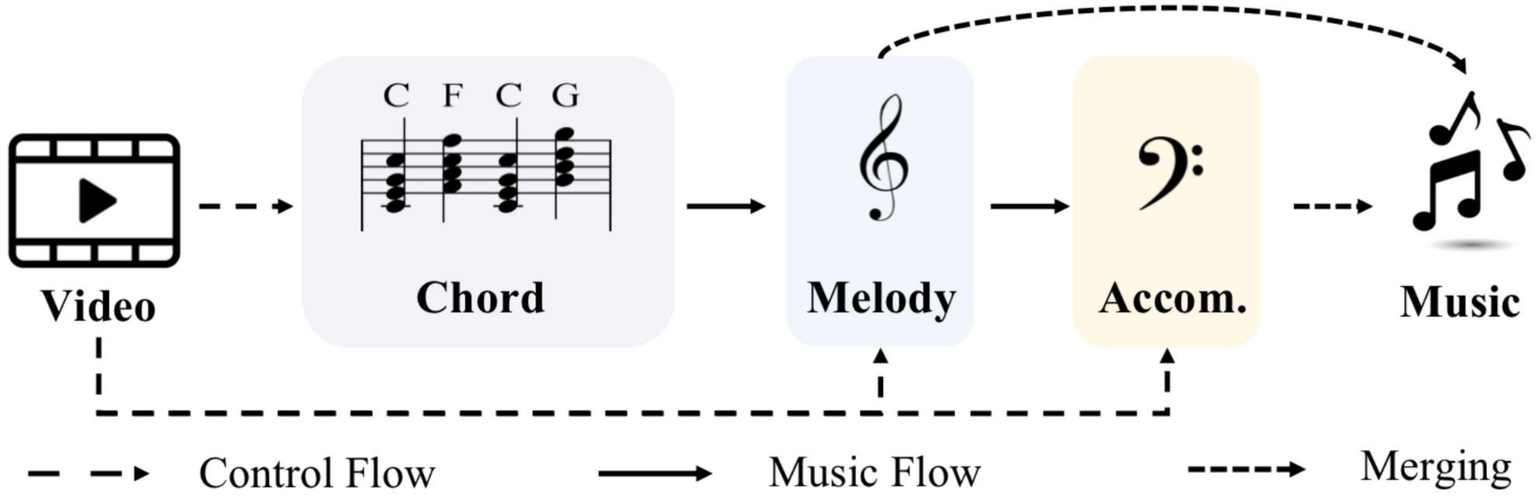
Genre & Rhythm



Genre & CLIP visual features

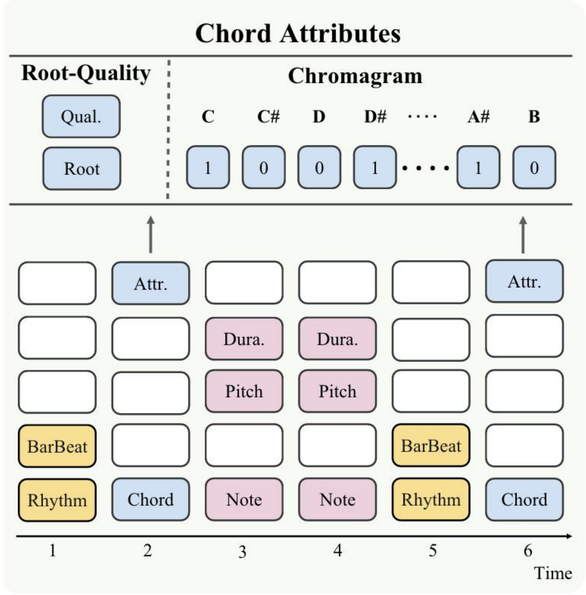
Video Background Music Generation: Dataset, Method and Evaluation

- Disentangle the generation of chord, melody and accompaniment
- Extract multiple video features to control the music generation process



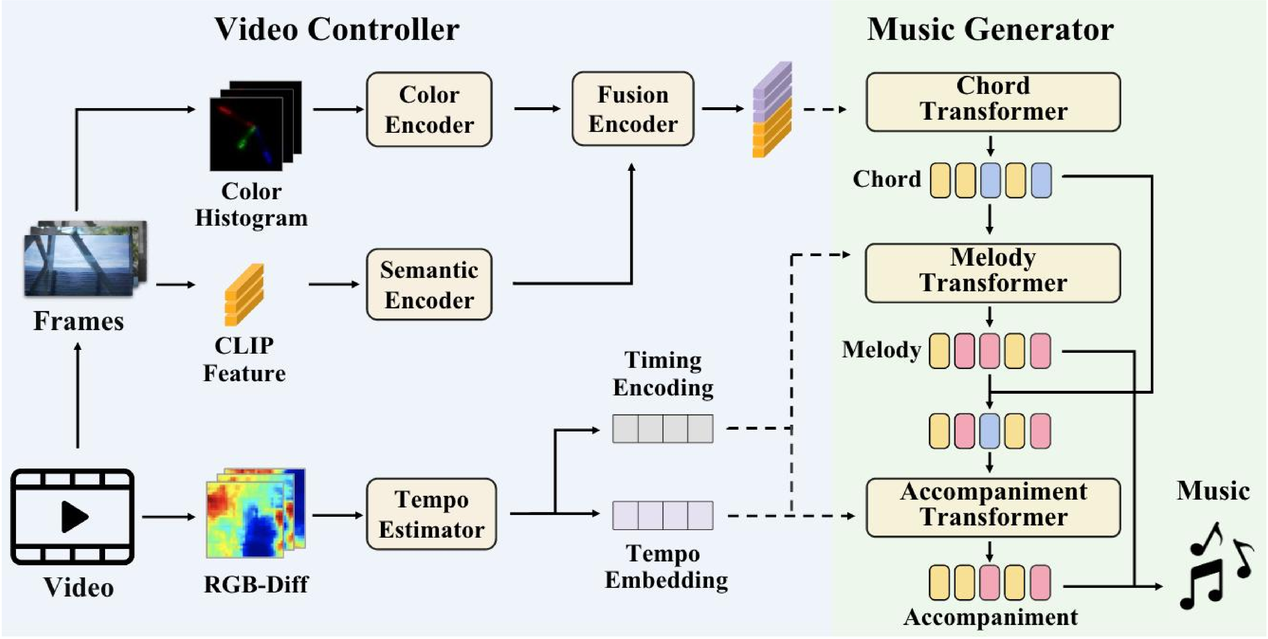
Video Background Music Generation: Dataset, Method and Evaluation

- Further refine music representations



Video Background Music Generation:
Dataset, Method and Evaluation

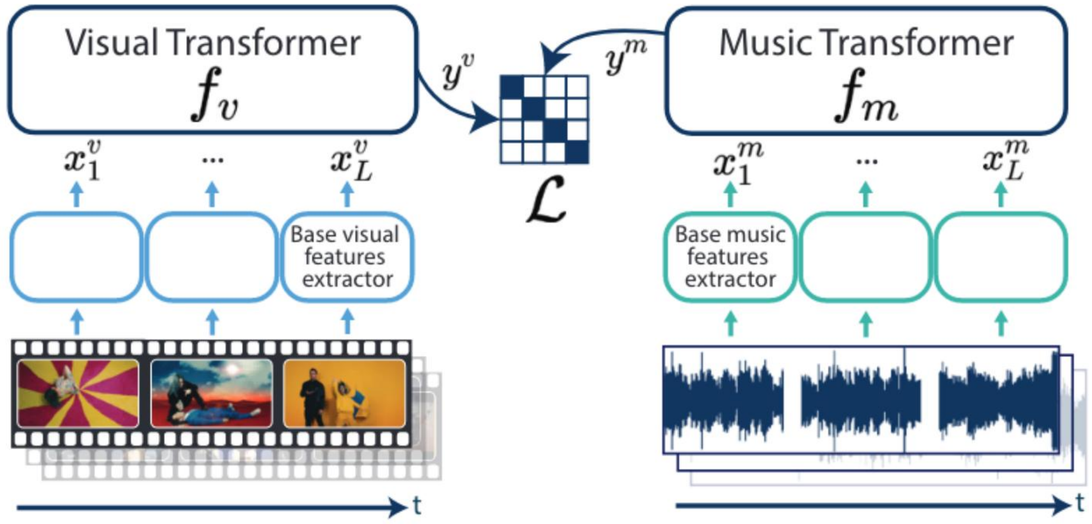
- Unconditional Music Generation:
  Disentangle to generate chord, melody and accompaniment in order
- Video-to-Music Generation:
  Extract color, semantics, rhythm features from videos to control different stages of the model and generate background music



Video Background Music Generation:
Dataset, Method and Evaluation

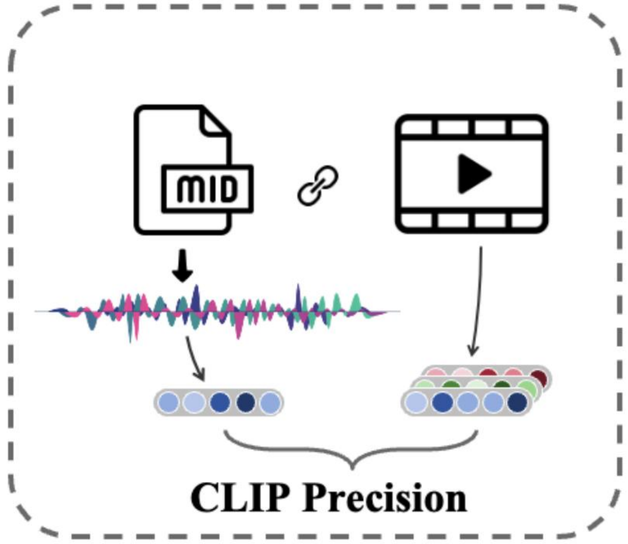
How to evaluate video-music correspondence?

- Subjective evaluation: Bias
- Video-music correspondence is more complex than text-image
- Apply contrastive learning to train music-video CLIP to evaluate video-music correspondence, like CLIP score



Video Background Music Generation:
Dataset, Method and Evaluation

- Use the music-video CLIP as the evaluator model
- Retrieve videos from the dataset with the generated music
- Correctly retrieval indicates strong correspondence
- Adopt retrieval-based metrics: R-Precision and Median Rank



Video Background Music Generation:
Dataset, Method and Evaluation

Methods	Video-Music Correspondence				Music Quality				
	P@5	P@10	P@20	AR	SC	PE	PCE	EBR	IOI
Real (SymMV)	-	-	-	-	0.986	4.197	2.633	0.023	0.184
CMT [9]	8.9	17.7	31.0	33.4	0.990	3.920	2.444	0.074	0.246
w/o semantic	11.6	23.9	42.0	26.1	0.955	2.892	2.310	**0.019**	0.358
w/o color	15.6	**26.6**	**44.8**	**25.1**	0.956	2.732	2.200	0.011	0.330
w/o motion	12.2	22.2	37.9	26.3	0.975	3.010	2.283	0.004	0.261
Video2music	10.8	19.7	33.3	30.0	0.981	**3.990**	**2.639**	0.010	0.229
Video2chord2music	13.7	23.1	43.6	26.0	0.996	2.497	2.036	0.081	0.985
V-MusProd	**15.7**	24.6	**44.8**	25.4	**0.983**	3.940	2.607	0.004	**0.174**

Table 2: **Objective evaluation on SymMV test set.** We evaluate video-music correspondence and music quality with VMCP and music quality metrics. P indicates Precision, where higher is better. AR indicates average rank, where lower is better. For music quality metrics, **closer** to Real is better.

Video Background Music Generation:
Dataset, Method and Evaluation

Metrics	Expert	Non-expert
Music Melody	77%	82%
Music Rhythm	63%	53%
Video Content	63%	63%
Video Rhythm	60%	57%
Chord Quality	63%	-
Accom. Quality	83%	-
Overall Ranking	73%	67%

Table 3: **Subjective evaluation for V-MusProd against CMT [9].** We show preference rates in music quality metrics, video-music correspondence metrics, and expertise metrics.

Methods	SC	PE	PCE	EBR	IOI
Real (POP909)	0.965	4.455	2.774	0.005	0.125
CP [21]	0.987	3.697	2.538	0.041	0.250
Music Trans. [22]	0.985	3.934	2.581	0.034	0.216
HAT [54]	0.989	3.856	2.550	0.040	**0.139**
V-MusProd	**0.967**	**4.070**	**2.774**	**0.005**	0.171

Table 4: **Results of unconditional generation on POP909 [50].** For all the metrics, **closer** to Real is better.

Video Background Music Generation:
Dataset, Method and Evaluation

Demo of Video-to-music Generation:
https://www.wzk.plus/slides/musprod_samples/conditional/processed/V-MusProd_001_processed.mp4

Demo of Unconditional Music Generation:
https://www.wzk.plus/slides/musprod_samples/unconditional/processed/005_processed.mp3

Overview of Video-to-Music

Initial Attempt: CMT (2021)

Advanced Method: MusProd (2023)

Recent Work: VMB (2025)

Discussion on Social Impact

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

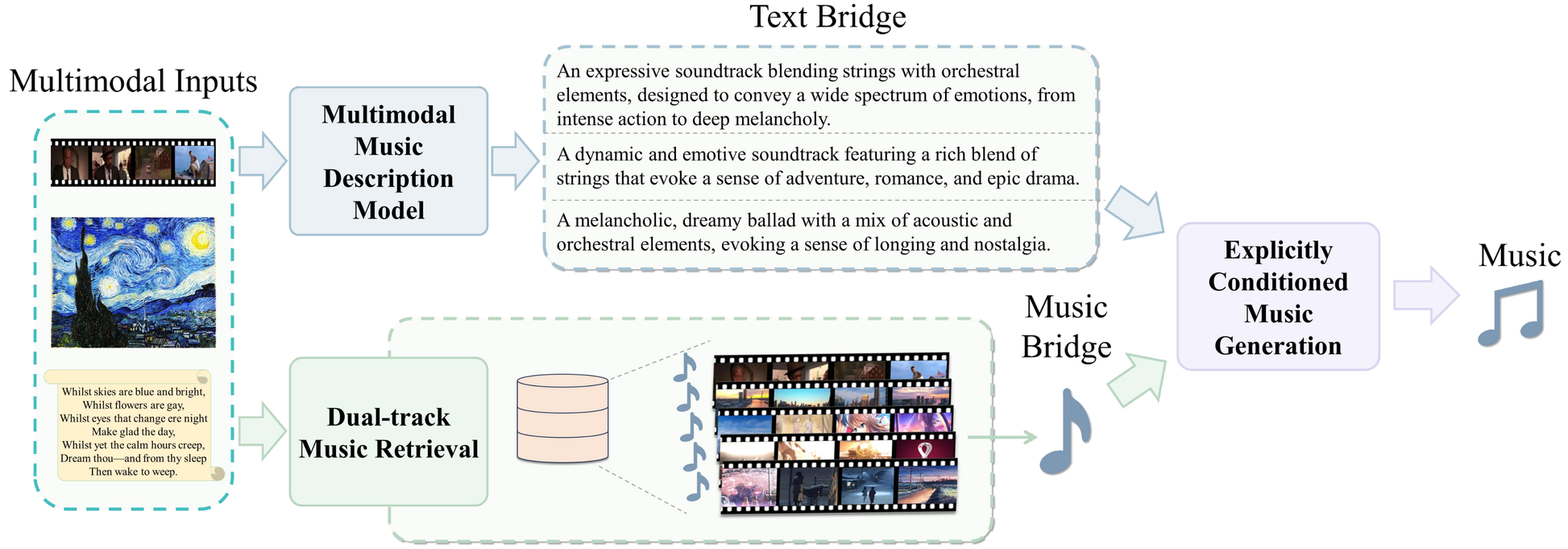
ISMIR 2025 LLM4MA Workshop

Introduction

- Background: Multimodal music generation — creating music from text, images, or videos, with applications in film, games, and XR.

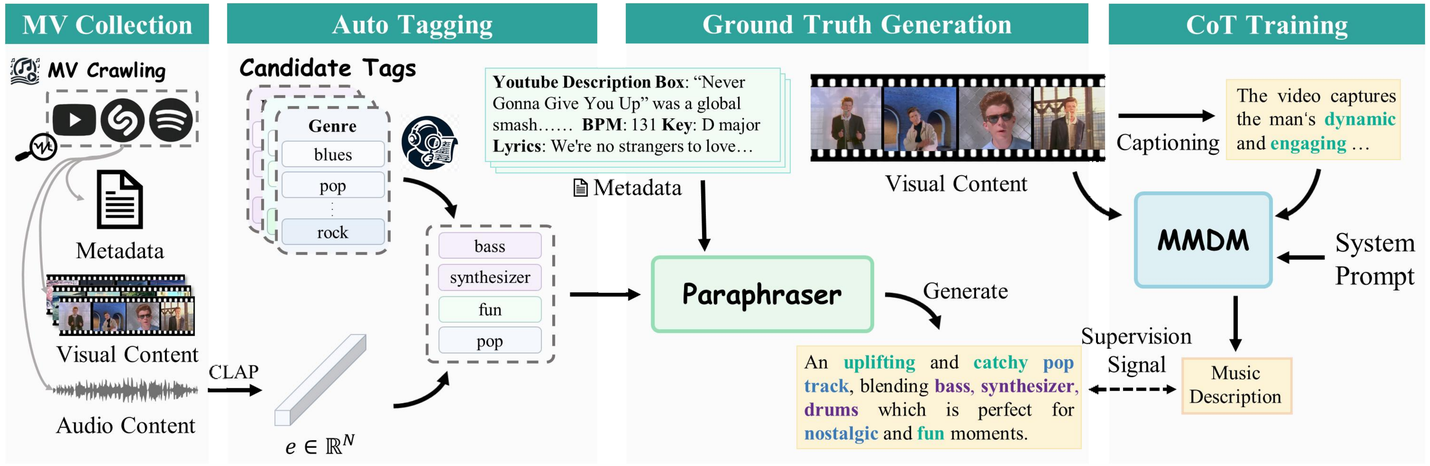
Introduction

- 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

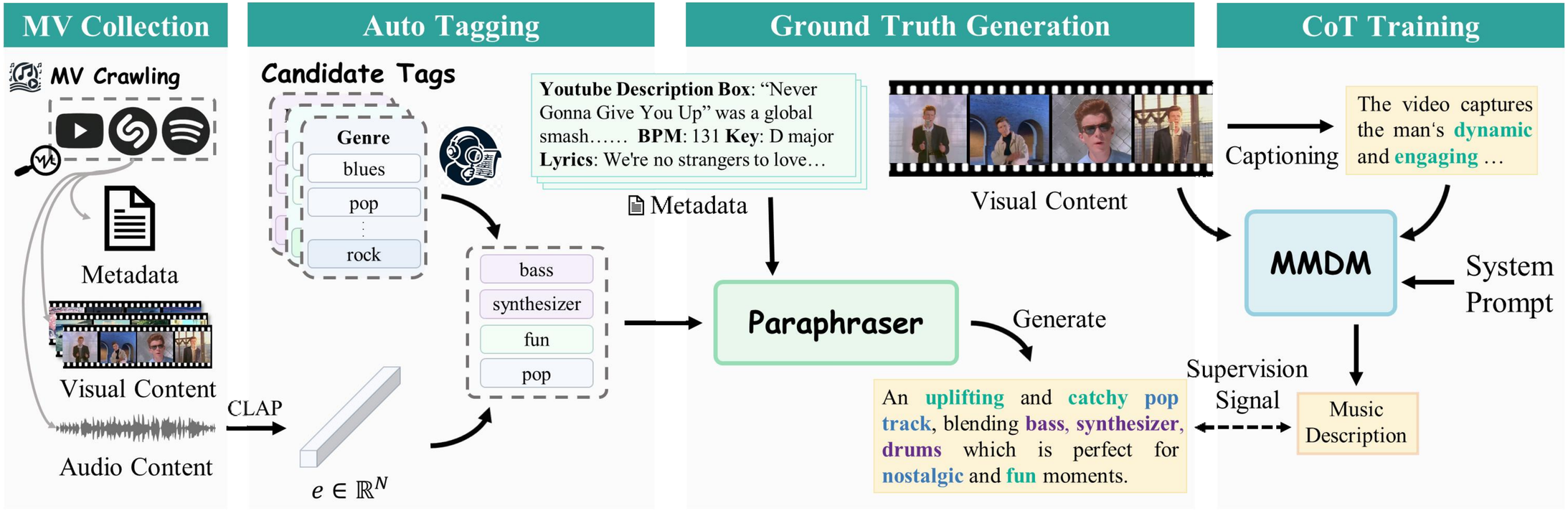
- MTV-24K: A curated video-music dataset with fine-grained alignment, used for training visual-to-music-description.
- MT-512K: A large-scale text-music dataset (500K+ pairs) with rich annotations, forming the foundation for RAG.



Dataset	Genre	Music Description	Source Separation	Music Attr.	Size
HIMV-200K [21]	✗	✗	✗	✗	200K
AIST++ [33]	✓	✗	✗	✗	1,408
TikTok [62]	✗	✗	✗	✗	445
SymMV [63]	✓	✗	✗	✓	1,140
DISCO-MV [31]	✗	✗	✗	✗	2200K
V2M [53]	✗	✗	✗	✗	360K
MUVideo [37]	✗	coarse-grained	✗	✗	14.5K
BGM909 [34]	✓	fine-grained	✓	✓	909
MTV-24K(Ours)	✓	fine-grained	✓	✓	24K

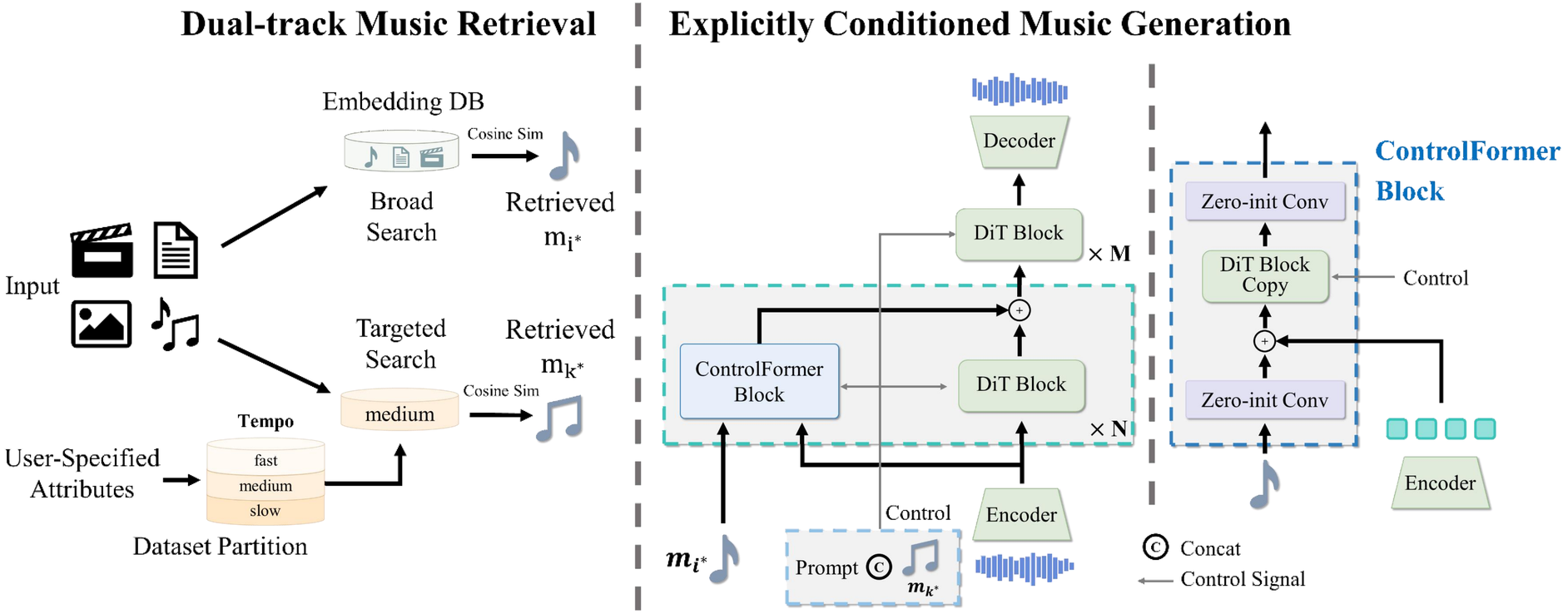
Method

- Text Bridge (MMDM): Translates visual inputs (image/video) into structured music descriptions, serving as the semantic bridge.



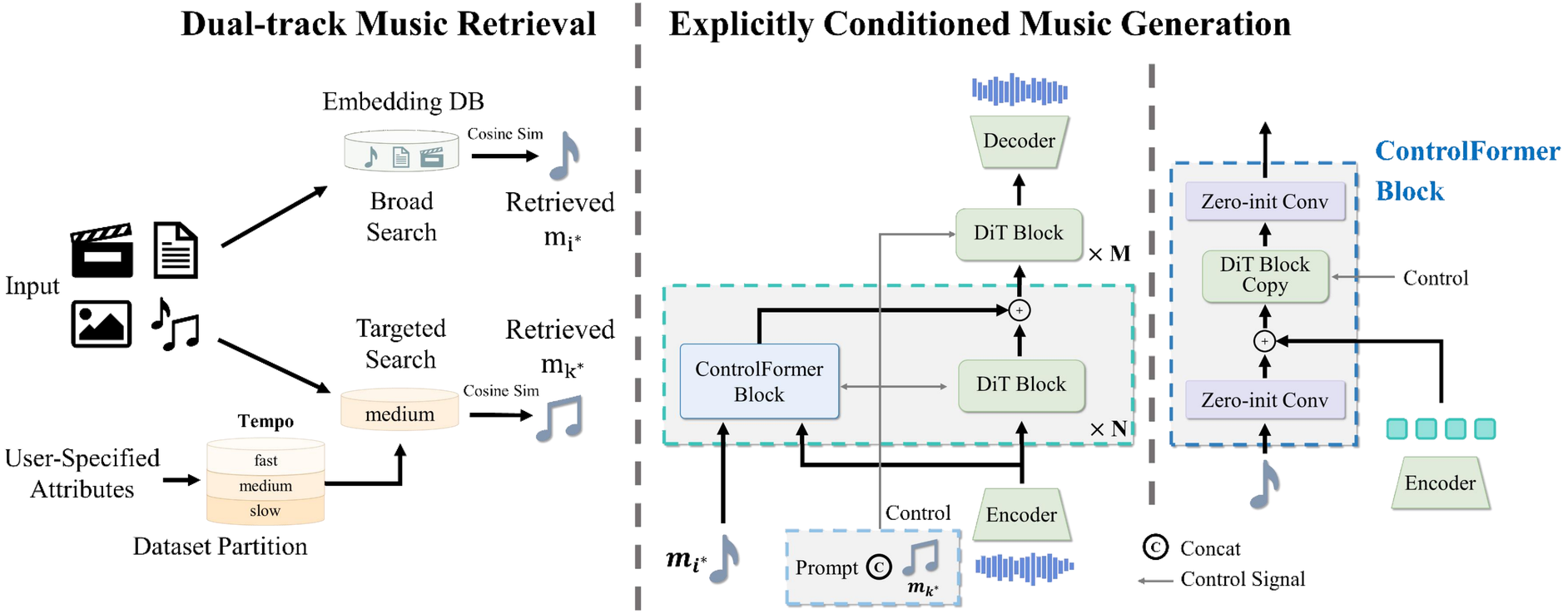
Method

- Music Bridge (Dual-track Retrieval): Broad retrieval provides melody/rhythm reference; targeted retrieval offers controllable attributes like genre, mood, tempo.



Method

- Explicitly Conditioned Music Generation (ECMG): Diffusion Transformer + ControlFormer, integrating both bridges for high-quality and controllable music generation.



Experiments

- Video-to-Music

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@SymMV

Experiments

- Text-to-Music

Method	Objective Metrics				Subjective Metrics↑	
	KL_{passt}↓	FD_{openl3}↓	CLAPScore↑	IB↑	MP	TMA
Stable Audio Open [17]	42.89	183.09	40.92	24.67	3.41	3.52
MusicGen [8]	46.89	181.59	33.95	22.46	3.11	3.35
AudioLDM [33]	99.85	293.86	17.61	20.01	2.34	2.71
M^2UGen [34]	49.03	188.84	28.76	16.70	3.19	3.27
VMB (ours)	37.43	132.16	39.66	29.36	3.78	3.48

Text2Music@SongDescriber

Experiments

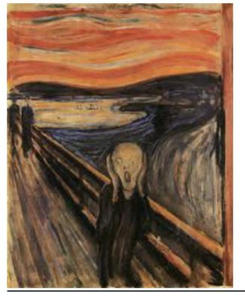
Method	$KL_{passt}\downarrow$	$FD_{openl3}\downarrow$	IB↑
CoDi [50]	216.48	251.52	9.60
M^2UGen [34]	128.33	247.42	2.28
VMB (ours)	105.60	119.76	11.88

Image-to-Music @ MUImage

	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.

Model	CLAPScore
GPT-4V [41]	44.41
InternVL [6]	44.21
MMDM	50.88

Video-to-Description

	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 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.

Experiments

Attribute	Change (Δ)
Instrument	+11.46
Genre	+3.03
Mood	+4.14

Controllability

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

Ablation

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

User Study For Controllability

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

Retrieval V.S. Generation

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

Low Resources

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

Text Modification

Demo

https://wzk1015.github.io/vmb/

Overview of Video-to-Music

Initial Attempt: CMT (2021)

Advanced Method: MusProd (2023)

Recent Work: VMB (2025)

Discussion on Social Impact

Impact on Music Industry

- Video2Music only simulates styles/rhythms, failing to meet film soundtrack demands for emotion, context, and coordination
- Key barriers
- disconnected workflows (lacking collaborative feedback)
- poor cultural context awareness
- cross-professional communication
- sound balance
- Need to evolve from "substitute" of human composers to **"intelligent collaborator"**
- collaboration rather than one-time generation
- inspire and assist

Function of AI Music

- AI-generated music **lacks genuine emotion** and **subjective expression**
- Suitable: **functional scenarios**, e.g. short video background music, restaurant/street background (functional pop)
- Not Suitable: indie music (rock, folk, etc.), demanding human creativity and ideas
- Film soundtrack is a middle ground: serves the work but allows composers' subtle personal expression
- The same goes for AI writing: we appreciate AI-written manuals/summaries (functional), but not AI-written novels/poetry (require human sentiment)
- AI can serve as a **creative assistant**, not a replacement for human expression in non-functional music

Thanks for listening!

Email: wangzhaokai@sjtu.edu.cn

Check out our repo: