香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Deep Music: Understanding, Generation, and Interaction

Presenter: Zijian Zhao

Ph.D. Student in Civil Engineering (Scientific Computation)

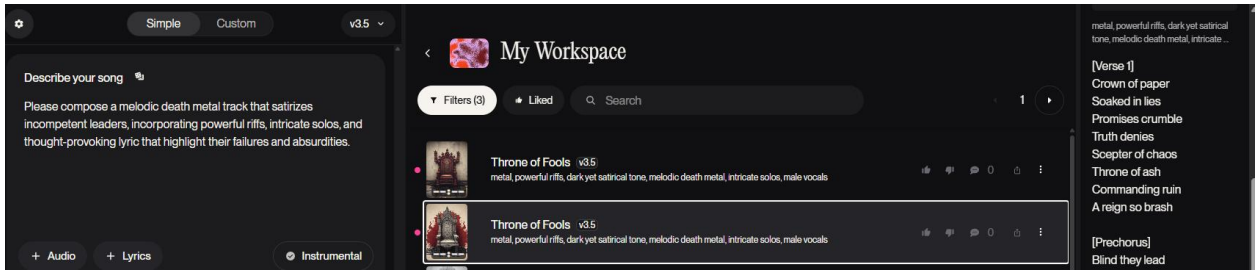
Department of Civil and Environmental Engineering

The Hong Kong University of Science and Technology

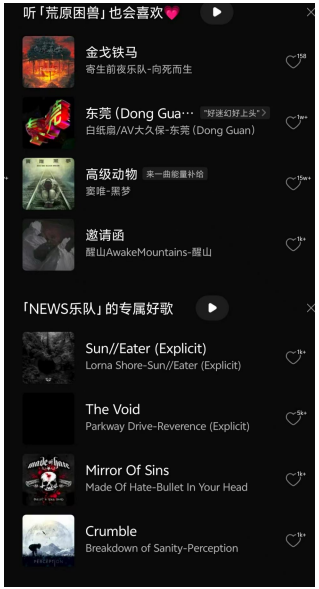
Part I: Background

Background: Music Information Retrieval (MIR)

- **What is MIR?**
  - MIR is the study and application of techniques to extract, analyze, and manage information from music. It encompasses audio analysis, music classification, recommendation systems, and content-based retrieval, aiming to improve how we search for and interact with music. (in short: computer + music)
- **Applications of MIR:**
  - Music Recommendation System
  - Music Generation Models
  - Music and Lycris Analysis, Music Education, ...



SUNO Music Generator



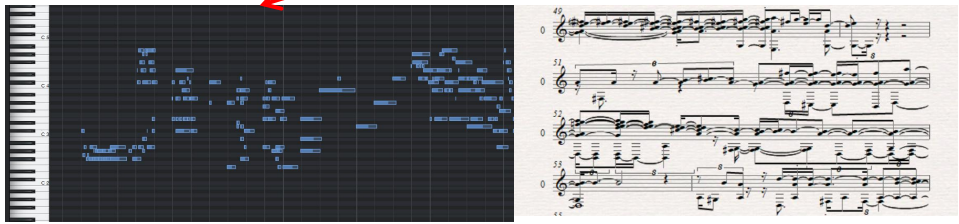
QQ Music Recommender

Background: Music Representation in MIR

- **Audio Music**
  - Original File (for computer programs):
    - Mp3
    - Wav
  - Representation Manners (for neural networks):
    - Tokenization
    - Continous Embedding
- **Symbolic Music**
  - Original File:
    - MIDI
    - MusicXML
    - ABC
  - Representation Manners:
    - CP [1]
    - REMI [2]
    - Octuple [3]

Audio Music

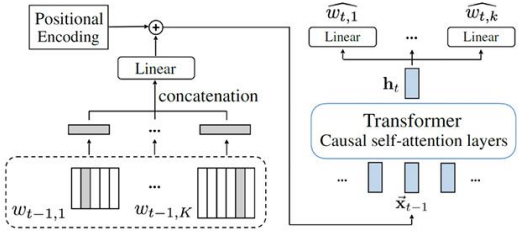
Symbolic Music



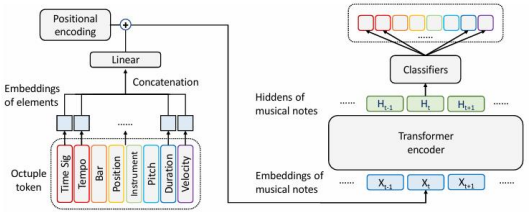
Background: Natural Language Models in MIR

-- “Music is the universal language of mankind.”

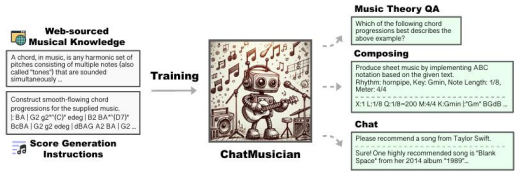
- **Representative MIR Models based on Nautral Language Models**
  - Early Stage
    - Music Transformer [4]
    - POP Music Transformer [2]
    - CP Transformer [1]
  - Pre-training Period
    - MidiBERT [5]
    - MusicBERT [3][6]
    - MuseBERT [7]
    - MERT [8]
  - LLM Era
    - ChatMusician [9]
    - YuE [10]
    - WavJourney [11]



CP Transformer



MusicBERT (symbolic)



ChatMusician

Research Questions

- **Music Understanding**
  - Sequence-Level Tasks
    - composer prediction
    - emotion classification
  - Token-Level Tasks
    - melody extraction
    - velocity prediction
- **Music Generation**
  - Music Continuation
- **Multi-Modal Music Interaction**
  - X to Music
  - Music to X

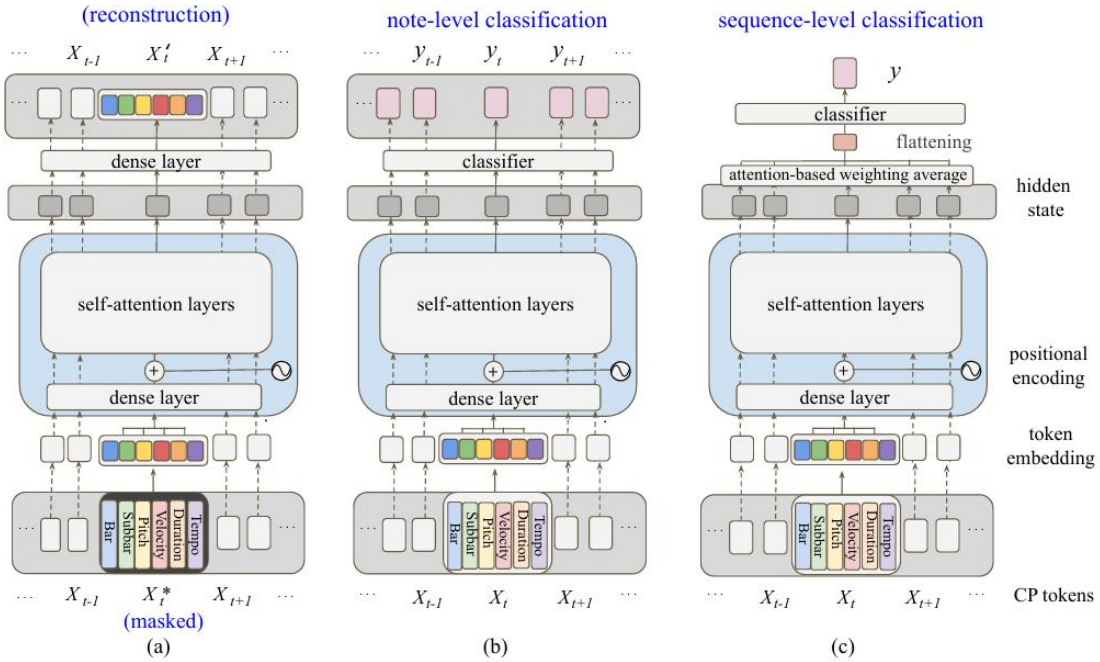
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Part II: Symbolicy Music Understanding Adversarial-MidiBERT

[1] **Zijian Zhao***, "Let Network Decide What to Learn: Symbolic Music Understanding Model Based on Large-scale Adversarial Pre-training", 2025 ACM International Conference on Multimedia Retrieval (ICMR), 2025

Preliminary: MidiBERT

- **Pre-training**
  - Mask Language Model (MLM)
- **Fine-tuning**
  - train specific classification heads



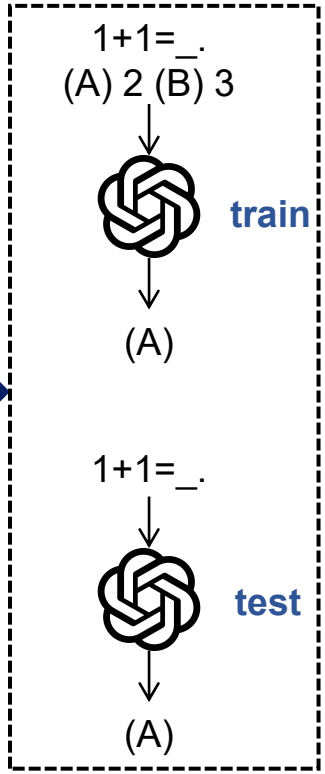
Problem of Conventional MLM: BIAS

- **Reason of Bias**
  - Not all tokens can be recovered from context.

Original Sentence	He	is	a	teacher	.	<EOS>
Bad MASK	He	is	a	<MASK>	.	<EOS>
Good MASK	He	is	<MASK>	teacher	.	<EOS>
Wrong Recovery	He	is	a	student	.	<EOS>
Correct Recovery	He	is	a	teacher	.	<EOS>

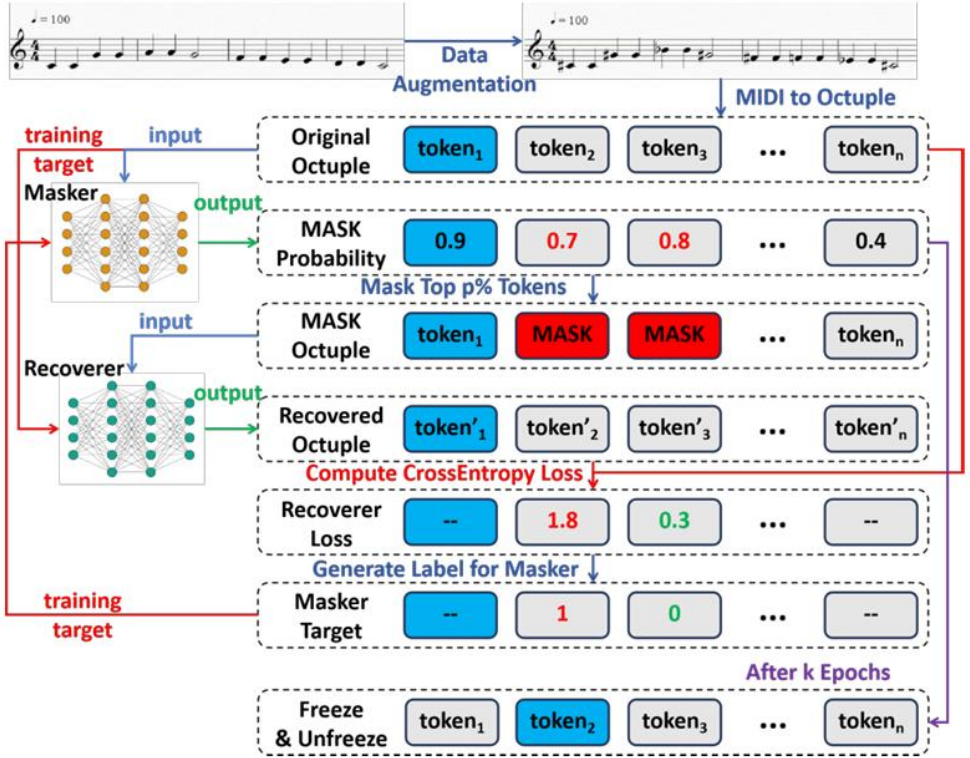
Take natural language for an example.

Potential Result:
overfit to training set



We aim for the model to learn common musical structures, relationships, and regularities—such as basic modes, riff patterns, and modulation techniques—rather than developing habits specific to particular composers.

Workflow



- **Recoverer**
  - Attempts to recover masked tokens.
- **Masker**
  - Aims to mask tokens that are challenging for the Recoverer to accurately recover.
- **Freeze Mechanism**
  - Regularly freezes tokens with the highest probability of being selected by the Masker.

- **Intuition:** The most challenging tokens to recover are context-free tokens. The Recoverer can only infer these tokens based on the distribution of the training set.

Experiment

Table 2: Dataset Decription

Dataset	Pieces	Task	Task Level	Class Number	Used in Pre-training
ASAP [4]	1068	–	–	–	✓
Pop1K7 [7]	1747	–	–	–	✓
Pianist8 [9]	865	Composer Classification	Sequence Level	8	✓
EMOPIA [8]	1078	Emotion Recognition	Sequence Level	4	✓
POP909 [19]	909	Melody Extraction	Token Level	3	✓
GiantMIDI [10]	10855	Velocity Prediction	Token Level	6	×

Table 3: Model Performance in Different Tasks: The bold and underlined value indicates the best and second best result within each task. (Note that the training times provided are for reference only, as the server is shared with other users.)

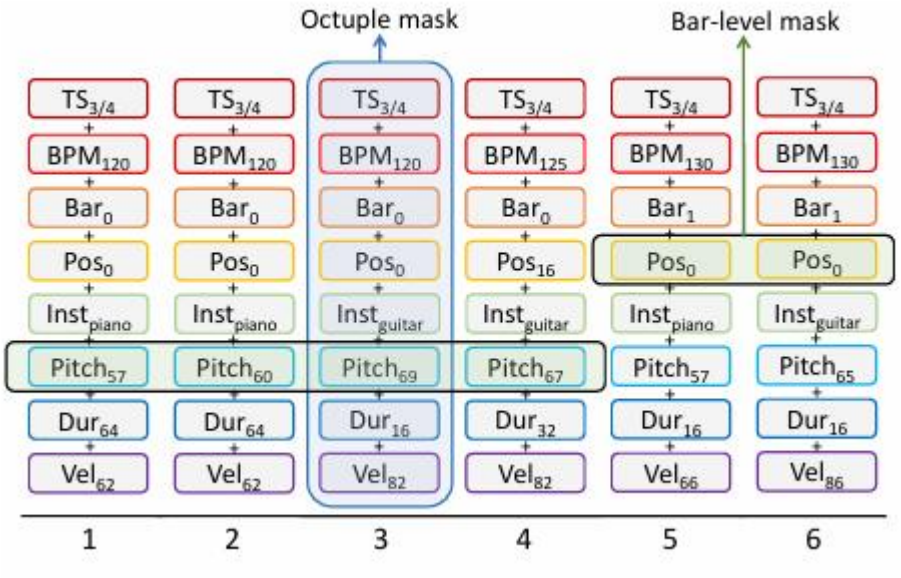
Model	Pre-train			Sequence-Level Classification		Token-Level Classification	
	Accuracy	Epochs	Time	Composer	Emotion	Velocity	Melody
MidiBERT [2]	79.60%	500	6.44d	79.07%	67.59%	44.88%	92.53%
MusicBERT-QM [16]	80.57%	500	9.47d	83.72%	69.52%	<u>46.71%</u>	<u>92.64%</u>
MusicBERT [20]	76.01%	500	10.06d	86.05%	71.06%	38.79%	92.47%
PianoBART [13]	**96.67%**	**268**	**3.19d**	88.37%	73.15%	**49.37%**	92.62%
Adversarial-MidiBERT (ours)	81.47%	<u>436</u>	9.82d	**97.92%**	**79.46%**	45.58%	**92.68%**
Adversarial-MidiBERT (fine-tune w/o mask)	81.47%	<u>436</u>	9.82d	65.98%	70.53%	45.30%	92.55%
Adversarial-MidiBERT (pre-train w/o adversary)	<u>83.51%</u>	500	<u>5.93d</u>	<u>88.91%</u>	<u>74.07%</u>	45.44%	<u>92.64%</u>
Adversarial-MidiBERT (w/o pre-train)	–	–	–	79.76%	68.75%	38.70%	87.98%

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Part III: Symbolicy Music Generation PianoBART

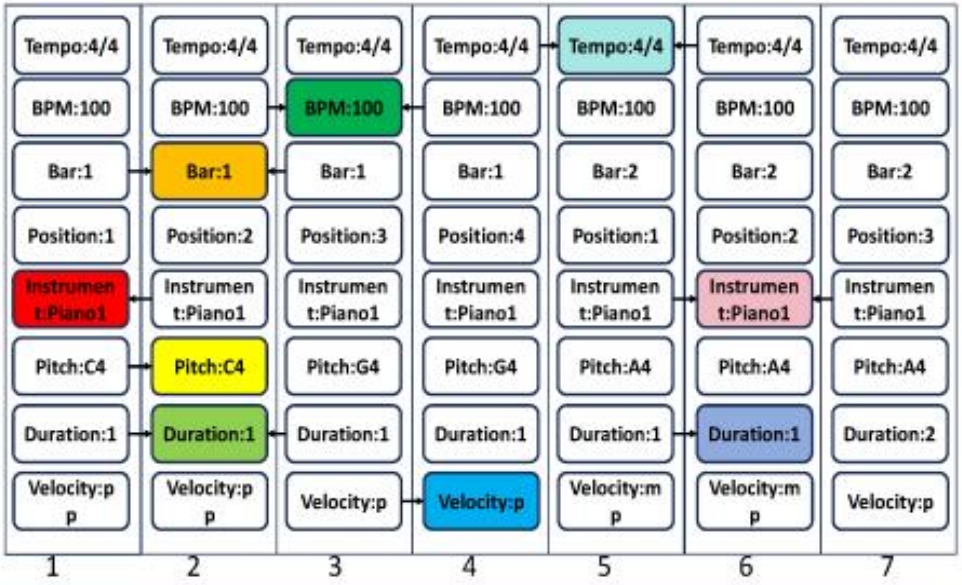
[2] Xiao Liang, **Zijian Zhao**, Weichao Zeng, Yutong He, Fupeng He, Yiyi Wang, Chengying Gao*, "PianoBART: Symbolic Piano Music Understanding and Generating with Large-Scale Pre-Training", 2024 IEEE Conference on Multimedia Expo (ICME), 2024

Preliminary: MusicBERT



- **Otuple Music Representation Manner**
  - Time signature
  - Tempo
  - Bar
  - Position
  - Instrument
  - Pitch
  - Duration
  - Velocity

Problem of Music MLM: INFORMATION LEAKAGE



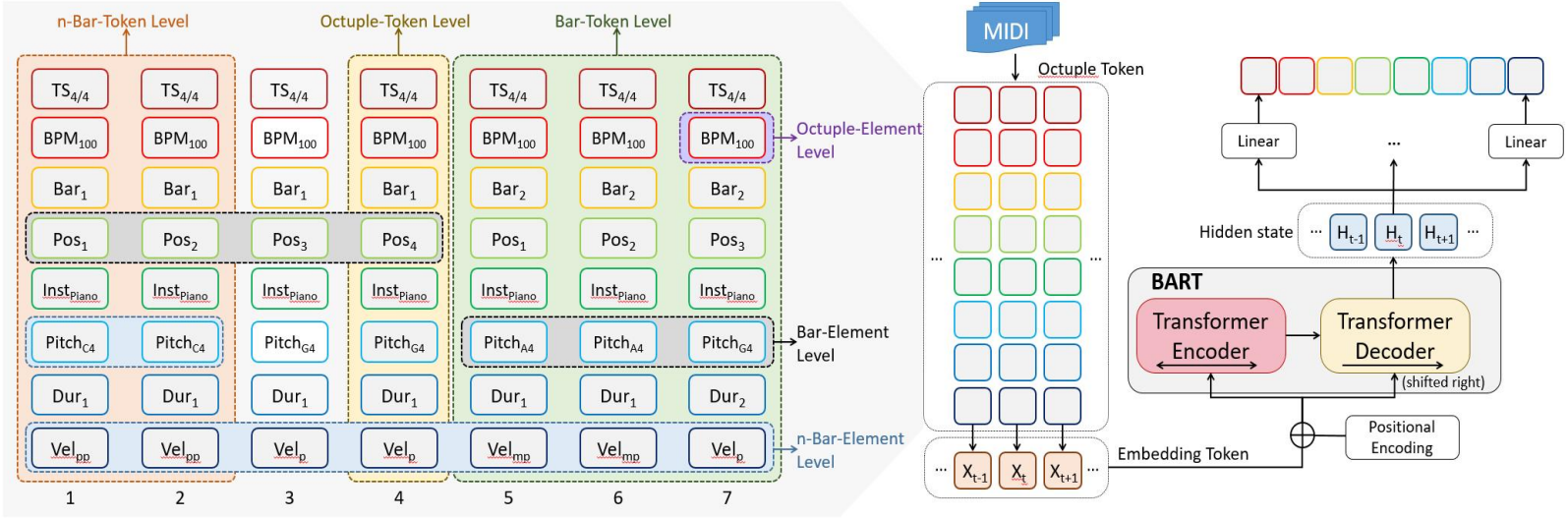
- **Finding**
  - There are many repetitive attributes in successive tokens of music.
- **Problem**
  - In vanilla MLM, the model can simply copy the missing token from neighboring tokens. With this strategy, even if the model learns nothing, it can still achieve relatively high accuracy.

Model Structure

TABLE I
MULTI-LEVEL OBJECT SELECTION STRATEGY.

	Element Level	Token Level
Octuple Level	Octuple-Element Level	Octuple-Token Level
Bar Level	Bar-Element Level	Bar-Token Level
n-Bar Level	n-Bar-Element Level	n-Bar-Token Level

- **Pre-training Tasks of BART**
  - Token Masking (MLM)
  - Token Deletion
  - Text Infilling
  - Sentence Permutation
  - Document Rotation



Experiment

TABLE II
PRETRAINING PERFORMANCE.

Model	Time	Epoch	Parameter	Accuracy
MusicBERT [1]	10.06 d	500	114 million	76.01%
MidiBERT [2]	6.44 d	500	111 million	79.60%
PianoBART	**3.19 d**	**268**	225 million	**96.67%**

TABLE V
THE TESTING CLASSIFICATION ACCURACY OF DIFFERENT MODELS ON FOUR MUSIC UNDERSTANDING TASK (VELOCITY PREDICTION, MELODY EXTRACTION, EMOTION CLASSIFICATION AND COMPOSER CLASSIFICATION). THE BEST ACCURACY IS ACHIEVED BY PIANOBART IN ALL THESE TASKS.

	Token-level Tasks		Sequence-level Tasks		
Model	**Velocity**	**Melody**	**Emotion**	**Composer (Pianoist8)**	**Composer (ASAP)**
MusicBERT [1]	51.23%	92.47%	71.06%	86.05%	94.27%
MidiBERT [2]	48.57%	92.53%	67.59%	79.07%	96.18%
PianoBART (w/o pre-training)	38.55%	82.40%	58.33%	69.77%	78.34%
PianoBART-simple	51.57%	92.50%	66.67%	83.72%	96.32%
PianoBART	**51.63%**	**92.62%**	**73.15%**	**88.37%**	**97.45%**

Experiment

TABLE III
RESULTS OF MUSIC CONTINUATION ON MAESTRO [19].

Model	$PFS_{GT}\uparrow$	$PFS_{prompt}\uparrow$	PCHE ↓	GS ↓
Music Transformer [3]	0.1721	0.1903	0.496	0.140
Pop Music Transformer [5]	0.7742	0.7647	0.360	0.015
PianoBART (w/o pre-train)	0.1502	0.1349	0.237	0.133
PianoBART-simple	**0.8495**	0.8427	0.253	0.006
PianoBART	0.8245	**0.8666**	**0.213**	**0.001**

TABLE IV
RESULTS OF MUSIC CONTINUATION ON GIANTMIDI [20].

Model	$PFS_{GT}\uparrow$	$PFS_{prompt}\uparrow$	PCHE ↓	GS ↓
Music Transformer [3]	0.1137	0.0867	0.476	**0.018**
Pop Music Transformer [5]	0.5762	0.5640	0.378	0.019
PianoBART (w/o pre-train)	0.1793	0.1658	**0.055**	0.165
PianoBART-simple	0.7334	0.6984	0.508	0.083
PianoBART	**0.7708**	**0.7354**	0.224	0.071

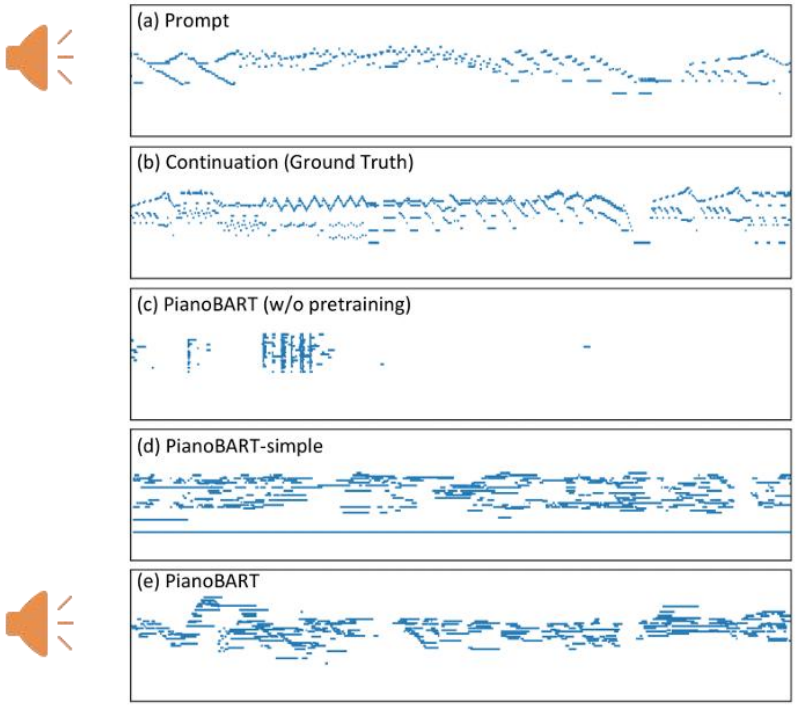


Fig. 2. Visualization results of generated examples on ablation variants. PianoBART (w/o pretraining), PianoBART-simple, and PianoBART are all continued from Prompt.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Part IV: Stage Light Generation from Music
Skip-BART

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

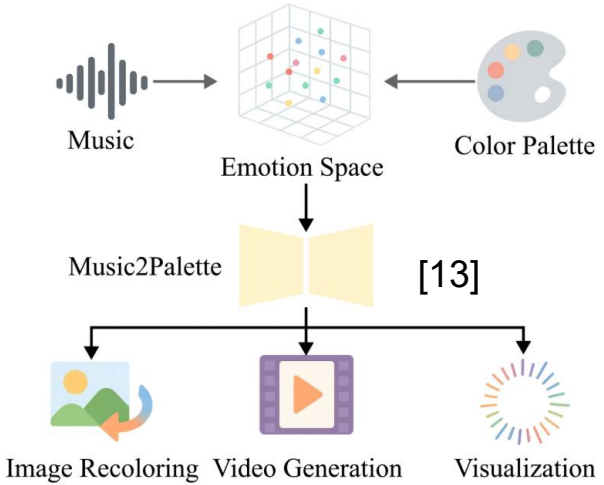
香 港 大 學 THE UNIVERSITY OF HONG KONG

CityU 香港城市大學 City University of Hong Kong

[3] **Zijian Zhao**, Dian Jin, Zijing Zhou, Xiaoyu Zhang*, "Automatic Stage Lighting Control: Is it a Rule-Driven Process or a Generative Task?" (under anonymous review, available in AriXiv 2025)

Previous Works

- **Rule-based Light Control**
  - Emotion
  - Style
  - Topic
  - Lycris
  - Chord
  - .....



- **Challenges & Shortcomings**
  - Corse-grained Classification -- Can the model effectively distinguish between:
    - C, Cm, Csus4, C7, Cm7, C9, ...
    - Atmosphere Black Metal, Melodic Black Metal, Depressive Black Metal, Folk Black Metal, ...
  - Low Accuracy
    - The effectiveness of rule-based methods is heavily reliant on accurate classification.
  - Low Interpretability [12]
    - The rationale behind the mappings is often unclear.

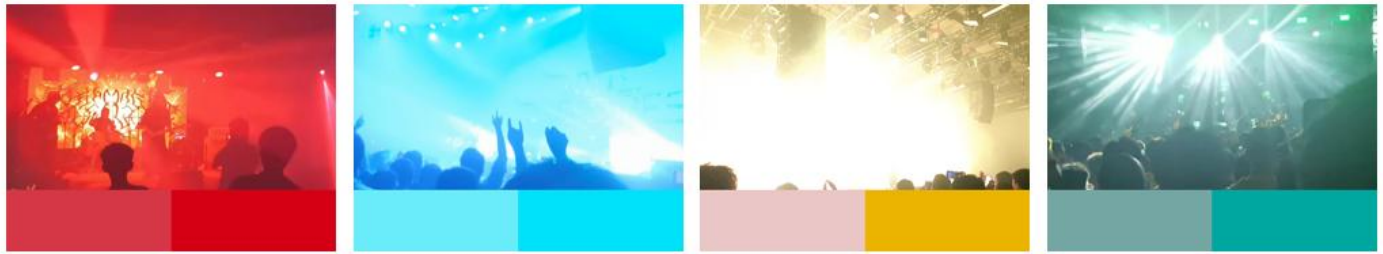
Dataset & Data Process

Code Link: https://github.com/RS2002/Skip-BART

Dataset Link: https://huggingface.co/datasets/RS2002/RPMC-L2

My Band Link: https://tokamak-disruption.netlify.app

- **End-to-End is all you need!**
- **Dataset**
  - Introducing the first stage lighting dataset: Rock, Punk, Metal, and Core - Livehouse Lighting (RPMC-L2).
  - Contains 699 videos collected from 2020 to 2024 in Guangzhou and Shenzhen.
  - To avoid copyright issues, only processed feature data is released.
- **Data Process**
  - Extract hue and value for each frame from the Hue, Saturation, and Value (HSV) color space.

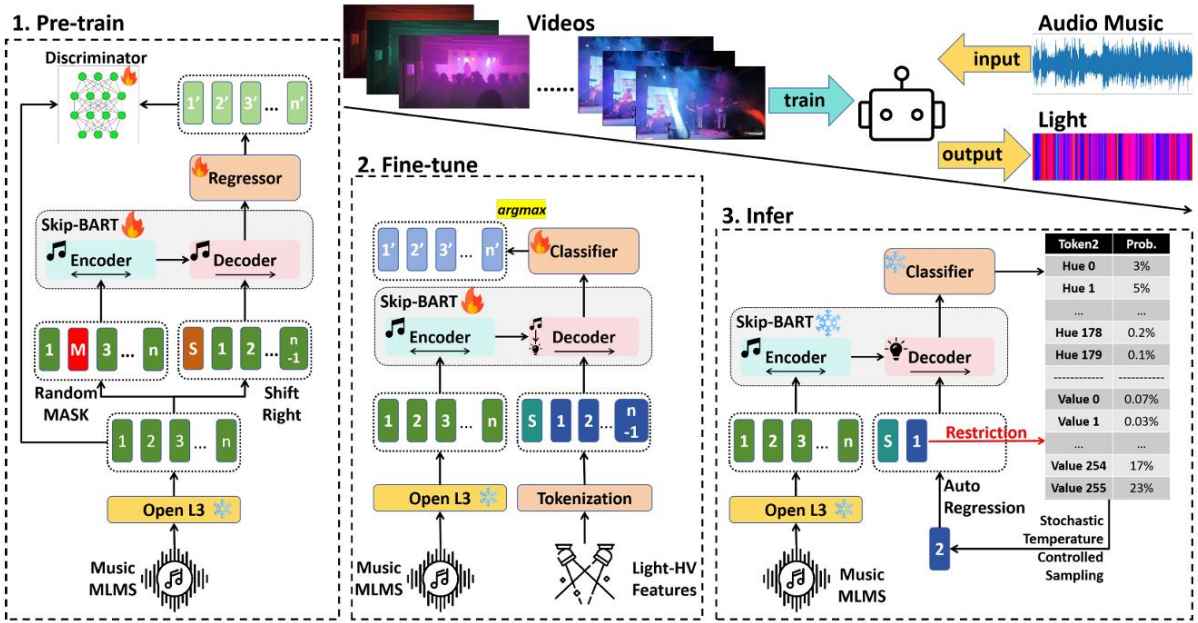


(a) Normal Warm Light (b) Normal Cold Light (c) Extreme Warm Light (d) Extreme Cold light

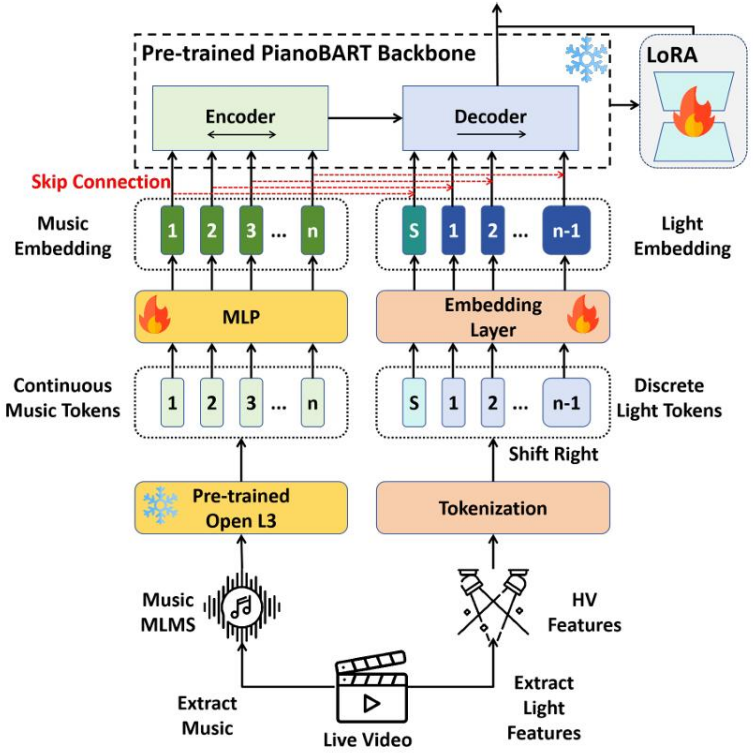
Figure 2: Each subfigure shows the input scene on the top, the result of the direct extraction method in the bottom-left, and our extraction method in the bottom-right. (a)-(b) Both methods accurately extract the dominant hue. (c)-(d) Our method extracts colors that are closer to the original appearance.

Workflow

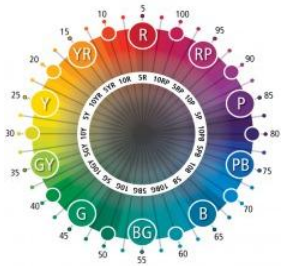
- **Pre-training**
  - pre-trained parameters from PianoBRT + music-only MLM pre-training
- **Fine-tuning**
- **Inference**
  - restricted stochastic temperature sampling



Model Structure



- **Music Embedding**
  - OpenL3
- **Light Embedding**
  - Discrete Embedding Layer
  - Why discretizing?
    - The Hue space is circular.



- **Skip-Connection Mechanism**
  - Connect each frame of music and light before inputting to the decoder.
- **Backbone**
  - Frozen PianoBART
  - Trainable LoRA

Quantitative Evaluation

- **Objects**
  - Groun Truth
  - Rule-based Method (based on emotion-mapping)
  - Skip-BART
  - Ablation Study (Skip-BART w/o skip-connection)

Table 1: Quantitative Results: The best result is indicated in **bold**, and the second best is indicated by underline. This notation will remain consistent in the following tables.

Methods	RMSE↓		MAE↓		$\text{corr}(\lvert\Delta\rvert)(\times 10^{-2})\uparrow$	
	Hue	Value	Hue	Value	Hue	Value
Rule-based	48.67	93.39	43.43	86.55	0.50	0.58
Skip-BART	**36.13**	**60.74**	**28.72**	**51.27**	0.88	**2.94**
from scratch	36.63	67.49	28.83	57.22	0.69	0.53
w/o skip connection	36.89	68.33	29.44	58.34	**1.15**	0.30
w/o light embedding	51.04	67.25	41.50	54.87	0.80	0.70

Human Evaluation

Table 2: Human Evaluation Scores: The six metric scores and overall evaluations of the four objects are expressed as Mean ± Standard Deviation (M ± SD) . The bold text represents the best result, while the underlined text indicates the second-best result.

Method	Emotion	Impact	Rhythm	Smoothness	Atmosphere	Surprise	Overall
Ground Truth	4.50±0.93	**4.48±0.99**	**4.61±0.99**	**4.62±1.07**	**4.49±0.89**	**4.34±1.10**	**4.51±0.88**
Skip-BART	**4.69±0.87**	4.39±0.95	4.50±1.06	4.32±1.12	4.32±0.93	3.83±1.06	4.35±0.87
Ablation Study	4.31±0.94	3.78±0.96	4.54±1.08	4.43±1.12	4.11±0.98	3.50±1.00	4.11±0.84
Rule-based	3.12±1.52	2.65±1.39	2.54±1.47	2.56±1.27	2.77±1.50	2.35±1.40	2.67±1.29

Table 6: Median and Plurality Scores: The bold text represents the best result, while the underlined text indicates the second-best result. For the plurality, we retain the highest score when multiple pluralities exist.

	Emotion	Impact	Rhythm	Smoothness	Atmosphere	Surprise	Overall
				Median			
Ground Truth	4.33	**4.67**	**4.67**	**4.67**	**4.67**	**4.50**	**4.58**
Skip-BART	**4.67**	4.50	4.50	4.33	4.33	4.17	4.42
Ablation Study	4.33	3.83	**4.67**	4.33	4.17	3.33	4.11
Rule-based	3.17	2.33	2.17	2.50	2.17	1.67	2.33
				Plurality			
Ground Truth	4.33	**5.33**	4.67	**5.00**	5.00	**5.00**	**5.17**
Skip-BART	**4.67**	3.67	**5.33**	4.33	**5.67**	4.33	4.72
Ablation Study	**4.67**	4.00	4.67	4.33	5.00	4.67	4.61
Rule-based	4.00	1.00	1.00	1.00	1.00	1.00	1.00

Human Evaluation

- Our method yields a p-value of 0.724 in a statistical comparison based on human evaluations with human lighting engineers, suggesting that the proposed approach does not significantly differ from human lighting engineering performance.

Table 3: Pairwise Comparisons of Overall Scores for the Four Objects: The symbols * and ** denote that the significance level (p) for the difference in means (ΔM) is less than 0.05 or 0.01, respectively.

Comparison	ΔM	SD	p
Ground Truth vs. Skip-BART	0.16	0.10	0.724
Ground Truth vs. Ablation Study	0.40	0.10	**0.003****
Ground Truth vs. Rule-based	1.84	0.21	**< 0.001****
Skip-BART vs. Ablation Study	0.23	0.10	0.152
Skip-BART vs. Rule-based	1.68	0.19	**< 0.001****
Ablation Study vs. Rule-based	1.44	0.16	**< 0.001****

Table 7: Pairwise Comparisons among Ground Truth (GT), Skip-BART (SB), Ablation Study (AS), and Rule-Based (RB): The symbols * and * denote that the significance level (p) for the difference in means (ΔM) is less than 0.05 or 0.001, respectively.**

Metrics	Comparison	ΔM	SD	p	Metrics	Comparison	ΔM	SD	p
Emotion	GT vs SB	-0.19	0.15	1.000	Impact	GT vs SB	0.09	0.13	1.000
	GT vs AS	0.19	0.11	0.598		GT vs AS	0.70	0.15	< 0.001***
	GT vs RB	1.38	0.25	< 0.001***		GT vs RB	1.83	0.23	< 0.001***
	SB vs AS	0.39	0.14	0.045*		SB vs AS	0.61	0.14	0.001***
	SB vs RB	1.57	0.23	< 0.001***		SB vs RB	1.75	0.23	< 0.001***
	AS vs RB	1.18	0.22	< 0.001***		AS vs RB	1.13	0.18	< 0.001***
Rhythm	GT vs SB	0.11	0.12	1.000	Smoothness	GT vs SB	0.30	0.14	0.234
	GT vs AS	0.08	0.14	1.000		GT vs AS	0.19	0.16	1.000
	GT vs RB	2.07	0.26	< 0.001***		GT vs RB	2.06	0.23	< 0.001***
	SB vs AS	-0.04	0.12	1.000		SB vs AS	-0.11	0.16	1.000
	SB vs RB	1.96	0.21	< 0.001***		SB vs RB	1.76	0.23	< 0.001***
	AS vs RB	1.99	0.25	< 0.001***		AS vs RB	1.87	0.20	< 0.001***
Atmosphere	GT vs SB	0.17	0.14	1.000	Surprise	GT vs SB	0.51	0.16	0.014*
	GT vs AS	0.38	0.15	0.088		GT vs AS	0.84	0.16	< 0.001***
	GT vs RB	1.72	0.25	< 0.001***		GT vs RB	1.99	0.23	< 0.001***
	SB vs AS	0.21	0.13	0.619		SB vs AS	0.33	0.17	0.318
	SB vs RB	1.55	0.21	< 0.001***		SB vs RB	1.48	0.21	< 0.001***
	AS vs RB	1.34	0.18	< 0.001***		AS vs RB	1.15	0.17	< 0.001***

Visualization Result

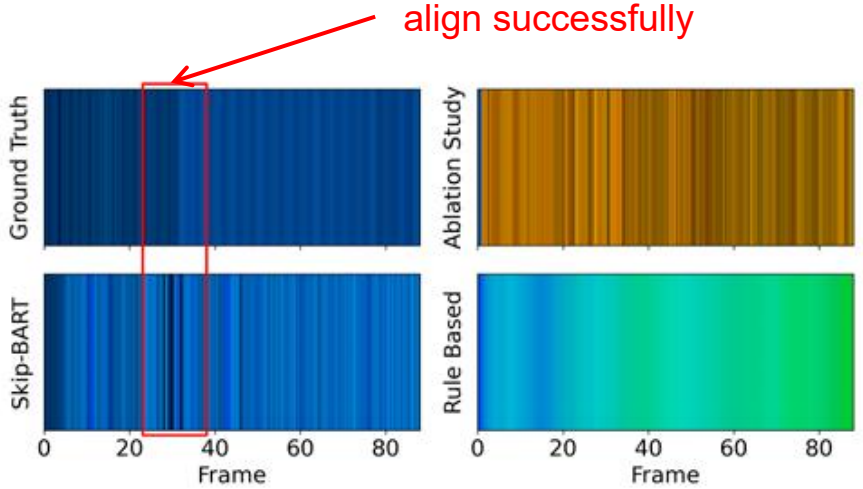


Figure 4: Visualization of lighting sequences generated by different methods.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Part V: Music Generation from Iamge
A Multi-Modal RAG Assisted VLM Framework

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

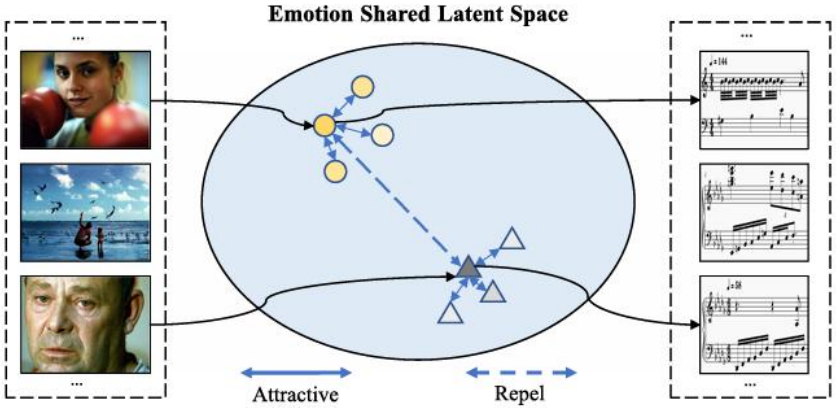
香 港 大 學 THE UNIVERSITY OF HONG KONG

[4] **Zijian Zhao***, Dian Jin, Zijing Zhou, "Zero-Effort Image-to-Music Generation: An Interpretable RAG-based VLM Approach" (under way, to be submitted to IEEE Signal Processing Letters (SPL))

Previous Works

- **Training-Based Methods**
  - Large Computation Resource Requirement
  - Dataset Limitations
  - Low Interpretability
- **Mapping-Based Methods**
  - Lack of Rationality
  - Low Interpretability

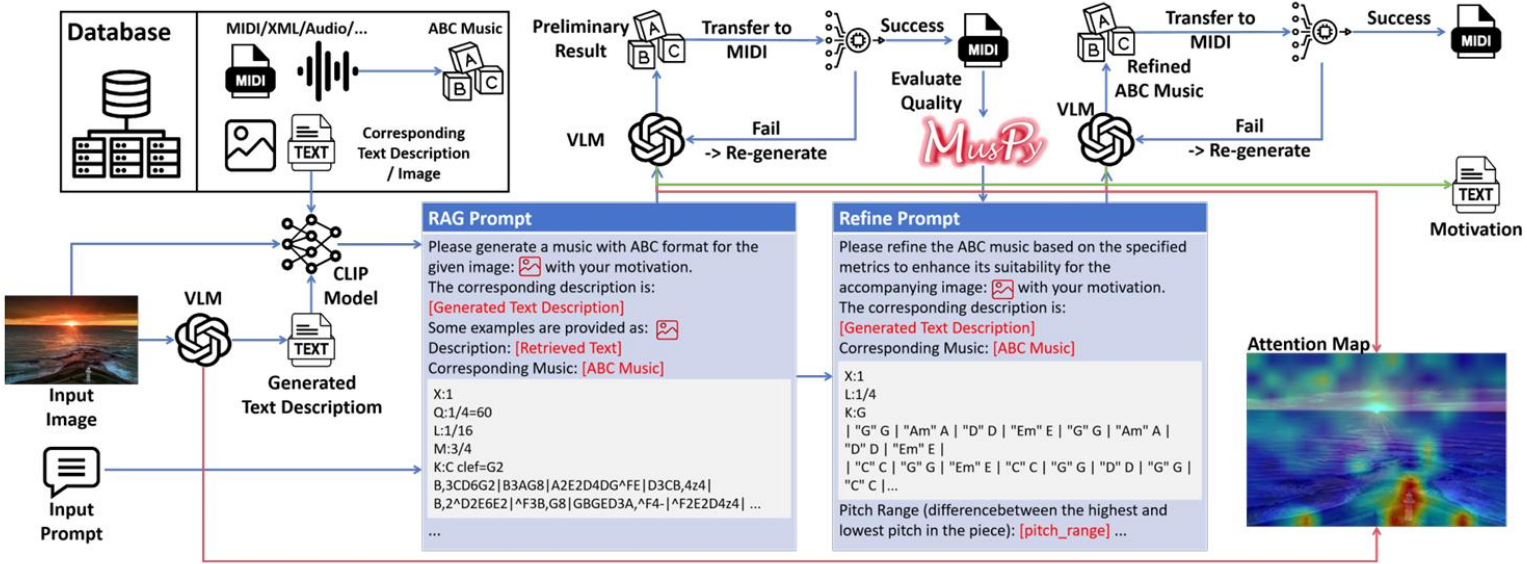
- **Our Method**
  - High Interpretability
  - Low Cost
    - No Need for Training
  - High Quality
    - RAG-Based Generation
    - Self-Refinement Mechanism



[14]

Workflow

- **Primary Music Generation**
  - Generate Image Description
  - Generate ABC Music via Multi-Modal RAG
- **Music Refinement Using a Model-Based Evaluator**
  - Self-Refine According to Music Quality Metrics
- **Explanation Generation with Text Output and Image Attention Map**



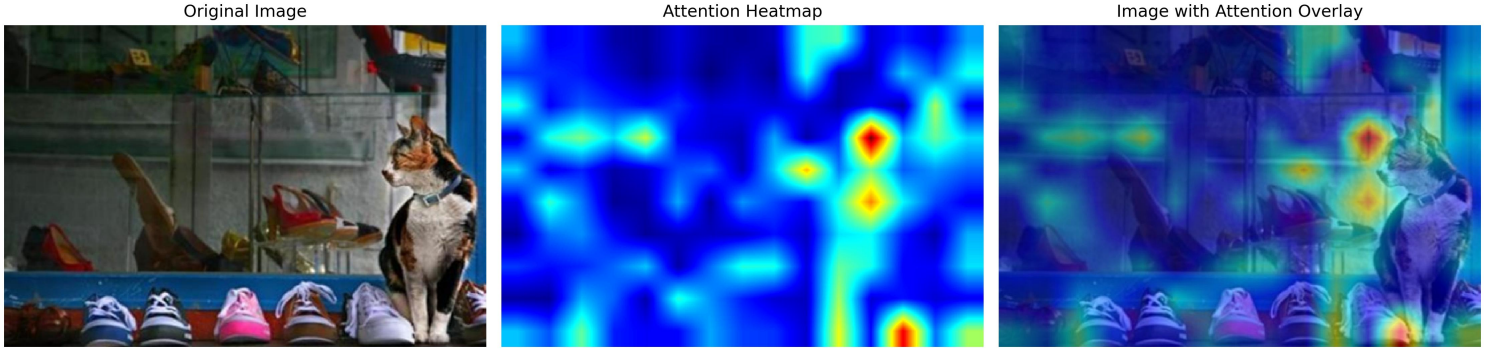
Machine Evaluation

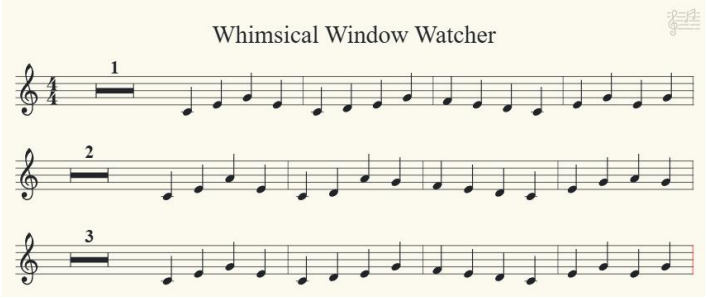
- **LM as Judger**
  - Generation VLM
    - Open Source SOTA: Keye-VL-8B [15]
  - Evaluation VLM
    - Latest Closed Source Model: Grok4 [16]

TABLE II
MACHINE EVALUATION RESULT

Methods \ Metrics	Music Quality						Music-Image Consistency			
	Overall	Melody	Rhythm	Authenticity	Harmony	Average	Overall	Semantics	Emotion	Average
Synesthesia [6]	3.0	2.4	3.6	2.8	2.2	2.8	3.8	2.8	3.8	3.5
Mozart's Touch [34]	4.4	3.8	4.6	4.6	4.2	4.3	5.4	4.0	5.6	5.0
Vanilla Keye [24]	4.8	4.4	4.6	5.0	4.4	4.6	5.4	4.8	5.4	5.2
Ours	**5.2**	**5.2**	**5.2**	**5.2**	**5.0**	**5.2**	**6.0**	**5.2**	**6.0**	**5.7**
w/o RAG	4.8	4.4	4.4	4.8	4.4	4.6	5.6	5.0	5.4	5.3
w/o Refinement	**5.2**	5.0	4.6	5.0	**5.0**	5.0	5.8	5.0	5.8	5.5

Generation Result





The music for "Whimsical Window Watcher" aims to evoke the serene curiosity of a calico cat observing a vibrant shoe display. The gentle 4/4 meter and C major key establish a calm, playful foundation, while the flowing melody (inspired by the cat's graceful movements) weaves through the scene. Light dynamics (mf to crescendo) mirror the cat's growing fascination, and simple ornaments (trills) add a whimsical touch, reflecting the vibrant yet peaceful atmosphere of the shop. The repeated motifs and soft resolution capture the cat's lingering gaze, blending the tranquil interior with the lively colors of the shoes.

Part VI: Future Directions

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Thanks for your attention!

Presented by Zijian Zhao

Email: zzhaock@connect.ust.hk

Homepage: https://zijianzhao.netlify.app