声学大讲堂
音频产业创新技术公益讲座
助力引领区建设 科创中国浦东行

文本驱动的大模型可控音乐生成与编辑

张逸霄 伦敦玛丽女王大学

2024.9.2

张逸霄

研究方向

可控音乐生成、跨模态音乐编辑，以及大语言模型在音乐中的应用

教育经历

- 2020-2024，伦敦玛丽女王大学，博士生
  - 导师：Prof. Simon Dixon & Dr. Mark Levy (Apple Inc.)
- 2015-2019，电子科技大学，工学学士

实习经历：

- 2024.5-至今，Stability AI
- 2023.10-2024.5，Sony AI Tokyo
- 2023.6-2023.9, Yamaha R&D
- 2019.9-2020.10, NYU Shanghai

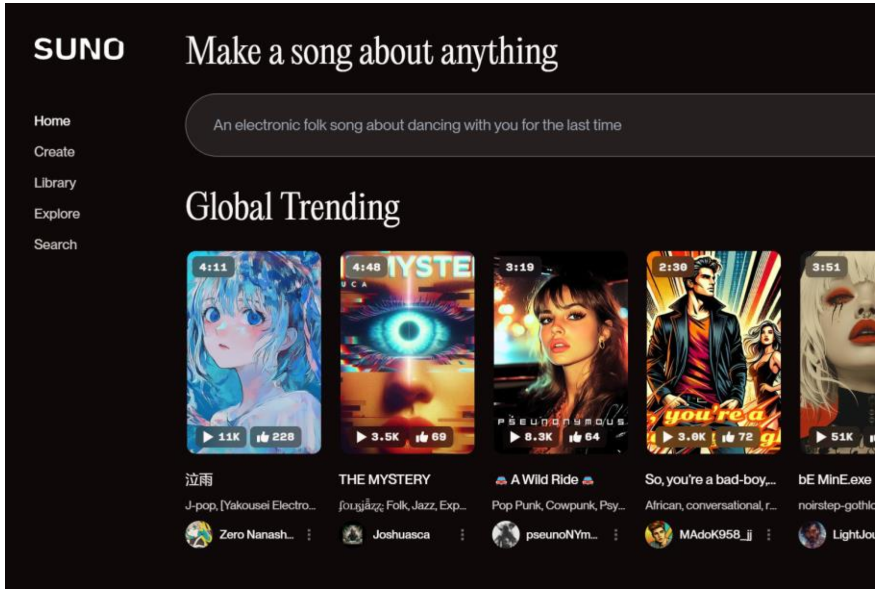
目录

- 背景
  - 文本到音乐生成：任务描述
  - 符号还是音频：两类表示法
  - 大模型时代的两种范式
  - 为什么需要增强模型的控制能力？
  - 四种可行的思路
  - 预训练模型：MusicGen
- 音乐大语言模型的内容控制
  - **Content-based Controls For Music Large Language Modeling**, ISMIR 2024
  - *Liwei Lin, Gus Xia, Junyan Jiang, Yixiao Zhang*
- 编配、内绘和润色：通过基于内容的控制来指导长期音乐音频生成和编辑
  - **Arrange, Inpaint, and Refine: Steerable Long-term Music Audio Generation and Editing via Content-based Controls**, IJCAI 2024
  - *Liwei Lin, Gus Xia, Yixiao Zhang, Junyan Jiang*
- 指令微调的音乐大模型：多任务音乐编辑
  - **Instruct-MusicGen: Unlocking Text-to-Music Editing for Music Language Models via Instruction Tuning**, arXiv/2405.18386
  - *Yixiao Zhang, Yukara Ikemiya, Naoki Murata, Woosung Choi, Marco Martínez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, Simon Dixon*
- 音乐大模型为什么仍然不能像人类一样合作？
  - **The Interpretation Gap in Text-to-Music Generation Models**, NLP4MusA Workshop @ ISMIR 2024
  - *Yongyi Zang*, Yixiao Zhang**
- Q&A

背景

文本到音乐生成：任务描述

- 给定音乐的文本描述，模型输出符合描述的音乐片段



Make a song about anything

An electronic folk song about dancing with you for the last time

Global Trending

泣雨
J-pop, [Yakousei Electro...

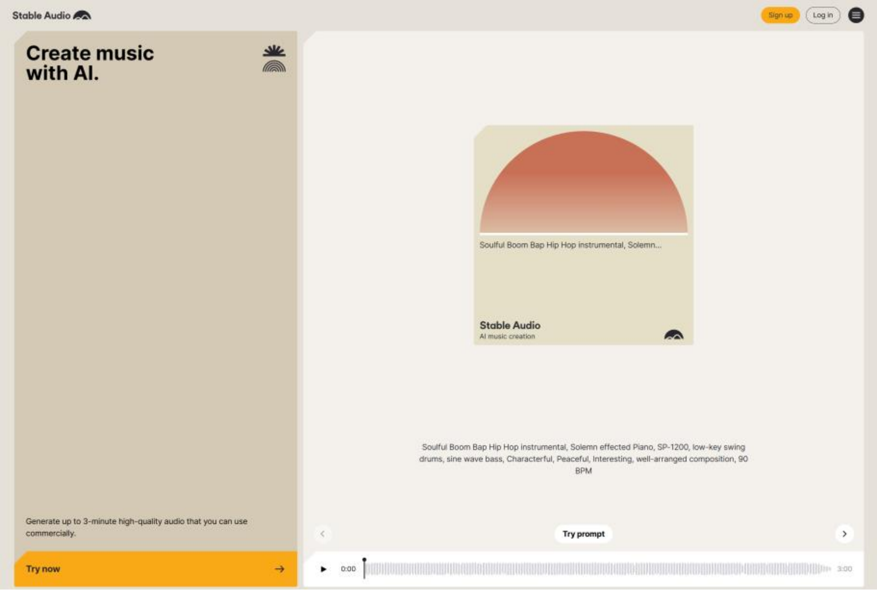
THE MYSTERY
ʆoʊʂjåʐʐ; Folk, Jazz, Exp...

A Wild Ride
Pop Punk, Cowpunk, Psy...

So, you're a bad-boy,...
African, conversational, r...

bE MinE.exe
noirstep-gothic

Suno AI



Create music with AI.

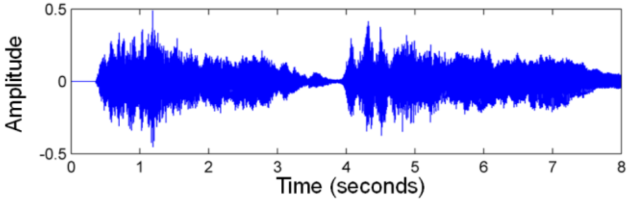
Soulful Boom Bap Hip Hop instrumental, Solemn...

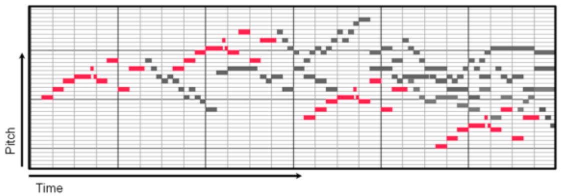
Soulful Boom Bap Hip Hop instrumental, Solemn effected Piano, SP-1200, low-key swing drums, sine wave bass, Characterful, Peaceful, Interesting, well-arranged composition, 90 BPM

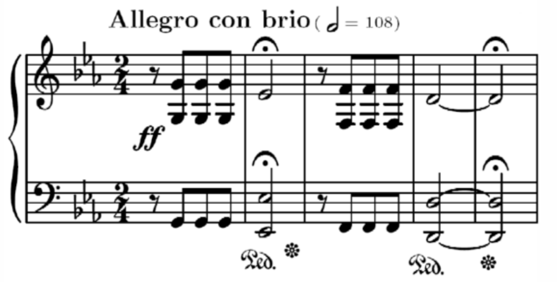
Generate up to 3-minute high-quality audio that you can use commercially.

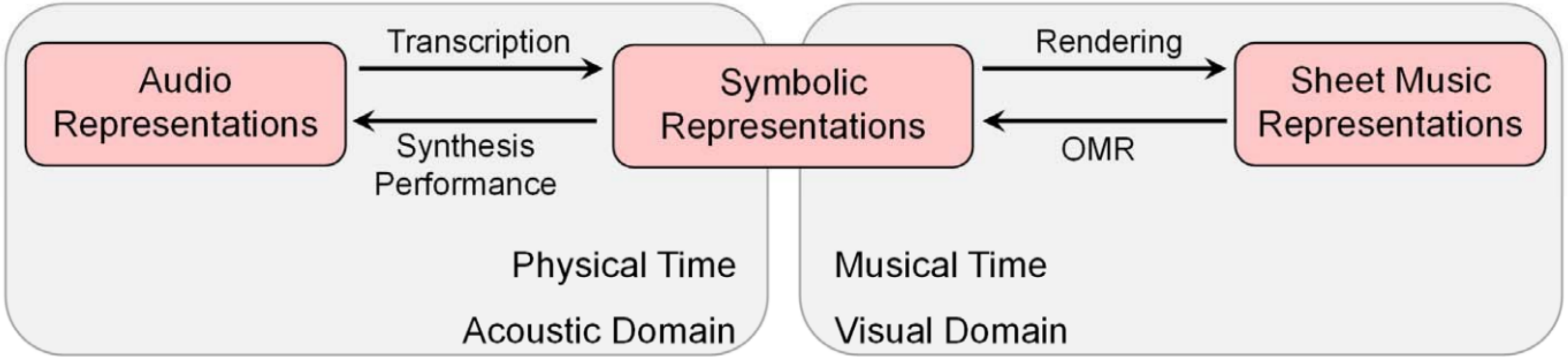
Stable Audio 2.0

符号还是音频：两类音乐表示法









Audio Representations

Transcription

Synthesis Performance

Symbolic Representations

Rendering

OMR

Sheet Music Representations

Physical Time

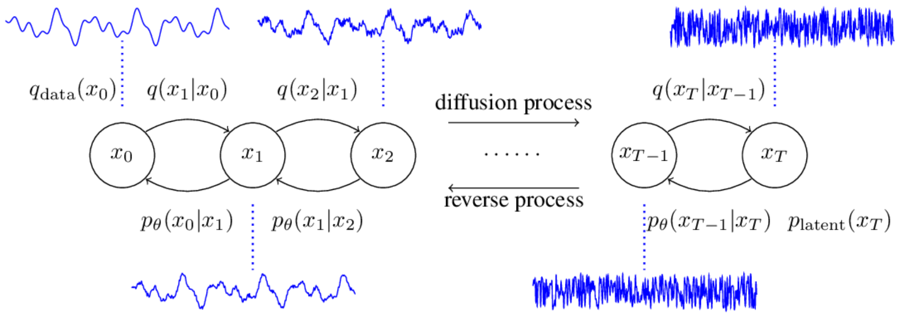
Acoustic Domain

Musical Time

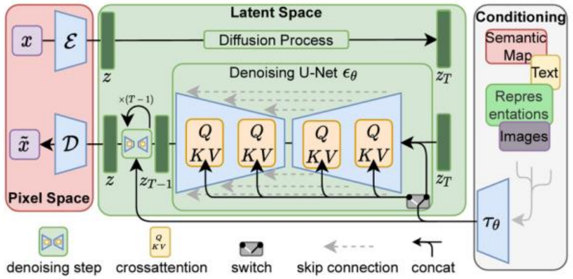
Visual Domain

[1] Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications* (Vol. 5, p. 62). Cham: Springer.

大模型时代的两种流行范式

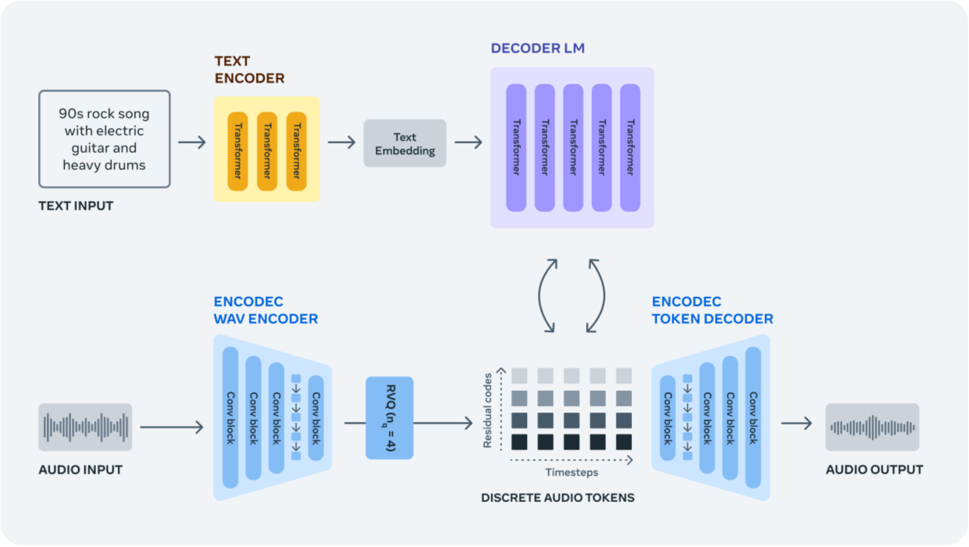


[2] Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In International Conference on Learning Representations.



[3] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

扩散模型
(Diffusion Models)



[4] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., ... & Défossez, A. (2024). Simple and controllable music generation. Advances in Neural Information Processing Systems, 36.

自回归语言模型
(Autoregressive Language Models)

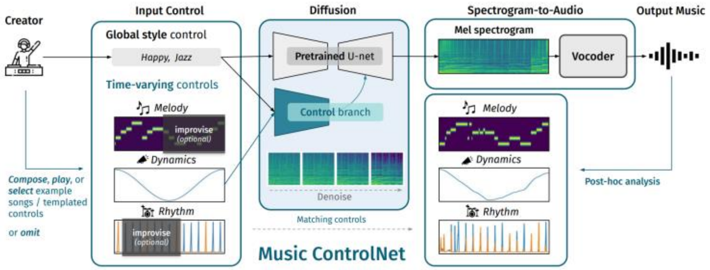
为什么需要增强模型的可控能力？

- 文本到音乐生成模型，只接受文本描述作为输入

- 对于音乐创作，用户需要更多控制：
  - 和弦进行、主旋律、鼓点、音色…

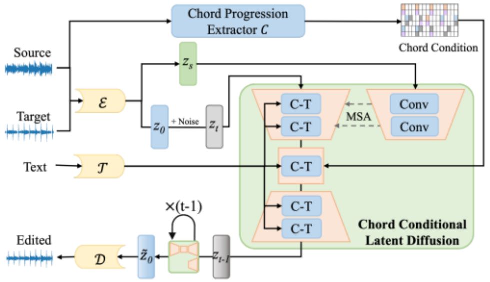
四种可行的思路

- **在预训练大模型上增加控制模块**
  - 在预训练阶段增加
  - 在微调阶段增加



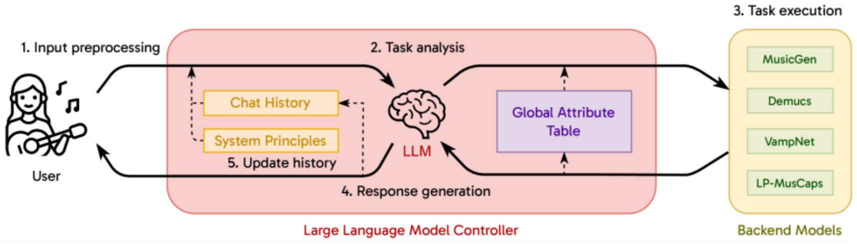
[5] Wu, S. L., Donahue, C., Watanabe, S., & Bryan, N. J. (2024). Music controlnet: Multiple time-varying controls for music generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, 2692-2703.

- **训练单独的音乐编辑模型**



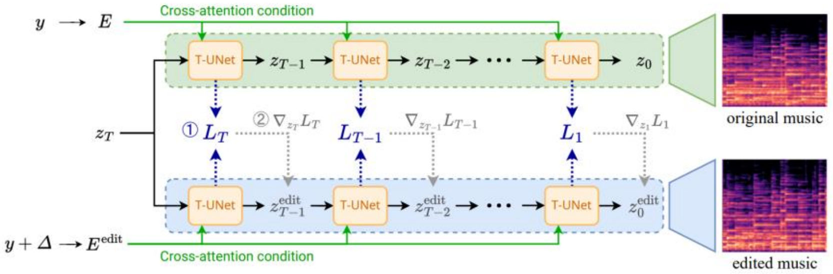
[6] Han, B., Dai, J., Hao, W., He, X., Guo, D., Chen, J., ... & Song, X. (2023). InstructME: An instruction guided music edit and remix framework with latent diffusion models. IJCAI 2024.

- **以代理(Agent)方式协调多种大模型**



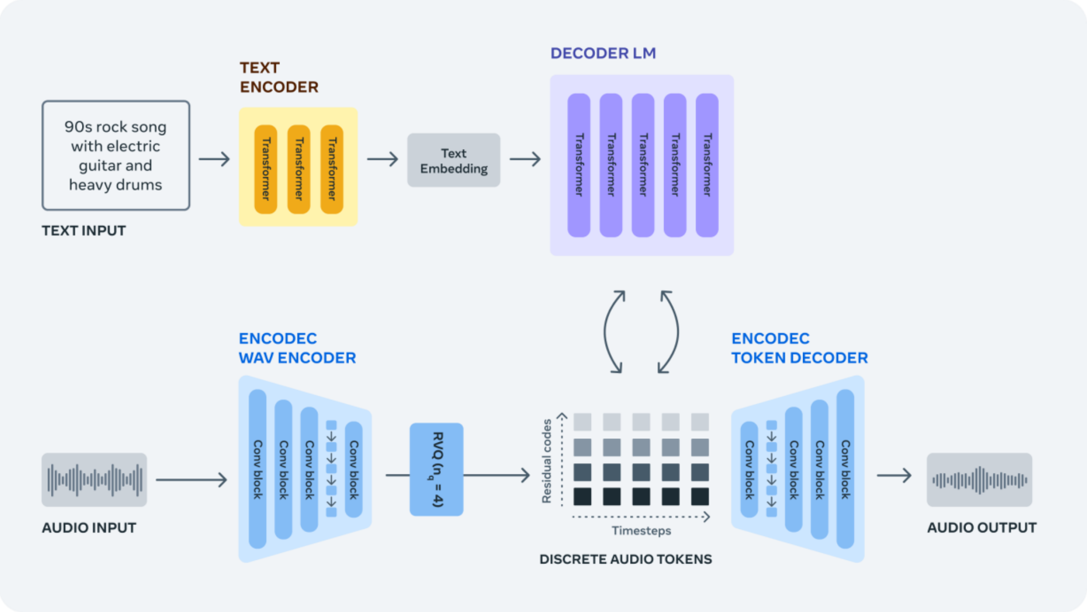
[7] Zhang, Y., Maezawa, A., Xia, G., Yamamoto, K., & Dixon, S. (2023). Loop copilot: Conducting ai ensembles for music generation and iterative editing. *arXiv preprint arXiv:2310.12404*.

- **在推理阶段介入控制**



[8] Zhang, Y., Ikemiya, Y., Xia, G., Murata, N., Martínez, M., Liao, W. H., ... & Dixon, S. (2024). MusicMagus: Zero-shot text-to-music editing via diffusion models. *arXiv preprint arXiv:2402.06178*.

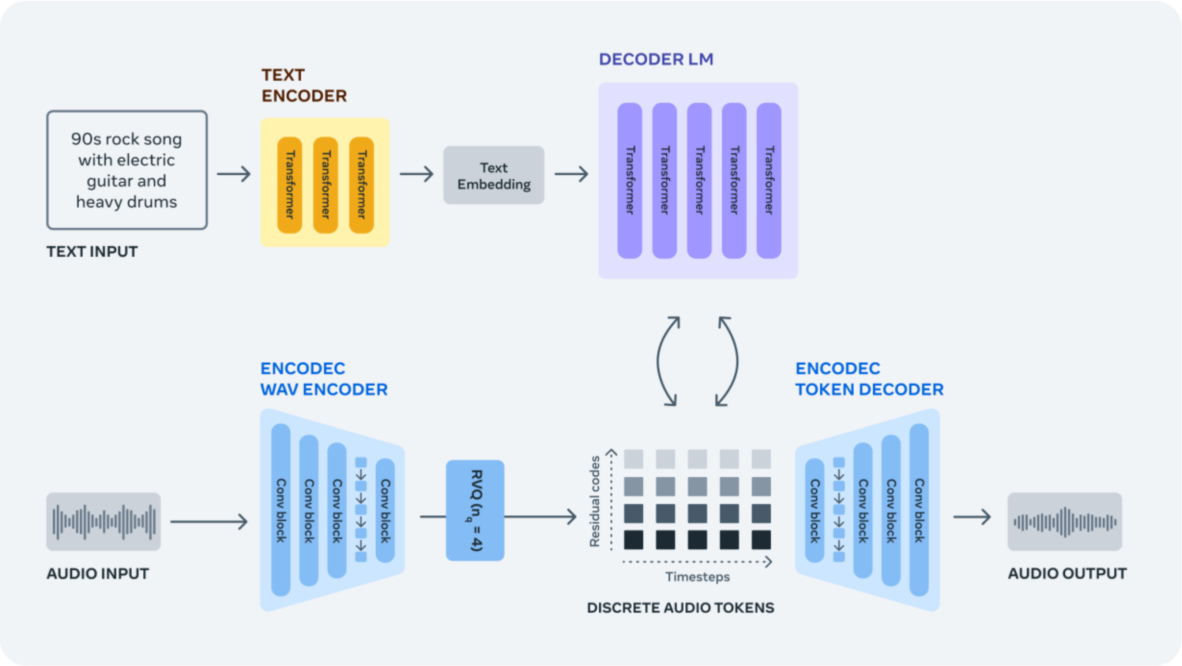
预训练模型： MusicGen



MusicGen包括**三个组件**：

1. 文本编码器
2. EnCodec
3. 潜空间的Transformer

预训练模型：MusicGen



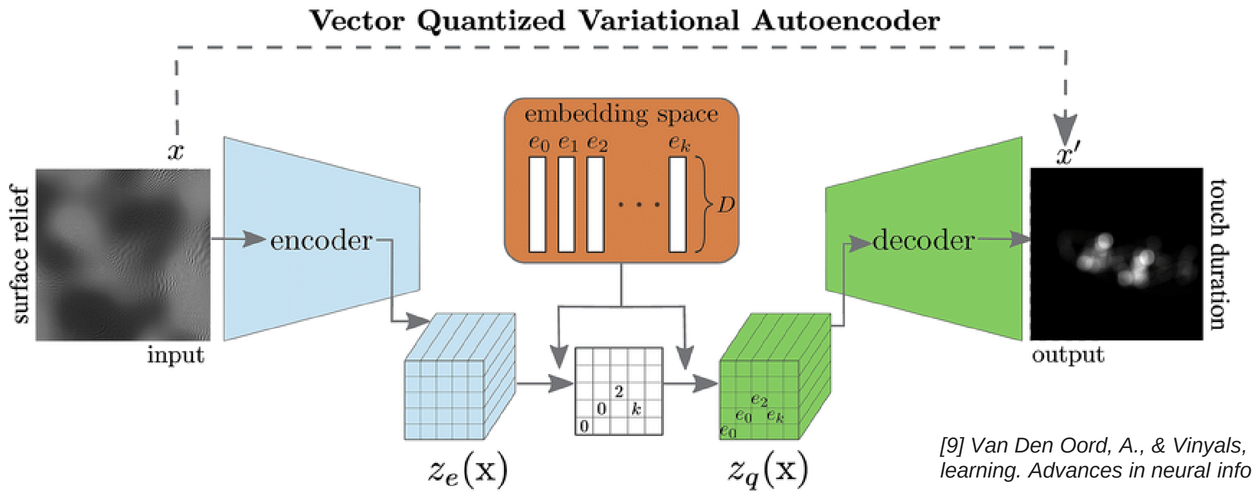
MusicGen包括**三个组件**：

1. 文本编码器：将**字符串**翻译为**编码向量**

2. EnCodec：将**连续的音乐波形**(32kHz)压缩成**离散的表示**(50Hz)

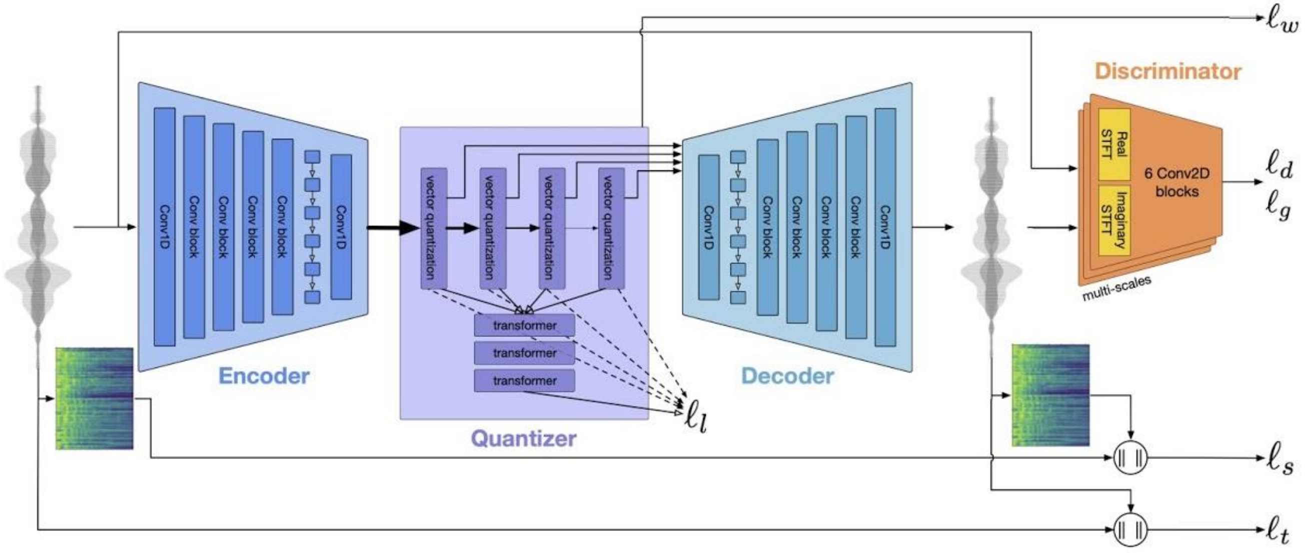
3. 潜空间的Transformer：负责建模音乐序列

EnCodec



[9] Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. Advances in neural information processing systems, 30.

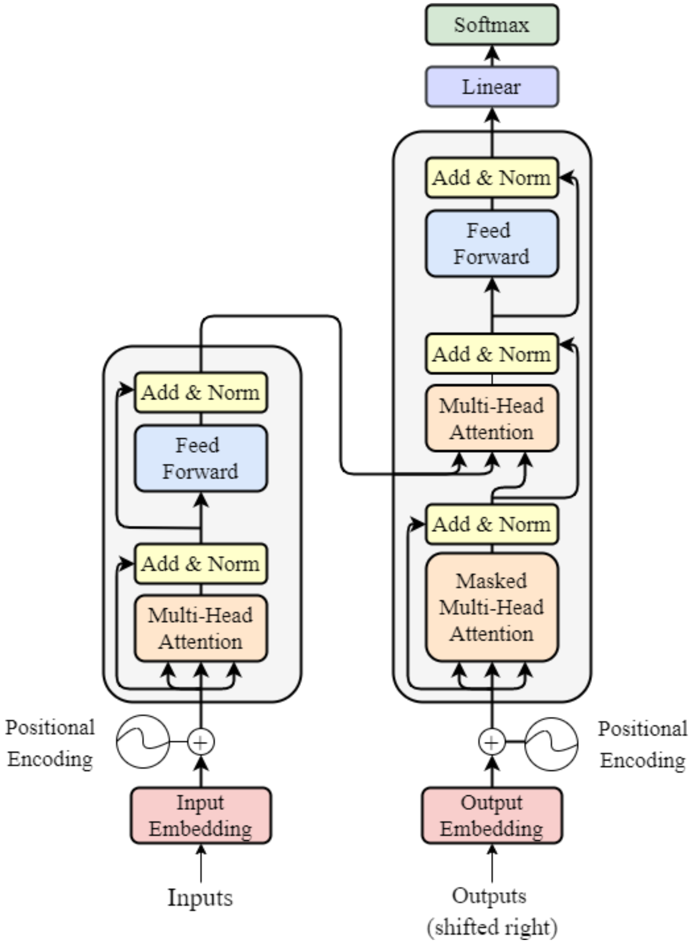
VQ-VAE



EnCodec

[10] Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. High Fidelity Neural Audio Compression. Transactions on Machine Learning Research.

Transformer



[11] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

每一层包含:

- 一个自注意力层
- 一个交叉注意力层
- 一个前馈层

音乐大语言模型的内容控制

Content-based Controls For Music Large Language Modeling, ISMIR 2024

Liwei Lin, Gus Xia, Junyan Jiang, Yixiao Zhang

论文主要贡献

- Coco-mulla在MusicGen上增加了**旋律、鼓点、和弦进行**的控制

- 仅在**300首**歌上微调原模型**4%**数量的新参数，**5小时**完成训练

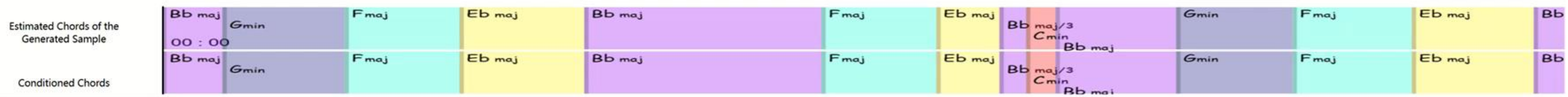
Liwei Lin, Gus Xia, Junyan Jiang, Yixiao Zhang @Music X Lab

We equip MusicGen, a text-to-music generation model, with direct and content-based controls on innate music languages such as pitch, chords and drum track. To this end, we contribute Coco-Mulla, a **co**ntent-based **co**ntrol method for **mu**sic **l**arge **la**nguage modeling. It uses a parameter-efficient fine-tuning (PEFT) method tailored for Transformer-based audio models. Our approach achieved low-resource semi-supervised learning, tuning with less than 4% parameters compared to the original model and training on a small dataset set with fewer than 300 *unannotated* songs. We illustrate the **chords** and **rhythms** control power of the model. Moreover, by combining **piano roll** and text descriptions, our system enables flexible music variation generation and style transfer.

See more details in our paper and try our model via github [coming soon] or huggingface [coming soon].

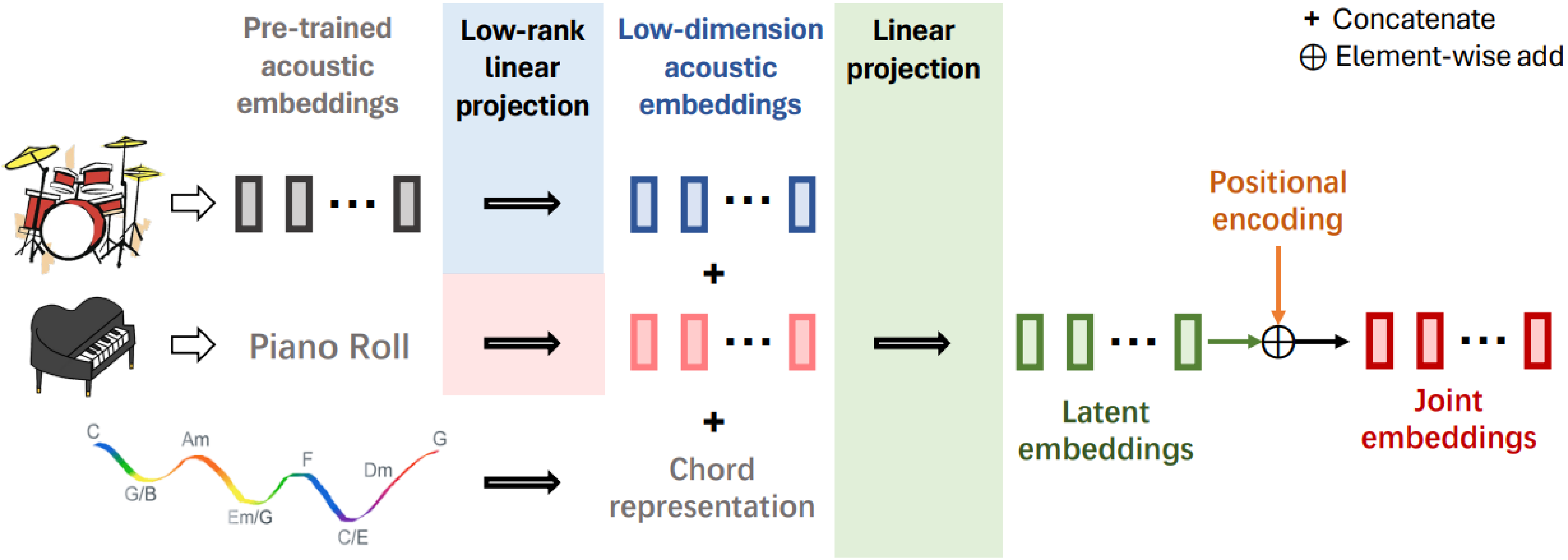
Chord and Drums Condition

Chord and rhythm controls via symbolic chord representation and acoustic drum tracks.
We transcribe chord progression of the generated audio samples using a chord recognition model.

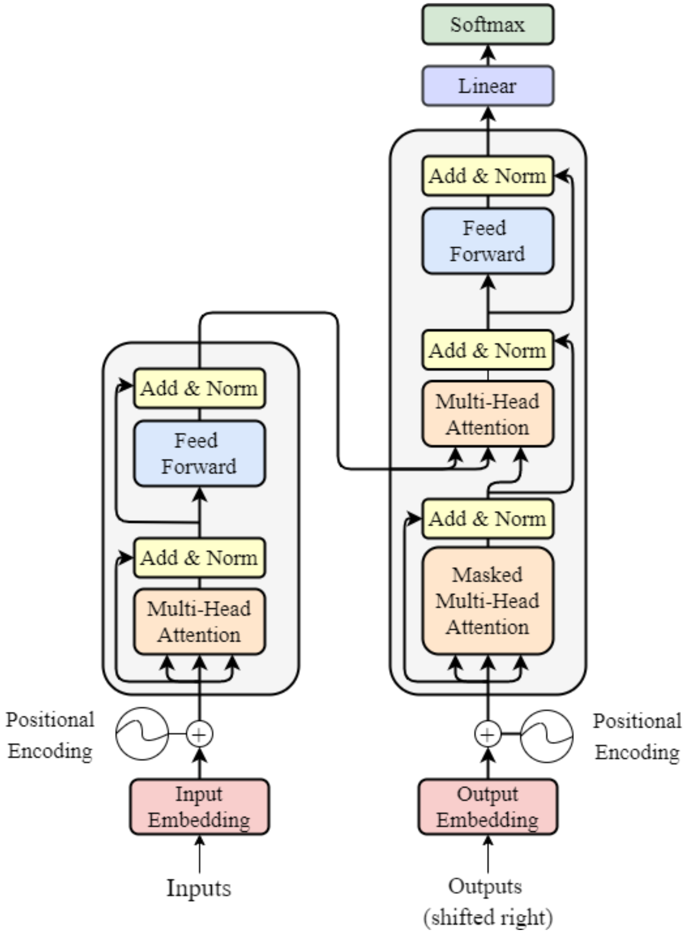


	Conditioned Drums	lazy jazz composition features a captivating saxophone solo that effortlessly melds with piano chords, skillfully weaving its way through the melody with languid grace. Instruments: saxophone, piano, drums.	relax folk song with a flute solo and acoustic guitar chords. instrument: flute, guitar, drums	happy piano with a swing melody. instrument: piano, drums	A grand orchestral arrangement with thunderous percussion, epic brass fanfares, and soaring strings, creating a cinematic atmosphere fit for a heroic battle.
Sample 001	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20
Sample 002	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20
Sample 003	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20
Sample 004	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20
Sample 005	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20	▶ 0:00 / 0:20

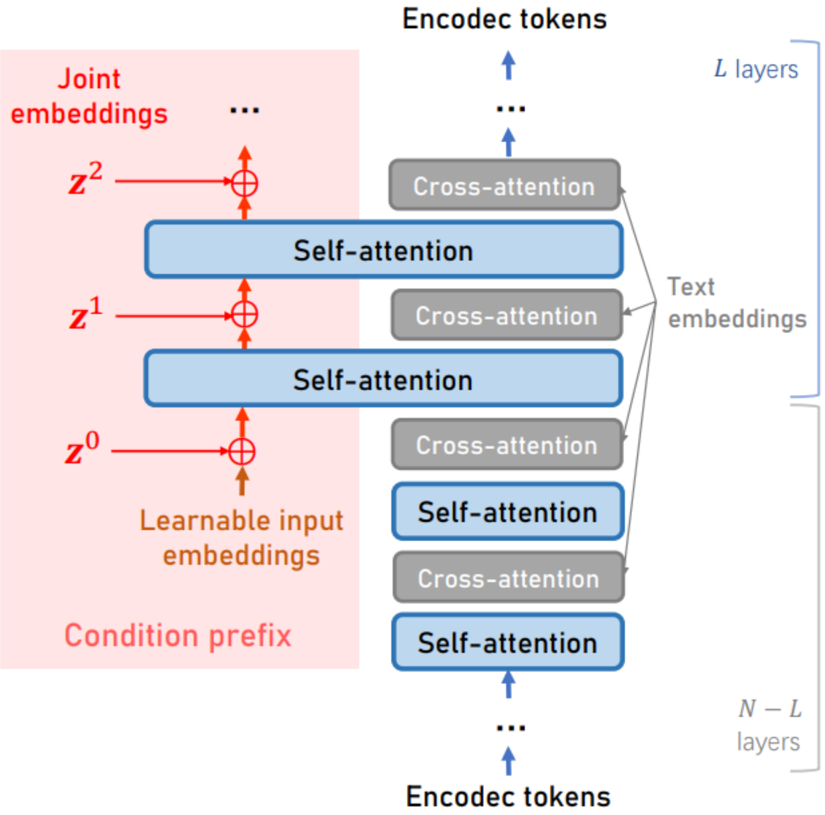
方法(1): 多个条件的编码嵌入



方法(2): 条件生成结构

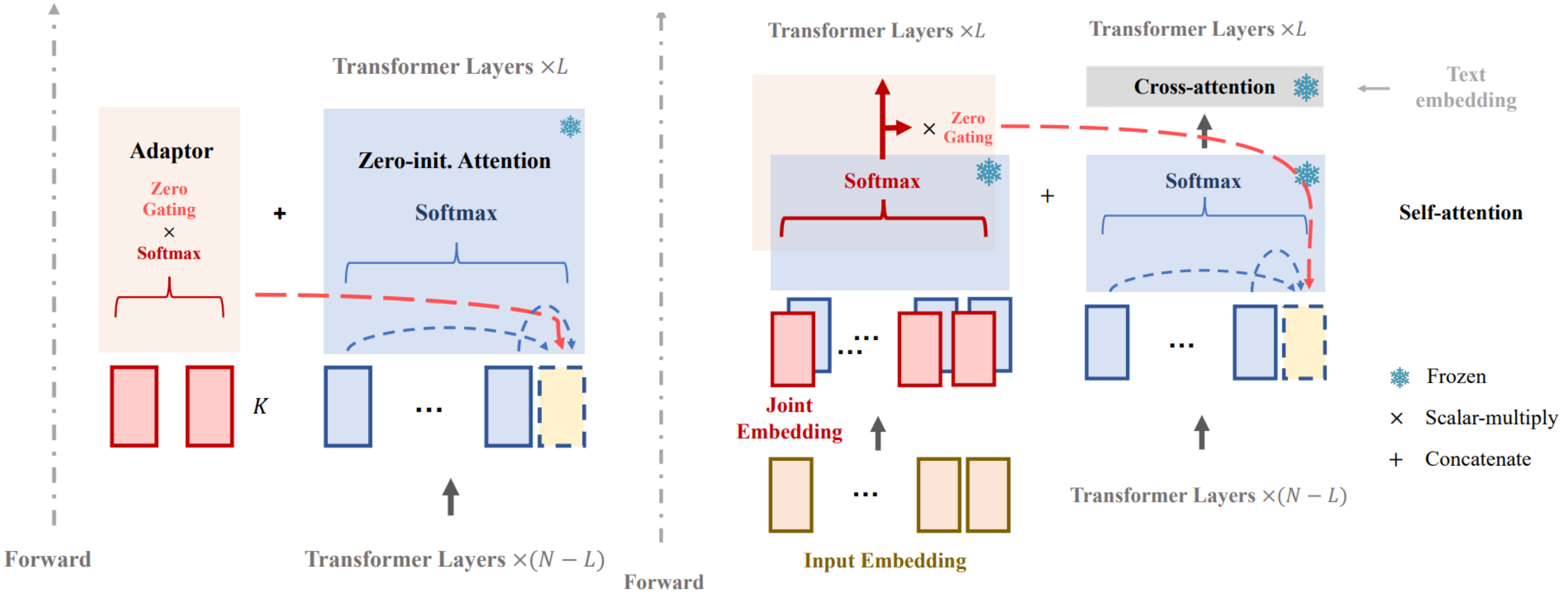


Transformer模型



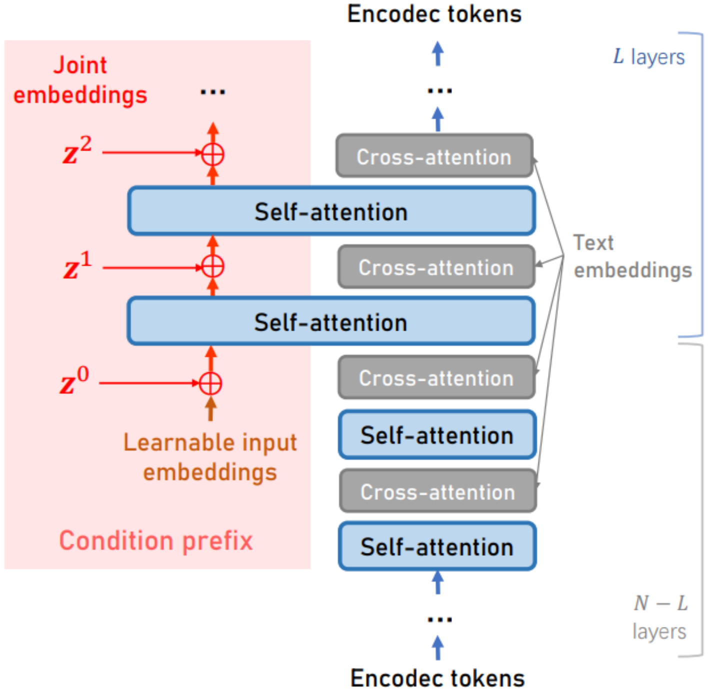
Coco-mulla模型

方法(2): 条件生成结构



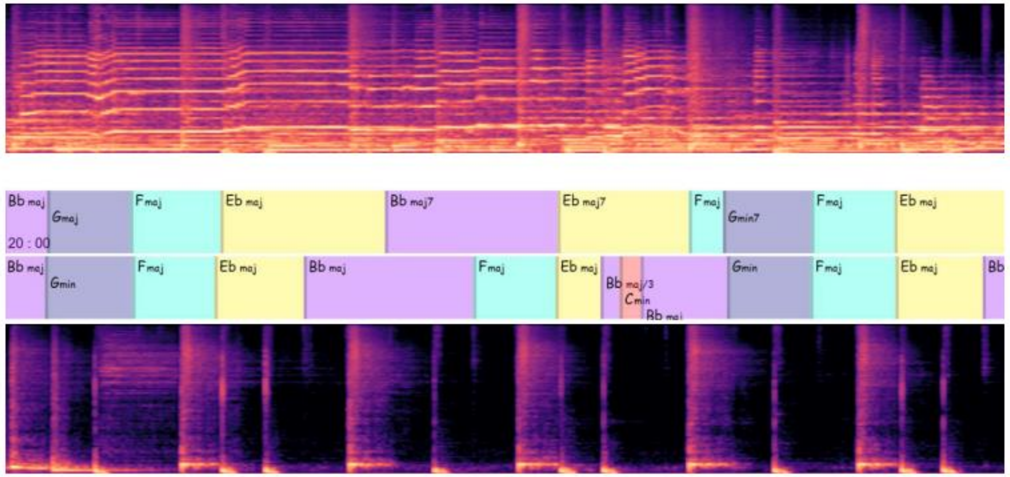
(a) LLaMA-Adapter

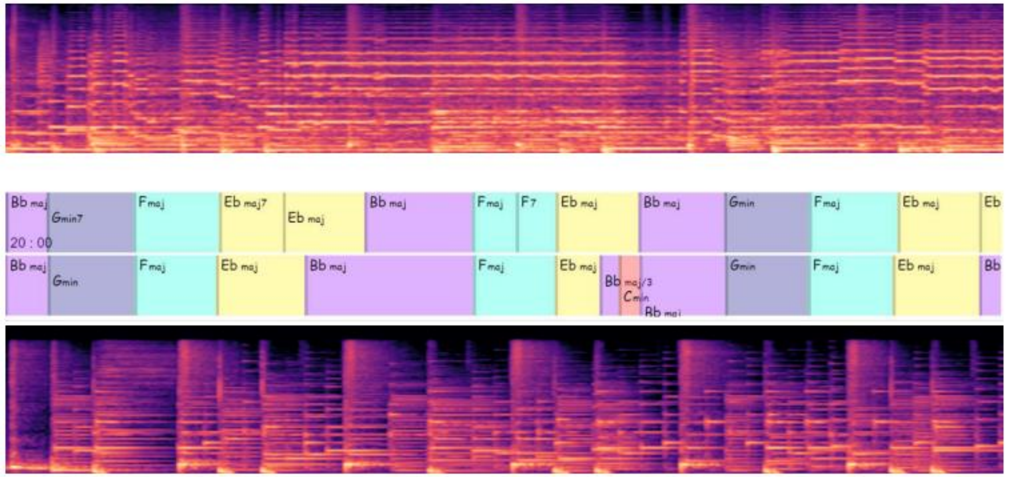
(b) The proposed conditional adaptor



实验评估

	$\textbf{Chord}_{\text{rec}}\uparrow$	$\textbf{Chord}^*_{\text{rec}}\uparrow$	$\textbf{Beat}_{F_1}\uparrow$	$\textbf{CLAP}_{\text{scr}}\uparrow$	$\textbf{FAD}_{\text{vgg}}\downarrow$	$\textbf{FAD}^*_{\text{vgg}}\downarrow$
Chord-only	0.412	0.195	-	**0.401**	6.209	6.695
MIDI-only	0.649	0.406	-	0.381	7.105	7.094
Drums-only	0.530	0.267	0.856	0.360	3.845	4.933
Full	**0.791**	**0.524**	**0.864**	0.351	**3.697**	**4.370**
MusicGen	-	-	-	0.441	6.434	6.847
Oracle	0.885	0.695	0.898	-	-	





编配、内绘和润色：通过基于内容的控制来指导音乐生成和编辑

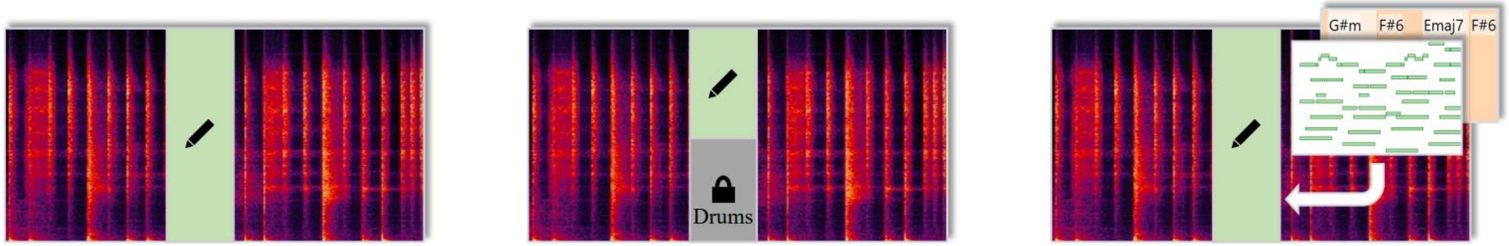
Arrange, Inpaint, and Refine: Steerable Long-term Music Audio Generation and Editing via Content-based Controls, IJCAI 2024

Liwei Lin, Gus Xia, Yixiao Zhang, Junyan Jiang

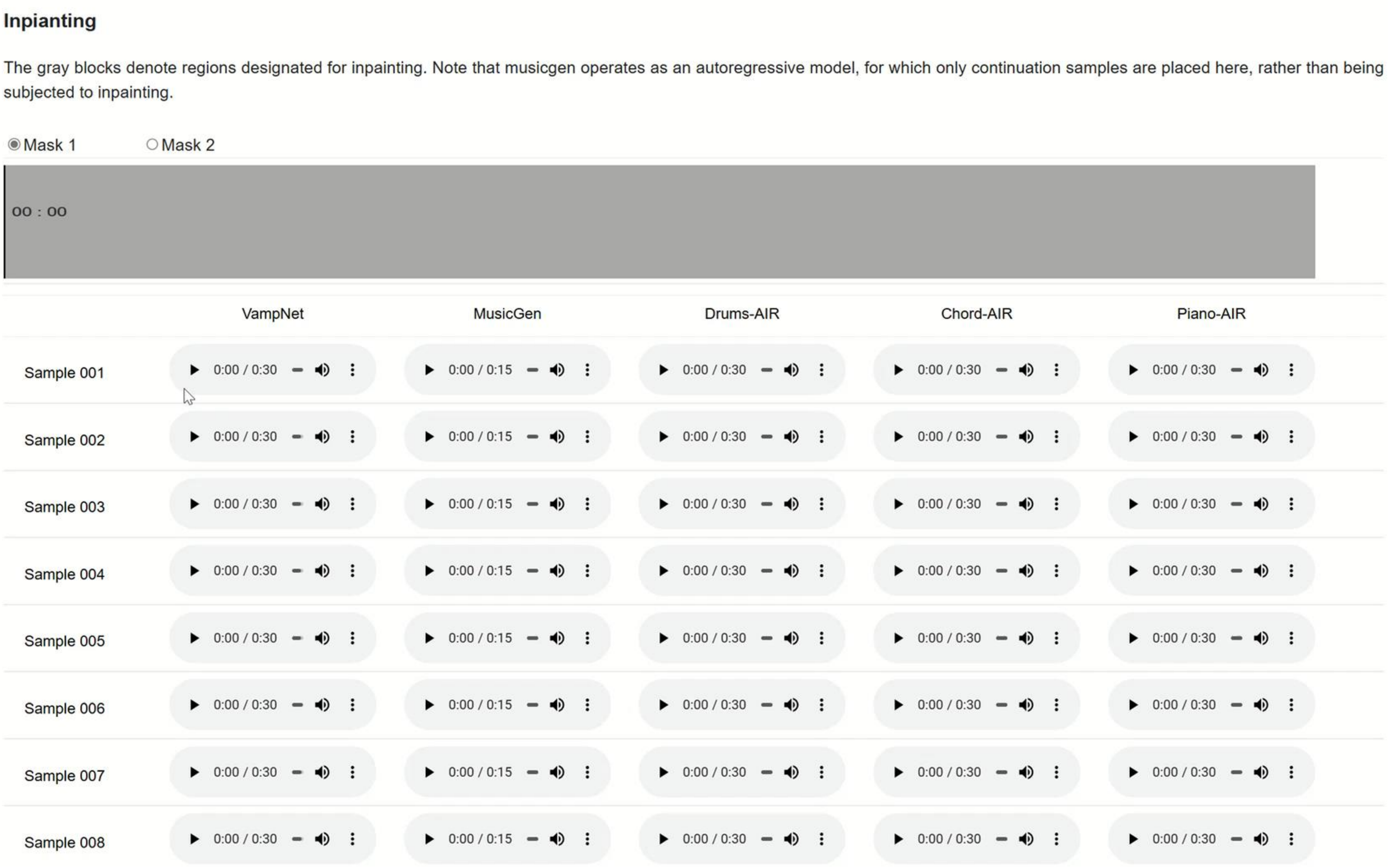
论文主要贡献

- 在Coco-mulla的基础上，新增：**音乐内绘、编配**任务

- 为此改进了模型结构，实现了更好的控制



(a) (b) (c)



Inpianting

The gray blocks denote regions designated for inpainting. Note that musicgen operates as an autoregressive model, for which only continuation samples are placed here, rather than being subjected to inpainting.

● Mask 1 ○ Mask 2

00 : 00

	VampNet	MusicGen	Drums-AIR	Chord-AIR	Piano-AIR
Sample 001	▶ 0:00 / 0:30	▶ 0:00 / 0:15	▶ 0:00 / 0:30	▶ 0:00 / 0:30	▶ 0:00 / 0:30
Sample 002	▶ 0:00 / 0:30	▶ 0:00 / 0:15	▶ 0:00 / 0:30	▶ 0:00 / 0:30	▶ 0:00 / 0:30
Sample 003	▶ 0:00 / 0:30	▶ 0:00 / 0:15	▶ 0:00 / 0:30	▶ 0:00 / 0:30	▶ 0:00 / 0:30
Sample 004	▶ 0:00 / 0:30	▶ 0:00 / 0:15	▶ 0:00 / 0:30	▶ 0:00 / 0:30	▶ 0:00 / 0:30
Sample 005	▶ 0:00 / 0:30	▶ 0:00 / 0:15	▶ 0:00 / 0:30	▶ 0:00 / 0:30	▶ 0:00 / 0:30
Sample 006	▶ 0:00 / 0:30	▶ 0:00 / 0:15	▶ 0:00 / 0:30	▶ 0:00 / 0:30	▶ 0:00 / 0:30
Sample 007	▶ 0:00 / 0:30	▶ 0:00 / 0:15	▶ 0:00 / 0:30	▶ 0:00 / 0:30	▶ 0:00 / 0:30
Sample 008	▶ 0:00 / 0:30	▶ 0:00 / 0:15	▶ 0:00 / 0:30	▶ 0:00 / 0:30	▶ 0:00 / 0:30

方法

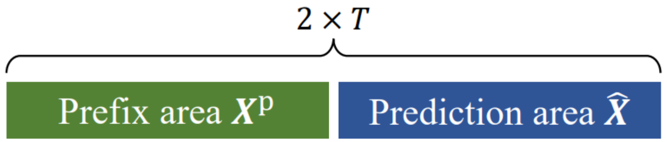
- 对于一个这样的内绘任务：

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{c}_3, \boldsymbol{c}_4, \boldsymbol{x}_5 \quad \rightarrow \quad \hat{\boldsymbol{x}}_3, \hat{\boldsymbol{x}}_4$$

- MusicGen是自回归模型，不支持内绘

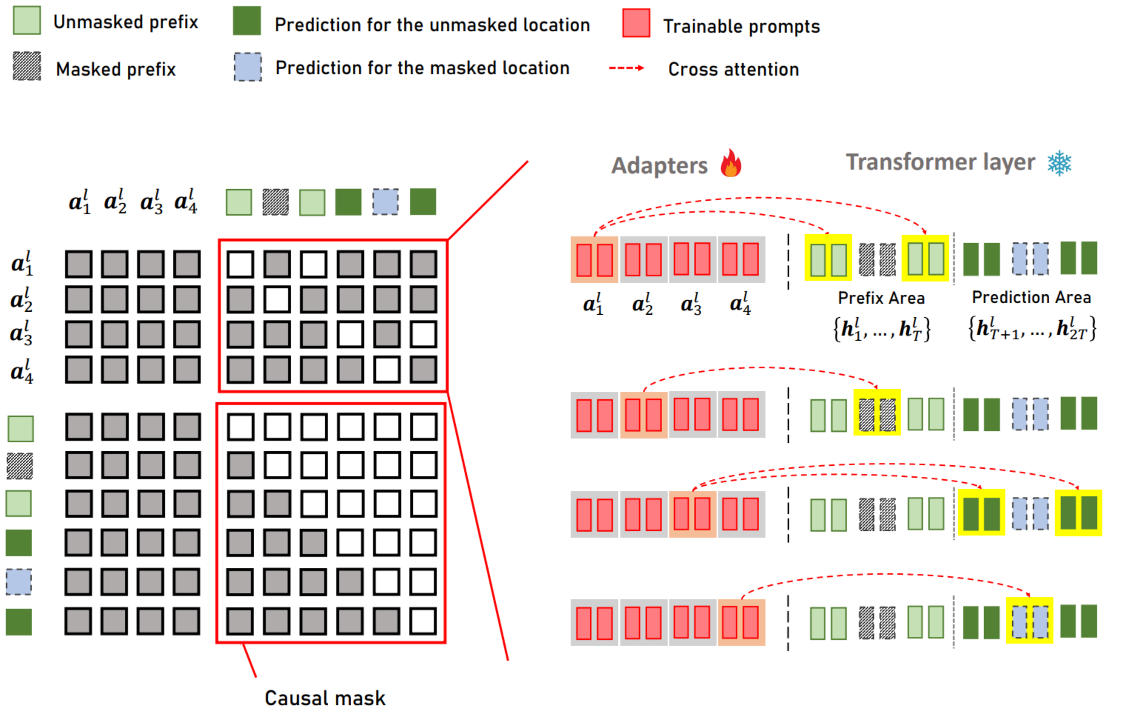
$\boldsymbol{x}_1, \boldsymbol{x}_2$ 可以输入，　$\boldsymbol{x}_5$ 不可以

- 我们通过这样的方式提供信息：


$2 \times T$
Prefix area $\boldsymbol{X}^{\mathrm{p}}$
Prediction area $\hat{\boldsymbol{X}}$

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{c}_3, \boldsymbol{c}_4, \boldsymbol{x}_5, \hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \hat{\boldsymbol{x}}_3, \hat{\boldsymbol{x}}_4, \hat{\boldsymbol{x}}_5$$

方法


Unmasked prefix
Masked prefix
Prediction for the unmasked location
Prediction for the masked location
Trainable prompts
Cross attention
a_1^l a_2^l a_3^l a_4^l
Causal mask
Adapters
Transformer layer
Prefix Area
$\{h_1^l, \dots, h_T^l\}$
Prediction Area
$\{h_{T+1}^l, \dots, h_{2T}^l\}$

$$r(t)\begin{cases}1, & \text{if } t \le T \text{ and } t\text{-th frame is unmasked,}\\ 2, & \text{if } t \le T \text{ and } t\text{-th frame is masked,}\\ 3, & \text{if } t > T \text{ and } (t-T)\text{-th frame is unmasked,}\\ 4, & \text{otherwise.}\end{cases}$$

输入的音乐表示：

$$\boldsymbol{H}^l = \{\boldsymbol{h}_1^l, \boldsymbol{h}_2^l, \dots, \boldsymbol{h}_{2T}^l\}$$

Transformer的self-attention：

$$\boldsymbol{S}^l = \text{Self-Attention}(\boldsymbol{H}^l).$$

$$\boldsymbol{S}^l = \{\boldsymbol{s}_1^l, \boldsymbol{s}_2^l, \dots \boldsymbol{s}_{2T}^l\}$$

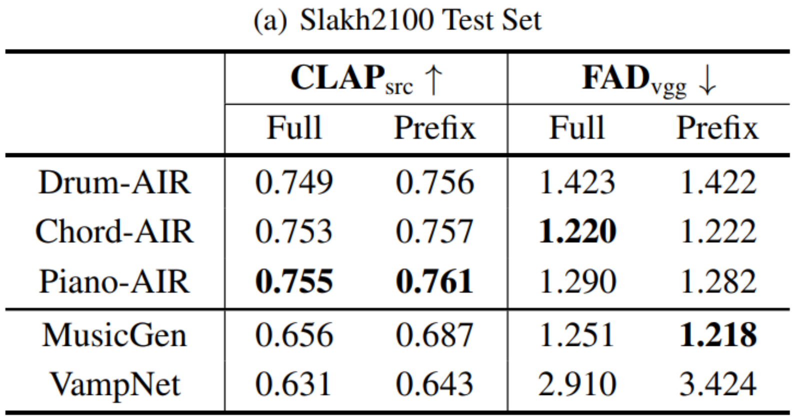
进行cross-attention计算：

$$\boldsymbol{u}_t^l = \text{Cross-Attention}(\boldsymbol{h}_t^l, \boldsymbol{a}_{r(t)}^l).$$

Adapter和Transformer信息融合：

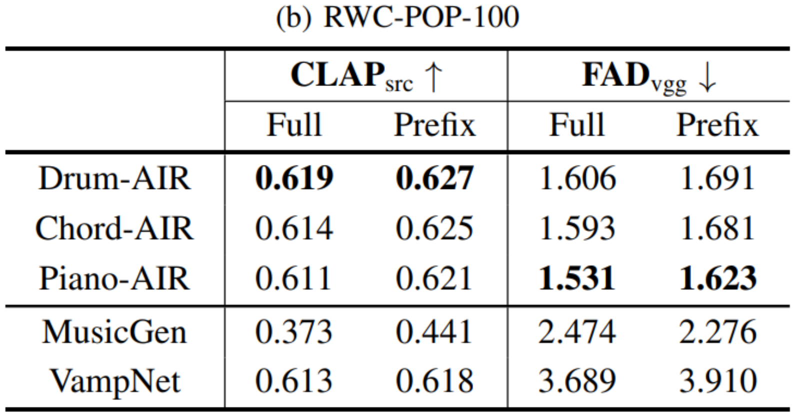
$$\boldsymbol{s}_t^{l,*} = \boldsymbol{s}_t^l + g_{r(t)}^l \cdot \boldsymbol{u}_t^l,$$

实验



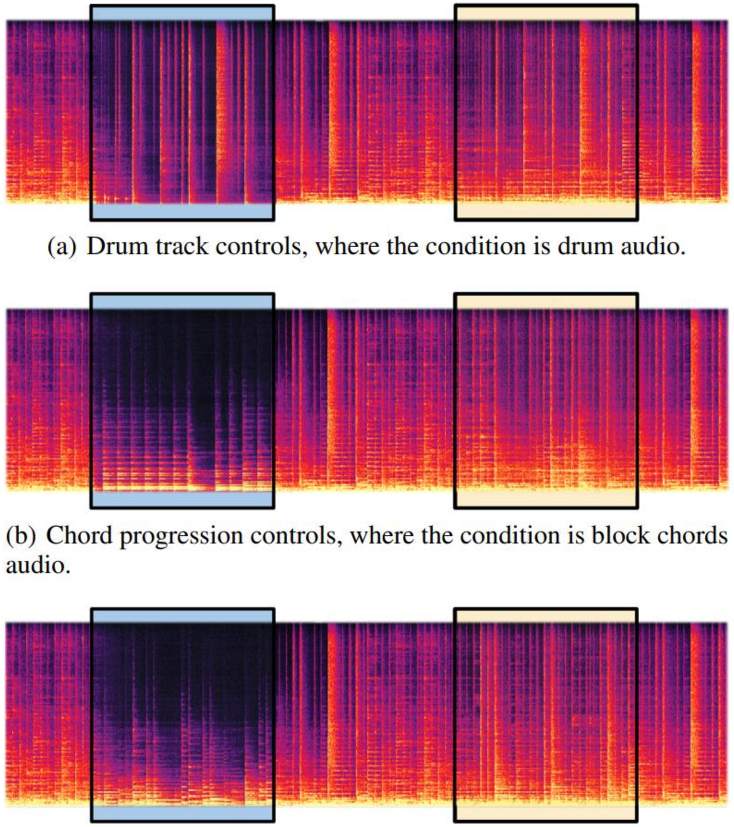
(a) Slakh2100 Test Set

	$\textbf{CLAP}_{src}\uparrow$		$\textbf{FAD}_{vgg}\downarrow$	
	Full	Prefix	Full	Prefix
Drum-AIR	0.749	0.756	1.423	1.422
Chord-AIR	0.753	0.757	**1.220**	1.222
Piano-AIR	**0.755**	**0.761**	1.290	1.282
MusicGen	0.656	0.687	1.251	**1.218**
VampNet	0.631	0.643	2.910	3.424



(b) RWC-POP-100

	$\textbf{CLAP}_{src}\uparrow$		$\textbf{FAD}_{vgg}\downarrow$	
	Full	Prefix	Full	Prefix
Drum-AIR	**0.619**	**0.627**	1.606	1.691
Chord-AIR	0.614	0.625	1.593	1.681
Piano-AIR	0.611	0.621	**1.531**	**1.623**
MusicGen	0.373	0.441	2.474	2.276
VampNet	0.613	0.618	3.689	3.910



(a) Drum track controls, where the condition is drum audio.

(b) Chord progression controls, where the condition is block chords audio.

(c) Arrangement and orchestration from piano cover, where the condition is piano cover audio.

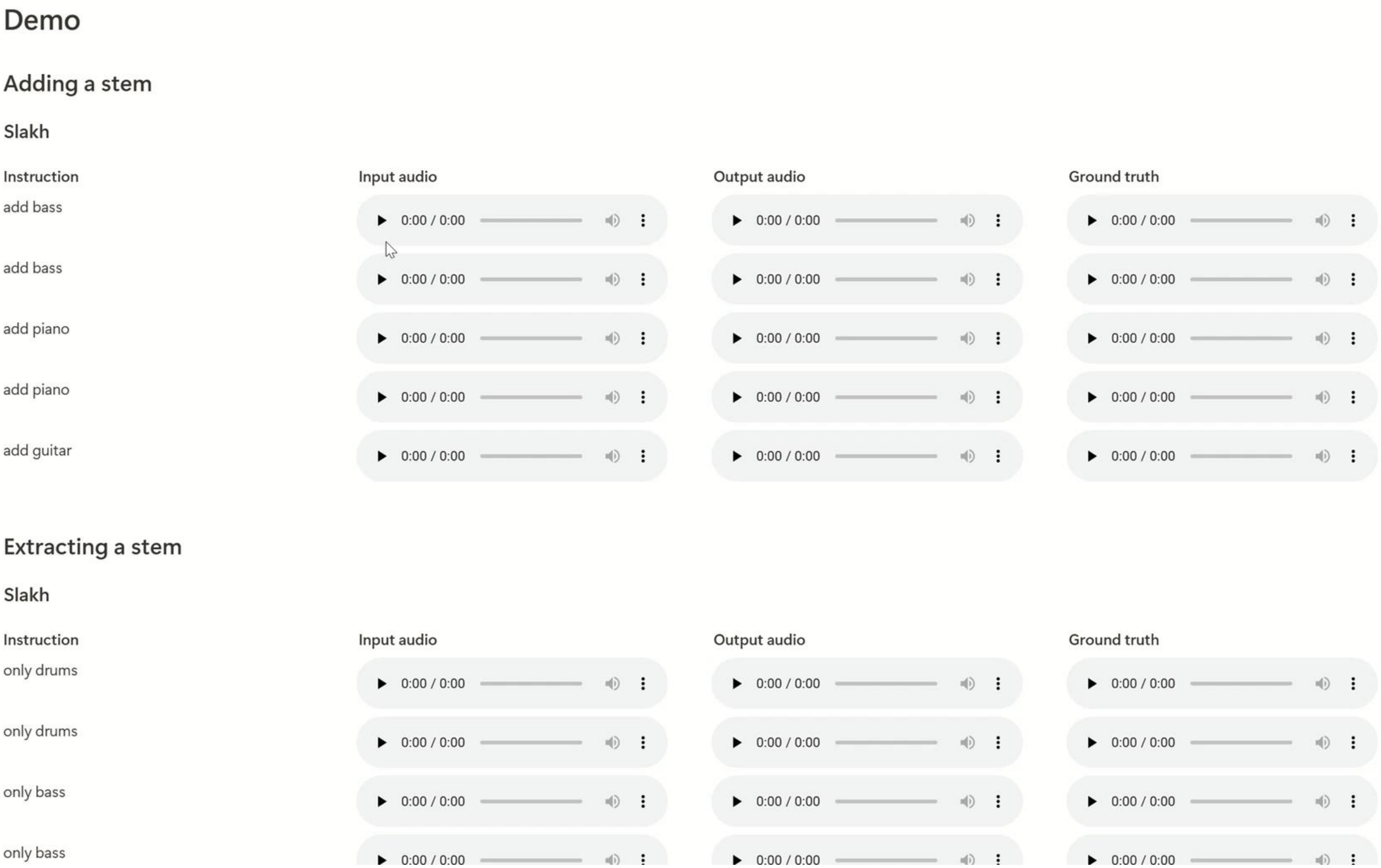
指令微调的音乐大模型：多任务音乐编辑

Instruct-MusicGen: Unlocking Text-to-Music Editing for Music Language Models via Instruction Tuning, arXiv/2405.18386

Yixiao Zhang, Yukara Ikemiya, Naoki Murata, Woosung Choi, Marco Martínez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, Simon Dixon

论文主要贡献

- 在Coco-mulla的基础上，新增：**添加**、**消除**、**提取**乐器任务
- 通过指令微调，让同一个模型能够适应多任务指令
- 同样只需要训练很短时间：**5000 steps**微调



Demo

Adding a stem

Slakh

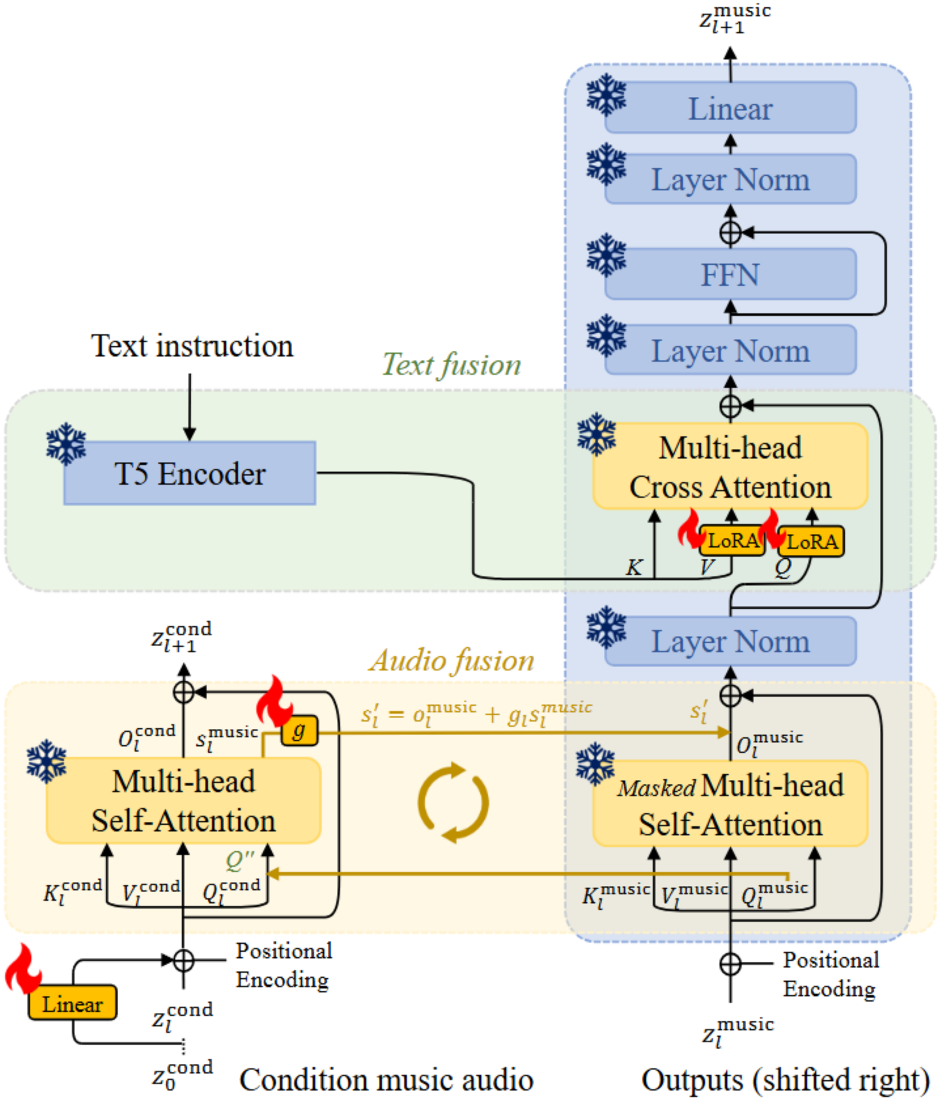
Instruction	Input audio	Output audio	Ground truth
add bass	▶ 0:00 / 0:00	▶ 0:00 / 0:00	▶ 0:00 / 0:00
add bass	▶ 0:00 / 0:00	▶ 0:00 / 0:00	▶ 0:00 / 0:00
add piano	▶ 0:00 / 0:00	▶ 0:00 / 0:00	▶ 0:00 / 0:00
add piano	▶ 0:00 / 0:00	▶ 0:00 / 0:00	▶ 0:00 / 0:00
add guitar	▶ 0:00 / 0:00	▶ 0:00 / 0:00	▶ 0:00 / 0:00

Extracting a stem

Slakh

Instruction	Input audio	Output audio	Ground truth
only drums	▶ 0:00 / 0:00	▶ 0:00 / 0:00	▶ 0:00 / 0:00
only drums	▶ 0:00 / 0:00	▶ 0:00 / 0:00	▶ 0:00 / 0:00
only bass	▶ 0:00 / 0:00	▶ 0:00 / 0:00	▶ 0:00 / 0:00
only bass	▶ 0:00 / 0:00	▶ 0:00 / 0:00	▶ 0:00 / 0:00

方法



输入的音乐表示：

$$Z^{\text{cond}} = \{z_0^{\text{cond}}, z_1^{\text{cond}}, \ldots, z_M^{\text{cond}}\},$$

$$Z^{\text{music}} = \{z_0^{\text{music}}, z_1^{\text{music}}, \ldots, z_M^{\text{music}}\},$$

Transformer的self-attention：

$$Q_l^{\text{music}}, K_l^{\text{music}}, V_l^{\text{music}} = \text{QKV-projector}(z_l^{\text{music}}),$$

$$o_l^{\text{music}} = \text{SelfAttn}(Q_l^{\text{music}}, K_l^{\text{music}}, V_l^{\text{music}}).$$

将条件音轨经过一层变换：

$$h = f_l(z^{\text{cond}}) + e_l.$$

条件Transformer的self-attention：

$$Q_l^{\text{cond}}, K_l^{\text{cond}}, V_l^{\text{cond}} = \text{QKV-projector}(z_l^{\text{cond}} + h),$$

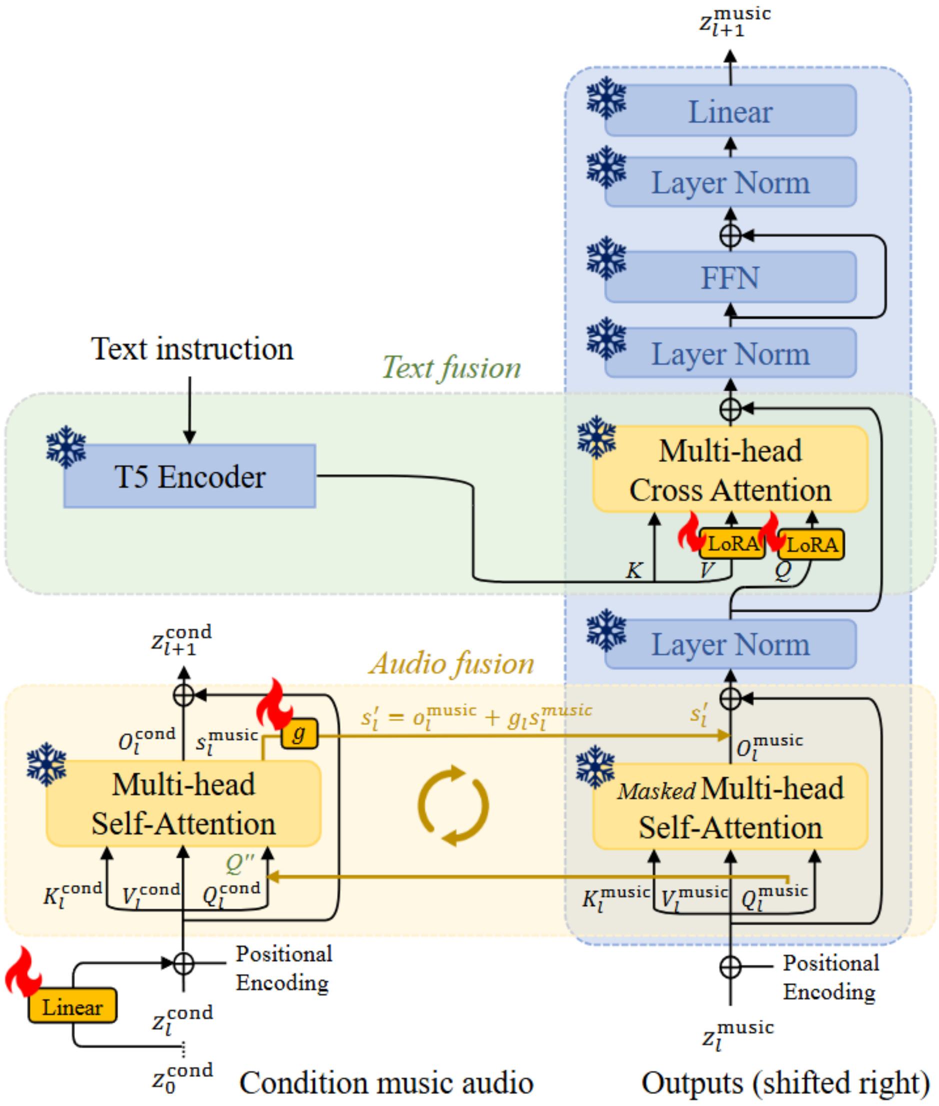
$$z_{l+1}^{\text{cond}} = \text{SelfAttn}(Q_l^{\text{cond}}, K_l^{\text{cond}}, V_l^{\text{cond}}).$$

两个self-attention信息进行融合：

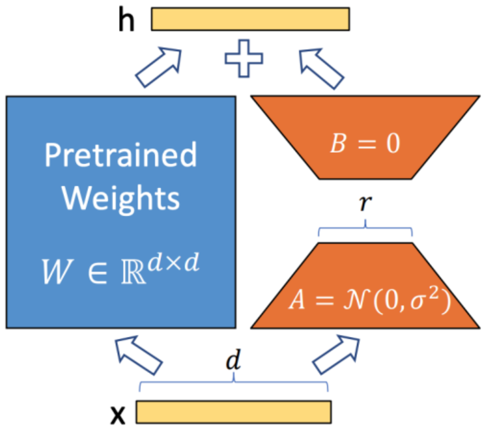
$$s_l^{\text{music}} = \text{CrossAttn}(Q_l^{\text{music}} + Q_l^{\text{cond}}, K_l^{\text{cond}}, V_l^{\text{cond}}).$$

$$s_l' = o_l^{\text{music}} + g_l \cdot s_l^{\text{music}},$$

方法



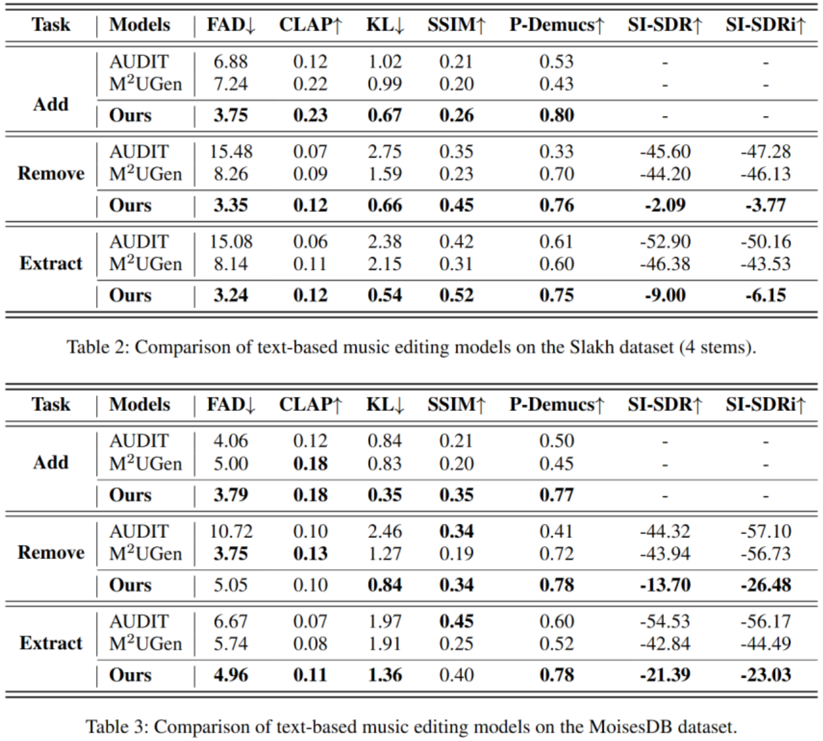
使用LoRA微调文本编码器的cross-attention：


h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

得到最终的音乐表示：

$$z_{l+1}^{\text{music}} = \text{TextFusion}(s_l', X^{\text{instruct}}),$$

实验

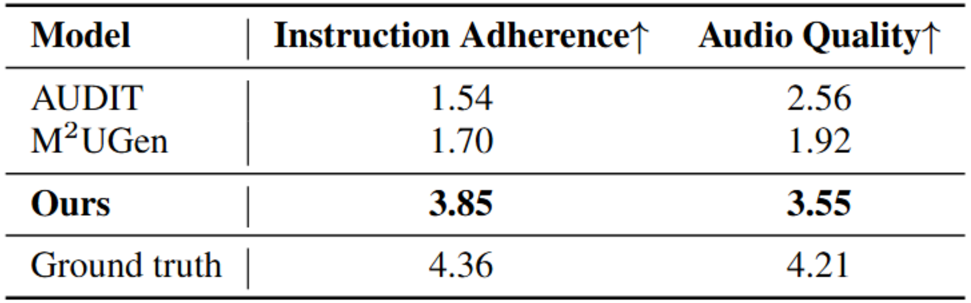


Task	Models	FAD↓	CLAP↑	KL↓	SSIM↑	P-Demucs↑	SI-SDR↑	SI-SDRi↑
Add	AUDIT	6.88	0.12	1.02	0.21	0.53	-	-
	M^{2}UGen	7.24	0.22	0.99	0.20	0.43	-	-
	Ours	**3.75**	**0.23**	**0.67**	**0.26**	**0.80**	-	-
Remove	AUDIT	15.48	0.07	2.75	0.35	0.33	-45.60	-47.28
	M^{2}UGen	8.26	0.09	1.59	0.23	0.70	-44.20	-46.13
	Ours	**3.35**	**0.12**	**0.66**	**0.45**	**0.76**	**-2.09**	**-3.77**
Extract	AUDIT	15.08	0.06	2.38	0.42	0.61	-52.90	-50.16
	M^{2}UGen	8.14	0.11	2.15	0.31	0.60	-46.38	-43.53
	Ours	**3.24**	**0.12**	**0.54**	**0.52**	**0.75**	**-9.00**	**-6.15**

Table 2: Comparison of text-based music editing models on the Slakh dataset (4 stems).

Task	Models	FAD↓	CLAP↑	KL↓	SSIM↑	P-Demucs↑	SI-SDR↑	SI-SDRi↑
Add	AUDIT	4.06	0.12	0.84	0.21	0.50	-	-
	M^{2}UGen	5.00	**0.18**	0.83	0.20	0.45	-	-
	Ours	**3.79**	**0.18**	**0.35**	**0.35**	**0.77**	-	-
Remove	AUDIT	10.72	0.10	2.46	**0.34**	0.41	-44.32	-57.10
	M^{2}UGen	**3.75**	**0.13**	1.27	0.19	0.72	-43.94	-56.73
	Ours	5.05	0.10	**0.84**	**0.34**	**0.78**	**-13.70**	**-26.48**
Extract	AUDIT	6.67	0.07	1.97	**0.45**	0.60	-54.53	-56.17
	M^{2}UGen	5.74	0.08	1.91	0.25	0.52	-42.84	-44.49
	Ours	**4.96**	**0.11**	**1.36**	0.40	**0.78**	**-21.39**	**-23.03**

Table 3: Comparison of text-based music editing models on the MoisesDB dataset.



Model	**Instruction Adherence↑**	**Audio Quality↑**
AUDIT	1.54	2.56
M^{2}UGen	1.70	1.92
Ours	**3.85**	**3.55**
Ground truth	4.36	4.21

客观实验

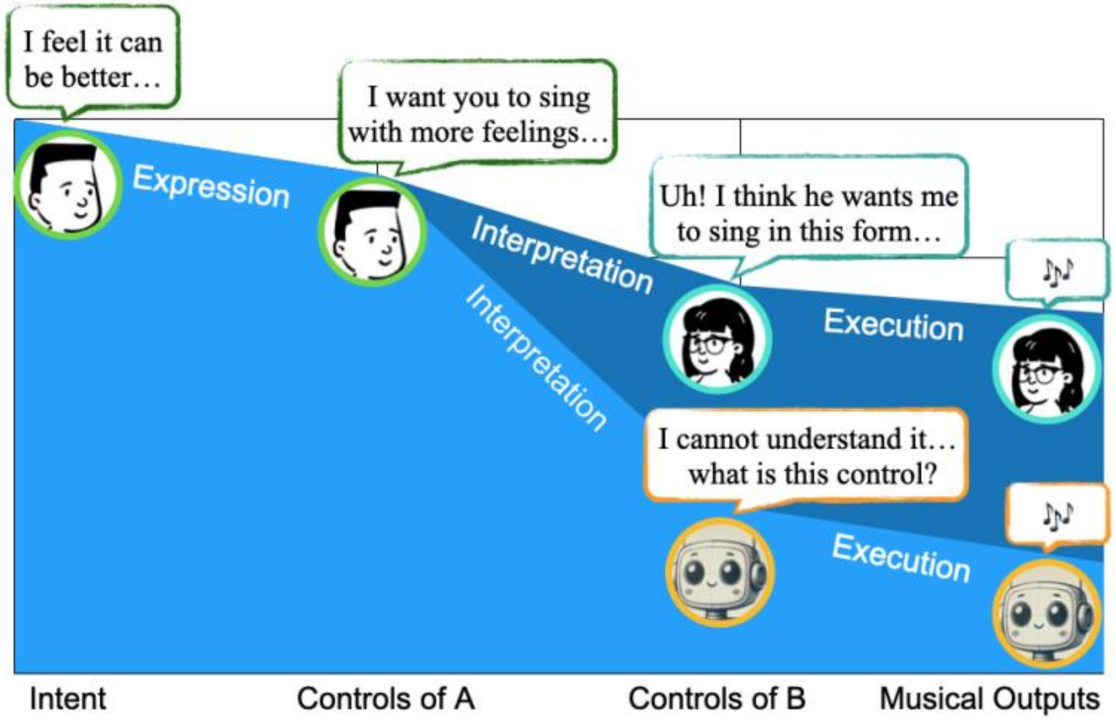
主观实验

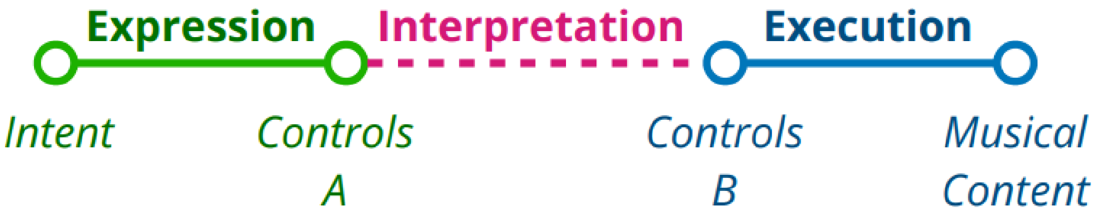
音乐大模型为什么仍然不能像人类一样合作?

The Interpretation Gap in Text-to-Music Generation Models, NLP4MusA Workshop @ ISMIR 2024

Yongyi Zang, Yixiao Zhang**

音乐协作的三个阶段

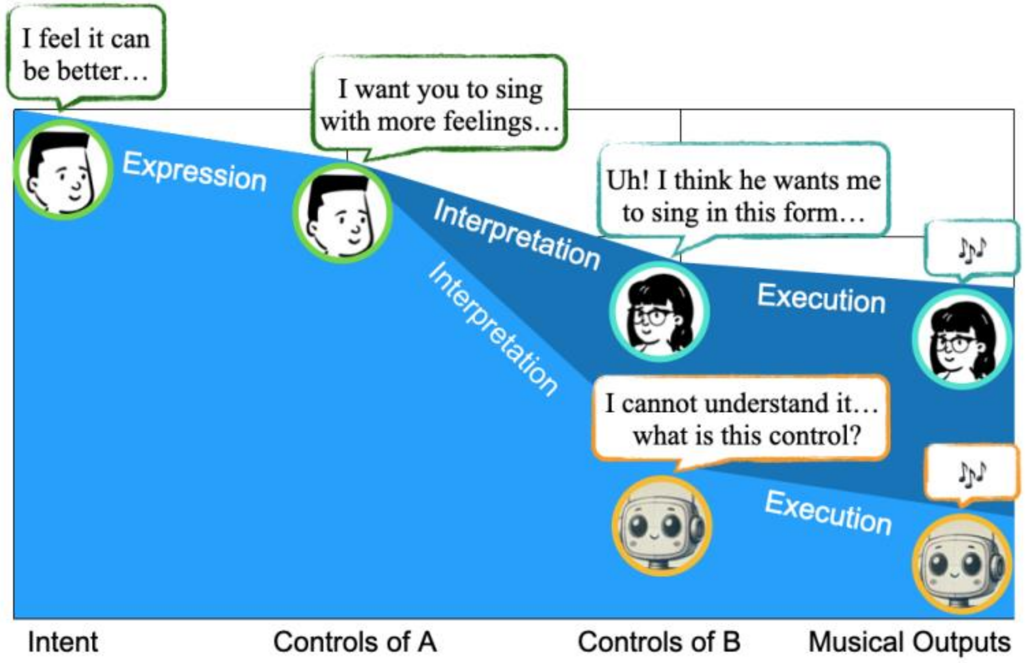

I feel it can be better…
I want you to sing with more feelings…
Uh! I think he wants me to sing in this form…
I cannot understand it… what is this control?
Expression
Interpretation
Interpretation
Execution
Execution
Intent
Controls of A
Controls of B
Musical Outputs

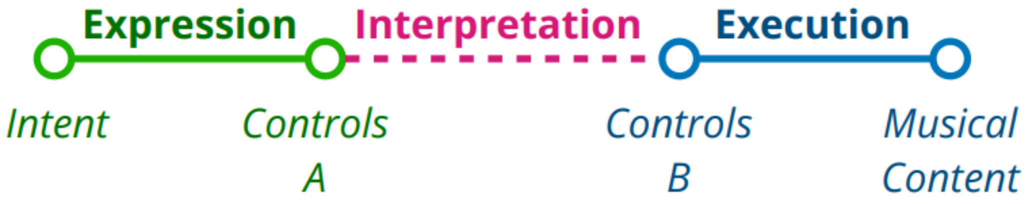


Expression Interpretation Execution

Intent *Controls A* *Controls B* *Musical Content*

音乐协作的三个阶段


I feel it can be better…
I want you to sing with more feelings…
Expression
Interpretation
Interpretation
Uh! I think he wants me to sing in this form…
Execution
I cannot understand it… what is this control?
Execution
Intent
Controls of A
Controls of B
Musical Outputs



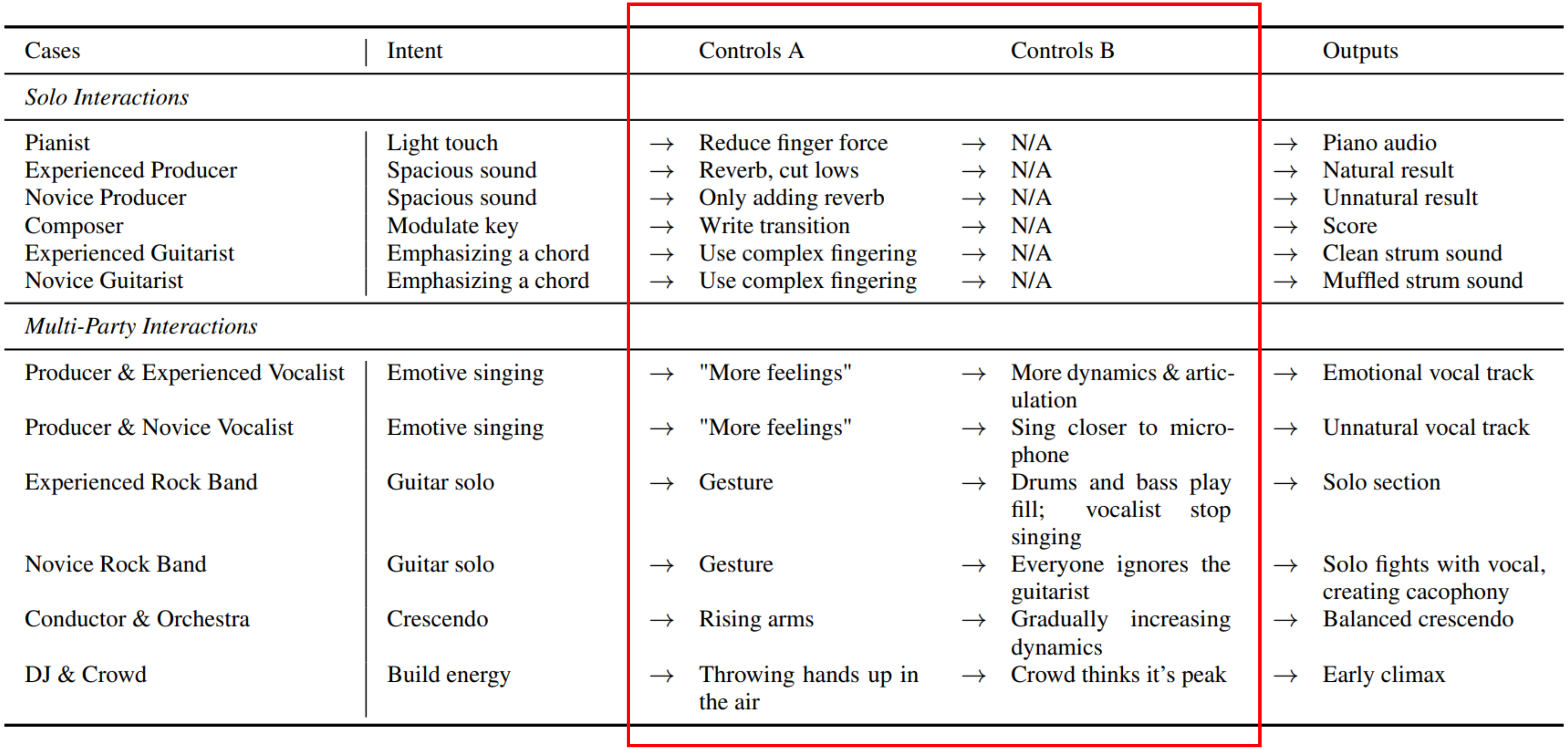
- **人工智能模型与人类存在“解释差距”**

- 音乐家在创作过程中会使用各种形式的控制信号，包括文字描述、旋律片段、和弦进程…

- 这些信号**可能非常具体**也可能带有**相当程度的模糊性**
- 例如：
  - “让这个贝斯音轨听起来更加温暖”
  - “让这段旋律听起来更具有表现力”

- 包含很多上下文信息和隐含的意义，需要接收者有一定的音乐素养才能正确理解并执行

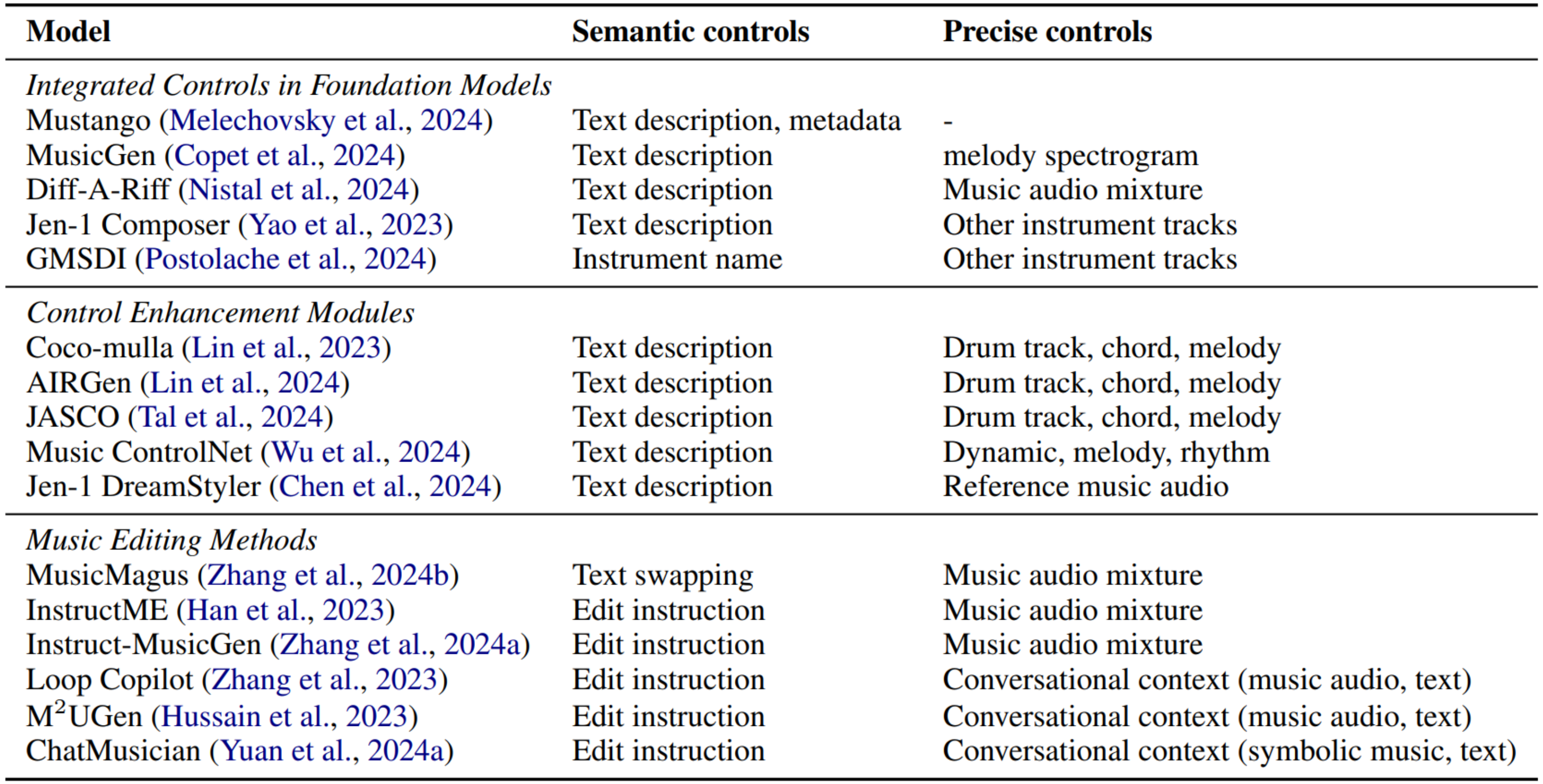
人类音乐协作的控制信号



Cases	Intent		Controls A		Controls B		Outputs
Solo Interactions							
Pianist	Light touch	→	Reduce finger force	→	N/A	→	Piano audio
Experienced Producer	Spacious sound	→	Reverb, cut lows	→	N/A	→	Natural result
Novice Producer	Spacious sound	→	Only adding reverb	→	N/A	→	Unnatural result
Composer	Modulate key	→	Write transition	→	N/A	→	Score
Experienced Guitarist	Emphasizing a chord	→	Use complex fingering	→	N/A	→	Clean strum sound
Novice Guitarist	Emphasizing a chord	→	Use complex fingering	→	N/A	→	Muffled strum sound
Multi-Party Interactions							
Producer & Experienced Vocalist	Emotive singing	→	"More feelings"	→	More dynamics & articulation	→	Emotional vocal track
Producer & Novice Vocalist	Emotive singing	→	"More feelings"	→	Sing closer to microphone	→	Unnatural vocal track
Experienced Rock Band	Guitar solo	→	Gesture	→	Drums and bass play fill; vocalist stop singing	→	Solo section
Novice Rock Band	Guitar solo	→	Gesture	→	Everyone ignores the guitarist	→	Solo fights with vocal, creating cacophony
Conductor & Orchestra	Crescendo	→	Rising arms	→	Gradually increasing dynamics	→	Balanced crescendo
DJ & Crowd	Build energy	→	Throwing hands up in the air	→	Crowd thinks it's peak	→	Early climax

现在的文本到音乐生成模型**不具备**必要的理解能力

现有的文本到音乐生成控制一览



Model	Semantic controls	Precise controls
Integrated Controls in Foundation Models		
Mustango (Melechovsky et al., 2024)	Text description, metadata	-
MusicGen (Copet et al., 2024)	Text description	melody spectrogram
Diff-A-Riff (Nistal et al., 2024)	Text description	Music audio mixture
Jen-1 Composer (Yao et al., 2023)	Text description	Other instrument tracks
GMSDI (Postolache et al., 2024)	Instrument name	Other instrument tracks
Control Enhancement Modules		
Coco-mulla (Lin et al., 2023)	Text description	Drum track, chord, melody
AIRGen (Lin et al., 2024)	Text description	Drum track, chord, melody
JASCO (Tal et al., 2024)	Text description	Drum track, chord, melody
Music ControlNet (Wu et al., 2024)	Text description	Dynamic, melody, rhythm
Jen-1 DreamStyler (Chen et al., 2024)	Text description	Reference music audio
Music Editing Methods		
MusicMagus (Zhang et al., 2024b)	Text swapping	Music audio mixture
InstructME (Han et al., 2023)	Edit instruction	Music audio mixture
Instruct-MusicGen (Zhang et al., 2024a)	Edit instruction	Music audio mixture
Loop Copilot (Zhang et al., 2023)	Edit instruction	Conversational context (music audio, text)
M^{2}UGen (Hussain et al., 2023)	Edit instruction	Conversational context (music audio, text)
ChatMusician (Yuan et al., 2024a)	Edit instruction	Conversational context (symbolic music, text)

现在的文本到音乐生成模型**不具备**必要的理解能力

可能的解决办法

- 收集解释数据(文本、视频)进行训练；或
- 利用大语言模型(LLM)的先验知识。

- 这仍然是一个**亟待解决的开放问题**

Q&A