歌曲生成与评估体系的探索

姚继珣

西北工业大学-音频与语音处理研究组（NPU@ASLP）

http://www.npu-aslp.org

背景介绍

- **音乐生成**
  - 音乐作为一种深刻的文化和艺术表达形式，具有重要的研究价值
  - 利用生成式模型来创作、编排和生成新音乐
- **应用场景**
  - 人机协同创作、游戏/直播个性化配乐、教育与疗愈

背景介绍

- **Sing Voice Synthesis**
  - Text/Lyric-To-Singing
    - DiffSinger, VISinger, StyleSinger,
  - Singing Voice Conversion
    - DiffSVC, LDM-SVC

Input: 跟着我左手右手一个慢动作，右手左手慢动作重播

- **Text-to-Music**
  - Music Generation
    - MusicLM, MusicGen, Noise2Music
  - Audio Generation
    - AudioLDM, Make-an-Audio, AudioGen, UniAudio

Input: Pop dance track with catchy melodies, tropical percussion, and upbeat rhythms, perfect for the beach

背景介绍

- **Sing Voice Synthesis**
  - Text/Lyric-To-Singing
    - DiffSinger, VISinger, StyleSinger,
  - Singing Voice Conversion
    - DiffSVC, LDM-SVC

Input: 跟着我左手右手一个慢动作，右手左手慢动作重播

只包含人声、律动感下降，缺少乐器带来的音域、氛围感受限

- **Text-to-Music**
  - Music Generation
    - MusicLM, MusicGen, Noise2Music
  - Audio Generation
    - AudioLDM, Make-an-Audio, AudioGen, Uniaudio

Input: Pop dance track with catchy melodies, tropical percussion, and upbeat rhythms, perfect for the beach

只包含伴奏，信息传递不明确、缺少记忆点、传播受限

背景介绍

- **Song Generation**
  - SongGen\SongCreator\SongEditor：同时生成人声加伴奏
    - 人声提供旋律焦点和情感叙事
    - 伴奏构建节奏基础、和声背景和氛围空间
    - 二者互为补充， 1 + 1 > 2

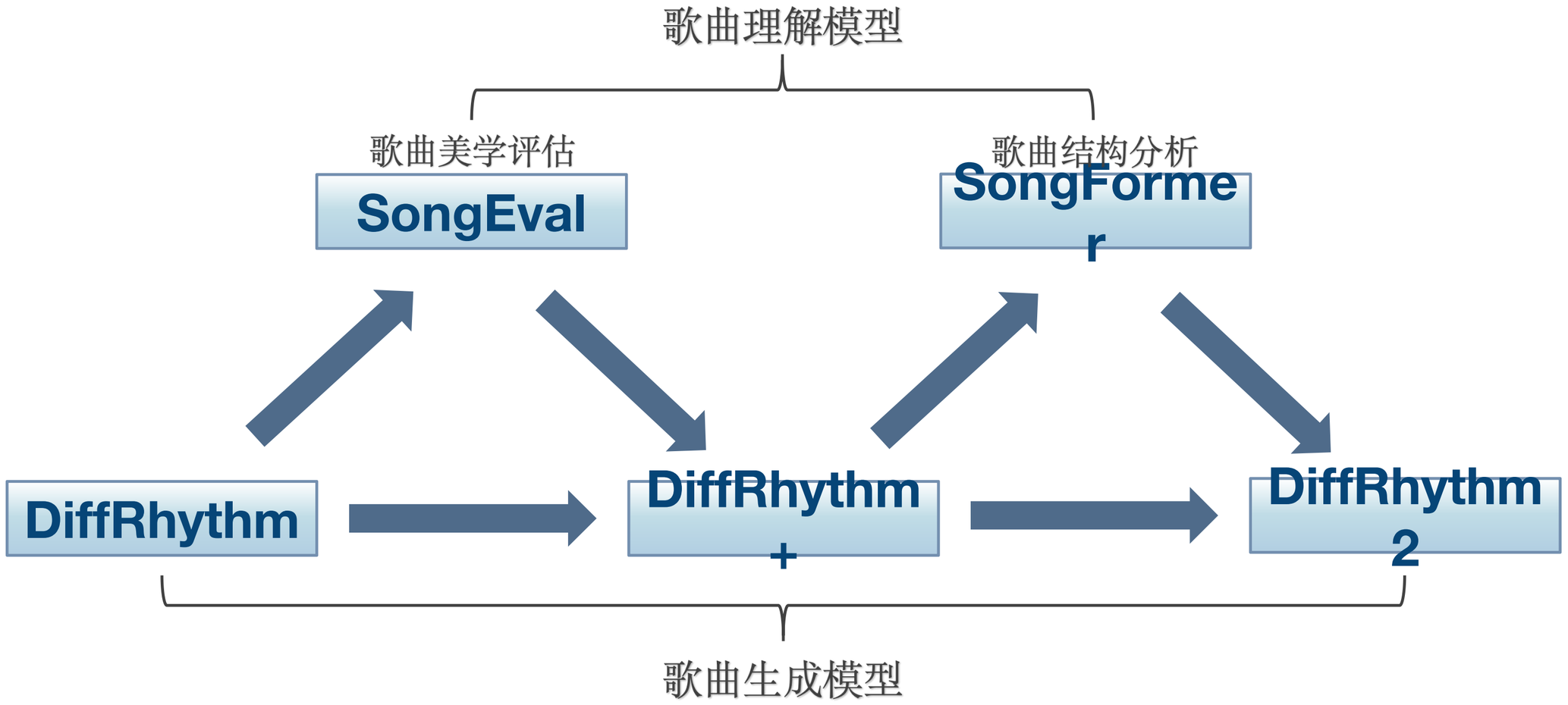
Input: Hey, it's a beautiful day. No clouds in the sky. The sun is shining bright.

背景介绍

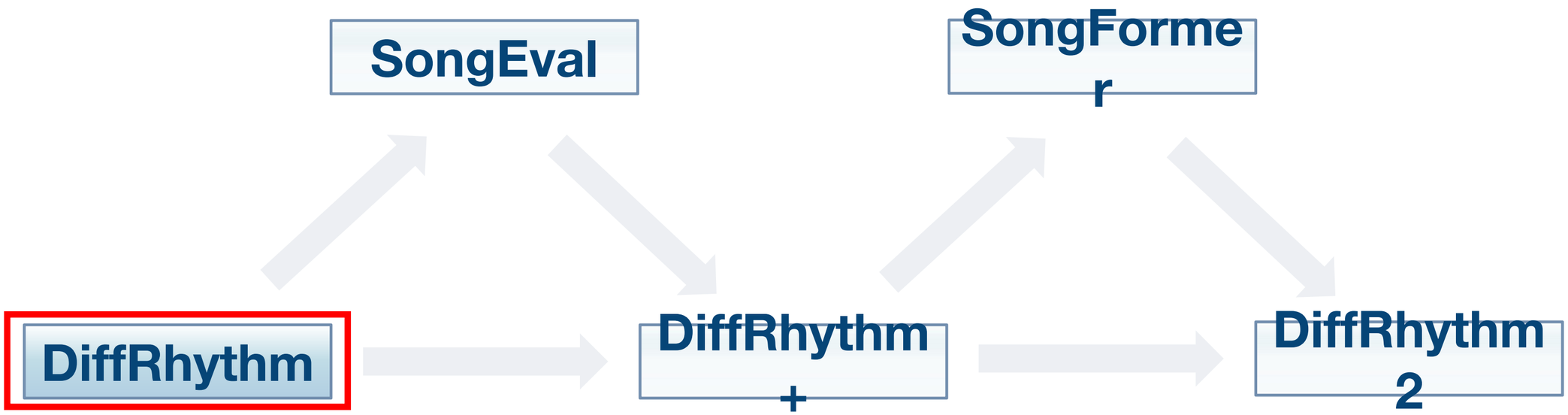
- **Song Generation**
  - SongGen\SongCreator\SongEditor：同时生成人声加伴奏
    - 人声提供旋律焦点和情感叙事
    - 伴奏构建节奏基础、和声背景和氛围空间
    - 二者互为补充， $1 + 1 > 2$
  - 时长影响，大部分时长都在30s以下
    - 缺乏结构完整性： 前奏、主歌、副歌、桥段、尾奏
    - 无法展现情感曲线：一首完整的歌代表了一个完整故事
    - 艺术表达不完整：编曲细节、递进

Input: Hey, it's a beautiful day. No clouds in the sky. The sun is shining bright.

研究路线

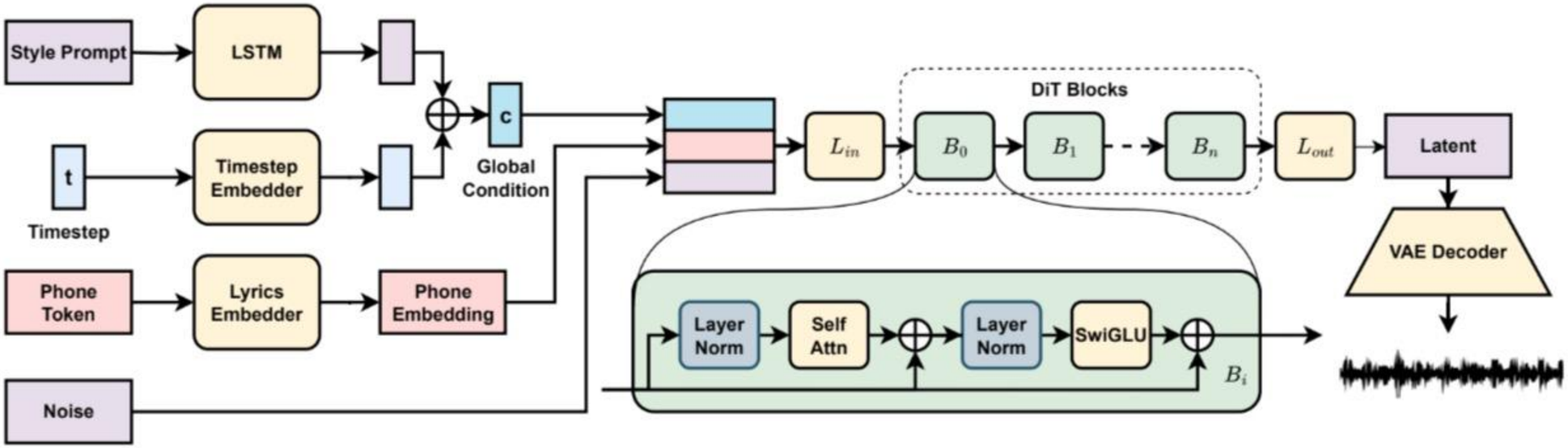

歌曲理解模型
歌曲美学评估
SongEval
歌曲结构分析
SongFormer
DiffRhythm
DiffRhythm+
DiffRhythm2
歌曲生成模型

DiffRhythm



DiffRhythm

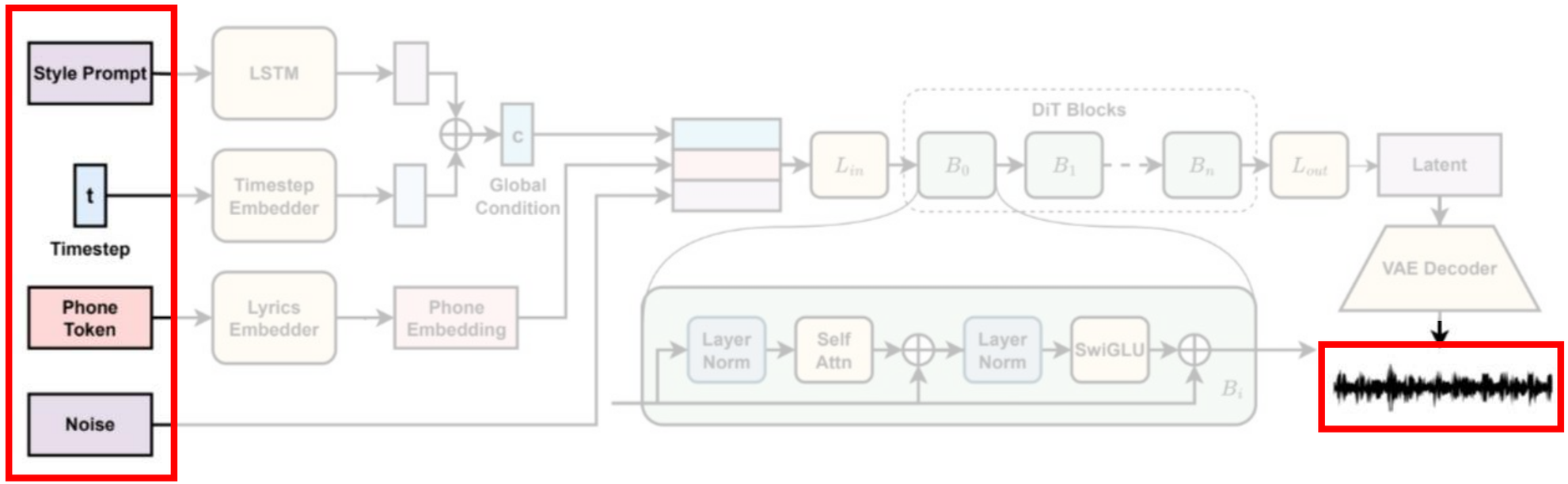
- **End-to-end full-length song generation with latent diffusion**
  - **全长生成**：能够生成长达 4分45秒 的包含人声和伴奏的完整歌曲
  - **极速生成**：生成一首完整歌曲(285秒)仅需约 **10秒**
  - **高质量输出**：44.1 kHz高保真立体声歌曲，音乐性和人声清晰度方面表现优异



Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

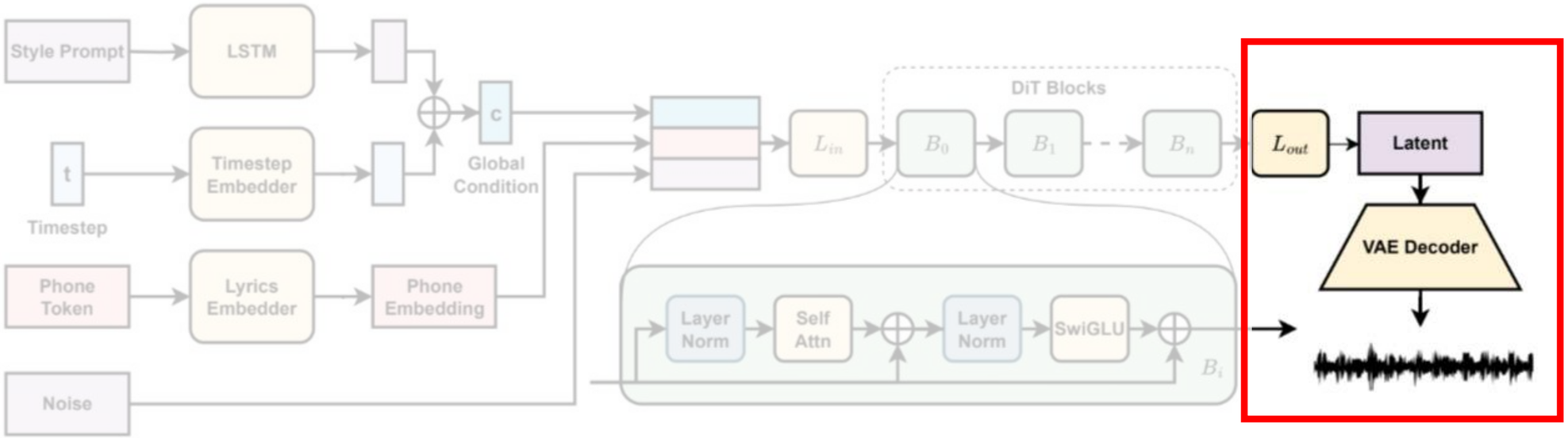
- **End-to-end full-length song generation with latent diffusion**
  - 输入:风格提示 (Style Prompt)、歌词 (Lyrics)，输出44.1kHz 的立体声完整歌曲



Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

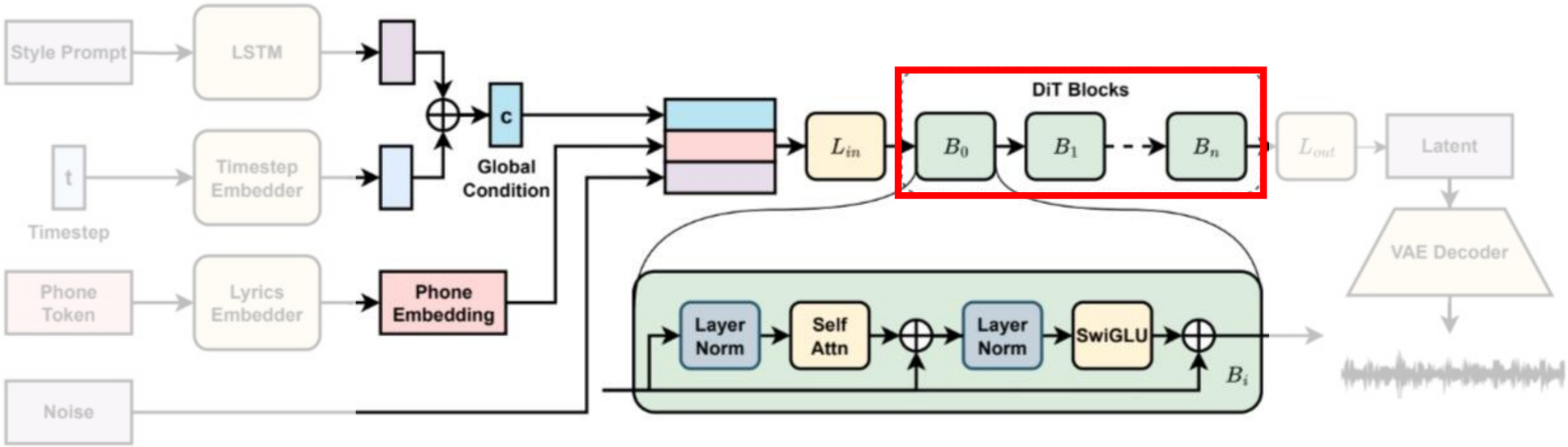
- **End-to-end full-length song generation with latent diffusion**
  - 输入:风格提示 (Style Prompt)、歌词 (Lyrics)，输出44.1kHz 的立体声完整歌曲
  - 两阶段架构
    - **VAE: 将高维的音频压缩为低帧率latent representation，缓解音频建模的序列长度难题**



Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

- **End-to-end full-length song generation with latent diffusion**
  - 输入:风格提示 (Style Prompt)、歌词 (Lyrics)，输出44.1kHz 的立体声完整歌曲
  - 两阶段架构
    - VAE: 将高维的音频压缩为低帧率latent representation，缓解音频建模的序列长度难题
    - **DiT: 以歌词和风格为condition，迭代去噪生成歌曲的latent representation**



Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

Lossy-to-Lossless VAE

- **动机**：大量音乐数据以有损的MP3格式存在
- **训练方法**：输入音频随机概率进行MP3压缩，但重建目标是原始的无损音频
  - 使VAE具备了强大的**修复能力**，能从有损输入中恢复高频细节
  - 与pre-trained共享相同的latent space，可以无缝替换

Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

Lossy-to-Lossless VAE

- **动机**：大量音乐数据以有损的MP3格式存在
- **训练方法**：输入音频随机概率进行MP3压缩，但重建目标是原始的无损音频
  - 使VAE具备了强大的**修复能力**，能从有损输入中恢复高频细节
  - 与pre-trained共享相同的latent space，可以无缝替换

	Lossless → Lossless			Lossy → Lossless					
	STOI↑	PESQ↑	MCD↓	STOI↑	PESQ↑	MCD↓	Sampling Rate	Frame Rate	Latent Channels
Music2Latent	0.584	1.448	8.796	-	-	-		10 Hz	
Stable Audio 2 VAE	0.621	1.96	8.033	-	-	-	44.1 kHz	21.5 Hz	64
DiffRhythm VAE	**0.646**	**2.235**	**8.024**	**0.639**	**2.191**	**9.319**		21.5 Hz	

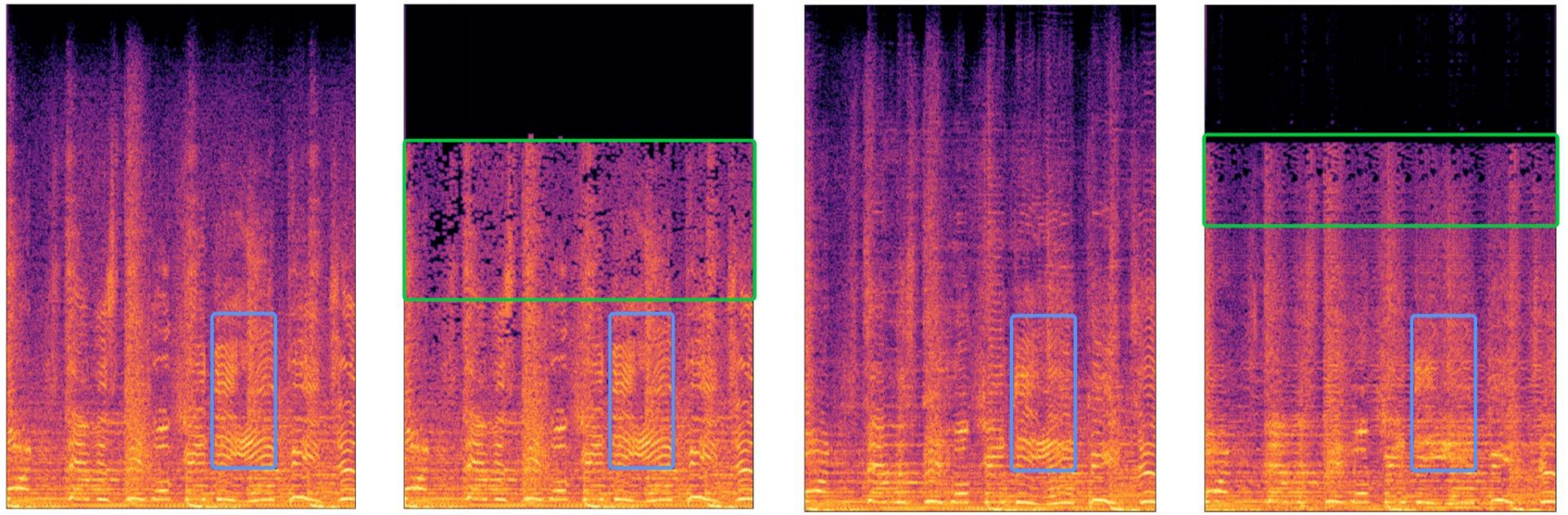
- 在无损和有损两种输入条件下，我们的VAE在各项客观指标上均优于基线模型

Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

Lossy-to-Lossless VAE

- **动机**：大量音乐数据以有损的MP3格式存在
- **训练方法**：输入音频随机概率进行MP3压缩，但重建目标是原始的无损音频
  - 使VAE具备了强大的**修复能力**，能从有损输入中恢复高频细节
  - 与pre-trained共享相同的latent space，可以无缝替换



(a) GT (Lossless) (b) GT (MP3 Compressed) (c) Proposed VAE (d) Stable Audio 2 VAE

处理MP3压缩输入时，能成功恢复被压缩掉的高频频谱信息

更好地重建人声的共振峰结构

Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

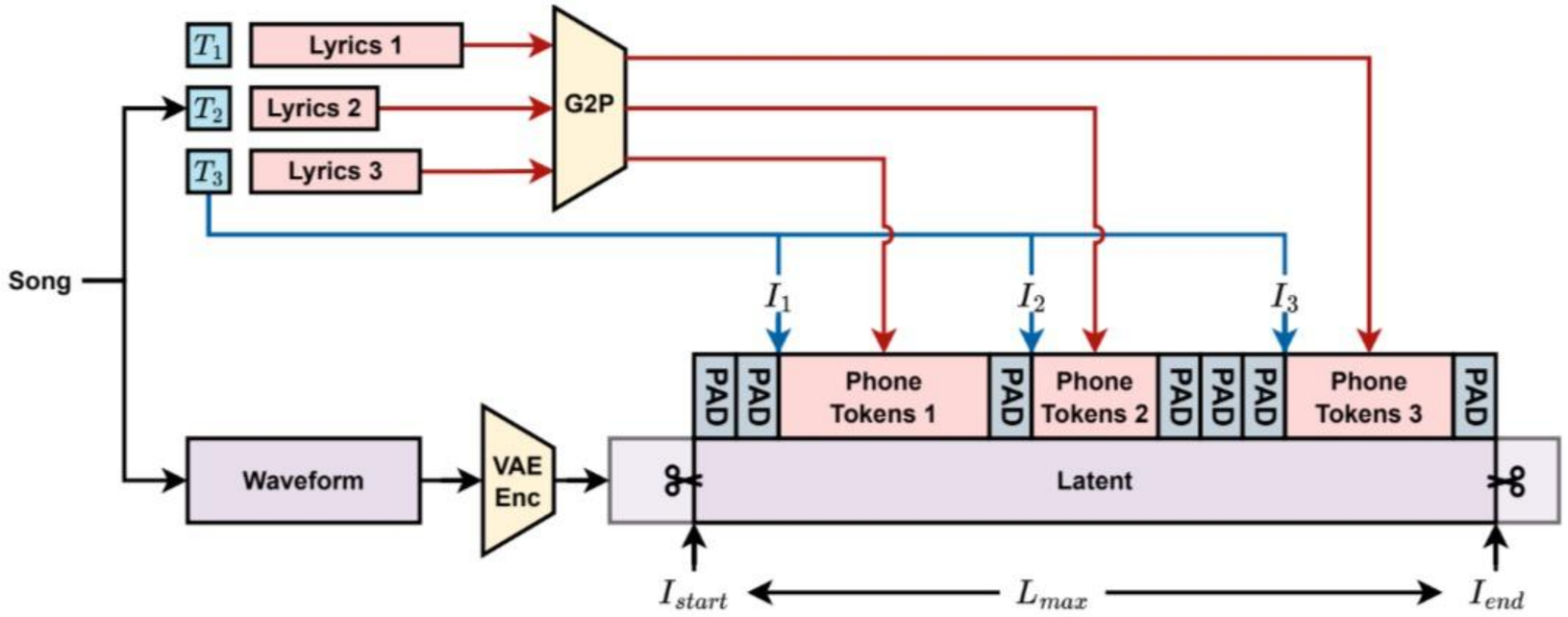
- **Lyric-latent alignment**
  - **面临的挑战**
    - **时间不连续性**：歌曲中人声片段常被长时间的乐器间奏隔开
    - **伴奏干扰**：相同的歌词在不同歌曲中对应完全不同的伴奏，增加了对齐难度

Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

Lyric-latent alignment

- Sentence-Level Alignment
  - **核心思想**：利用带有时间戳的**句子级歌词标注**，将对齐的难度降至最低
  - **实现方式**：将歌词文本转换为音素序列，根据其起始时间戳，直接对齐到与latent长度一致的序列中的相应位置



Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

实验设置

- **数据集**：约100万首歌曲（总计6万小时），包含中、英、纯音乐三种类型
- **模型配置**：
  - **VAE**：模型大小157M，下采样2048倍，帧率为21.5Hz
  - **DiT**：模型大小1.1B，16层LLaMA decoder layers
- **评估指标**：
  - **客观指标**：STOI, PESQ, MCD (用于VAE重建评估)；PER, FAD (用于歌曲生成评估)
  - **主观指标**：平均意见分 (MOS)，从音乐性、质量、清晰度三方面进行评估
- **基线系统**：SongLM
- DiffRhythm-base: 95s & DiffRhythm-full: 285s

Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

- 歌曲生成质量
  - **音质与可懂度**：DiffRhythm在歌曲质量和人声清晰度上表现更佳，音素错误率 (PER) 优于Baseline

	PER↓	FAD↓	Musicality↑	Quality↑	Intelligibility↑	Generation Length	RTF↓
GT (VAE-reconstructed)	16.14%	0.88	4.68±0.06	4.43±0.06	4.17±0.03	-	-
SongLM	21.35%	1.92	4.27±0.04	4.06±0.03	3.44±0.03	120 s	1.717
DiffRhythm-base	17.47%	2.11	4.14±0.07	4.19±0.05	3.80±0.04	95 s	0.037
DiffRhythm-full	18.02%	2.25	4.02±0.02	4.21±0.04	3.68±0.07	285 s	0.034
w/o align	-	3.16	4.07±0.05	3.04±0.02	-	95 s	0.037

Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

- 歌曲生成效率
  - **速度优势**：推理速度比SongLM快约 **50倍** (RTF < 0.04 vs 1.717)，突显了非自回归扩散模型的巨大效率优势

	PER↓	FAD↓	Musicality↑	Quality↑	Intelligibility↑	Generation Length	RTF↓
GT (VAE-reconstructed)	16.14%	0.88	4.68±0.06	4.43±0.06	4.17±0.03	-	-
SongLM	21.35%	1.92	4.27±0.04	4.06±0.03	3.44±0.03	120 s	1.717
DiffRhythm-base	17.47%	2.11	4.14±0.07	4.19±0.05	3.80±0.04	95 s	0.037
DiffRhythm-full	18.02%	2.25	4.02±0.02	4.21±0.04	3.68±0.07	285 s	0.034
w/o align	-	3.16	4.07±0.05	3.04±0.02	-	95 s	0.037

Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

DiffRhythm

- 如何评估生成的歌曲
  - **客观指标**：基于度量或某些距离
  - **主观指标**：成本

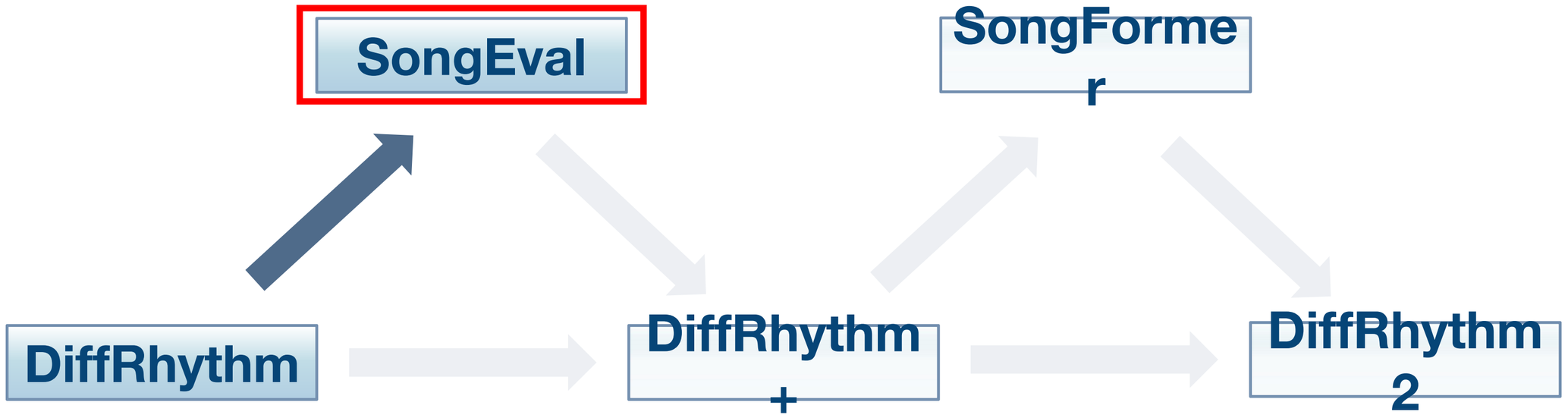
Method	PER(%)↓	FAD↓	Musicality↑	Quality↑
GT(restore)	5.51	0.86	3.63	3.59
SongLM	20.63	**1.99**	2.95	3.06
SongEditor	**18.33**	2.24	**3.02**	**3.19**

Model	FAD ↓	MCD ↓	Musicality ↑	Similarity ↑
Ground Truth	-	-	4.04 ± 0.06	3.79 ± 0.09
MusicGen	**1.90**	9.78	3.46 ± 0.11	3.27 ± 0.11
SongCreator	2.06	**8.44**	$\mathbf{4.01 \pm 0.07}$	$\mathbf{3.82 \pm 0.08}$

	PER↓	FAD↓	Musicality↑	Quality↑	Intelligibility↑	Generation Length	RTF↓
GT (VAE-reconstructed)	16.14%	0.88	4.68±0.06	4.43±0.06	4.17±0.03	-	-
SongLM	21.35%	1.92	4.27±0.04	4.06±0.03	3.44±0.03	120 s	1.717
DiffRhythm-base	17.47%	2.11	4.14±0.07	4.19±0.05	3.80±0.04	95 s	0.037
DiffRhythm-full	18.02%	2.25	4.02±0.02	4.21±0.04	3.68±0.07	285 s	0.034
w/o align	-	3.16	4.07±0.05	3.04±0.02	-	95 s	0.037

Z Ning, H Chen, Y Jiang, C Hao, G Ma, S Wang, J Yao, L Xie. DiffRhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. https://arxiv.org/pdf/2503.01183

SongEval



SongEval

背景动机

- **核心问题:** 如何有效评估AI生成歌曲的“美学质量”是一个巨大挑战
- **挑战来源:** 歌曲欣赏具有高度的主观性和多维度特性
- **现有方法的局限:**
  - **客观指标不足:** 基于嵌入距离等客观指标无法捕捉人类对音乐情感表达、连贯性的主观感知
  - **现有数据集的空白:** 已有的主观评估数据集主要关注语音生成、或在歌曲完整性、评估维度和数据规模上存在局限，例如仅关注伴奏或短小的音乐片段
- **我们的目标:** 创建一个更具可解释性、与专业音乐标准对齐的歌曲生成评估基准

J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

SongEval

- **数据集介绍**
  - **规模:** 包含2,399首完整歌曲，总时长超过140小时

		MusicEval	AES-Natural	SongEval
Language		-	EN	EN & ZH
Total Hours		16.67	29.44	140.32
Utt. Average Duration (min)		0.36	1.77	3.51
Components		Accompaniments only	Accompaniments +Vocal	Accompaniments +Vocal
Annotation Aspects	Musicality	✓	✓	✓
	Clarity	✗	✓	✓
	Naturalness	✗	✓	✓
	Memorability	✗	✗	✓
	Coherence	✗	✗	✓

J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

SongEval

数据集介绍

- **规模:** 包含2,399首完整歌曲，总时长超过140小时
- **语言与风格:** 覆盖英语和中文歌曲，涵盖九大主流音乐流派
- **专业标注:** 由16位具备专业音乐背景的标注员进行评分

		MusicEval	AES-Natural	SongEval
Language		-	EN	EN & ZH
Total Hours		16.67	29.44	140.32
Utt. Average Duration (min)		0.36	1.77	3.51
Components		Accompaniments only	Accompaniments +Vocal	Accompaniments +Vocal
Annotation Aspects	Musicality	✓	✓	✓
	Clarity	✗	✓	✓
	Naturalness	✗	✓	✓
	Memorability	✗	✗	✓
	Coherence	✗	✗	✓

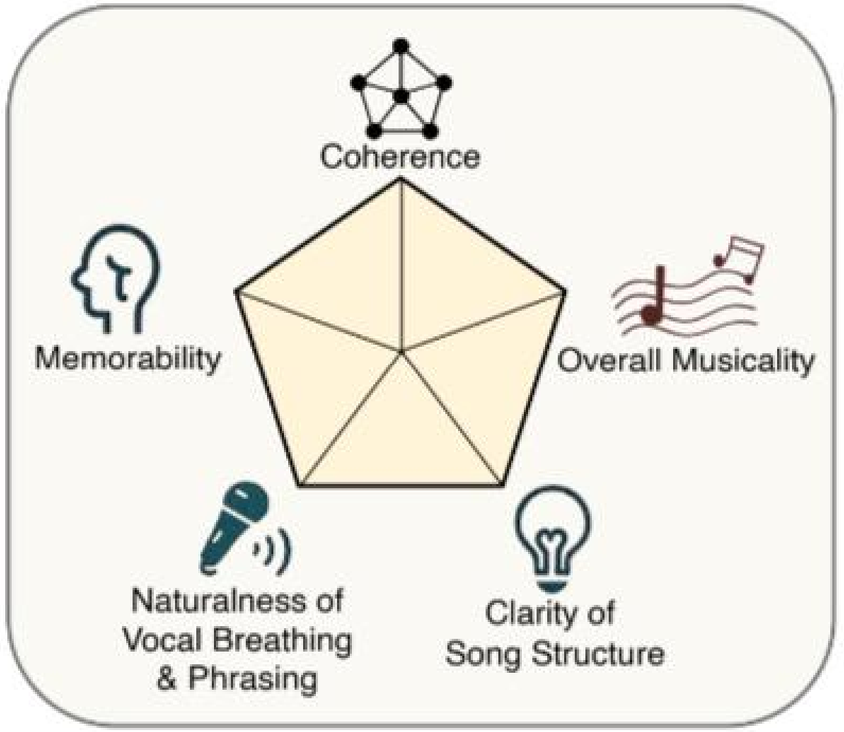
Genre	Language	Duration (Hours)	Samples	Gender Duration Ratio
Pop	ZH	15.74	284	52% / 48%
	EN	11.28	175	57% / 43%
Rock	ZH	5.29	91	61% / 39%
	EN	14.33	233	64% / 36%
Electronic	ZH	6.78	123	55% / 45%
	EN	6.96	126	50% / 50%
Blues	ZH	3.62	60	66% / 34%
	EN	8.70	135	74% / 26%
World Music	ZH	5.21	103	59% / 41%
	EN	6.34	125	55% / 45%
Hip-hop/Rap	ZH	4.35	83	65% / 35%
	EN	3.31	62	79% / 21%
Country	ZH	4.19	84	61% / 39%
	EN	4.74	71	53% / 47%
Jazz	ZH	4.13	69	50% / 50%
	EN	4.09	64	60% / 40%
Classical	ZH	3.71	62	43% / 57%
	EN	2.77	43	32% / 68%
Others	ZH	9.58	134	75% / 25%
	EN	15.21	272	56% / 44%
All	-	**140.32**	**2399**	**60% / 40%**

J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

SongEval

评估维度

- **整体连贯性:** 评估歌曲不同部分（主歌、副歌）在音乐和情感上的连续性
- **记忆点:** 衡量歌曲是否拥有易于记忆的旋律、节奏或歌词片段
- **人声自然度:** 评估演唱中的呼吸控制和乐句处理是否自然流畅。
- **结构清晰度:**衡量歌曲结构（如主歌-副歌-桥段）是否清晰、合乎逻辑
- **整体音乐性:** 对旋律、乐器以及人声与伴奏融合度的综合听感评价

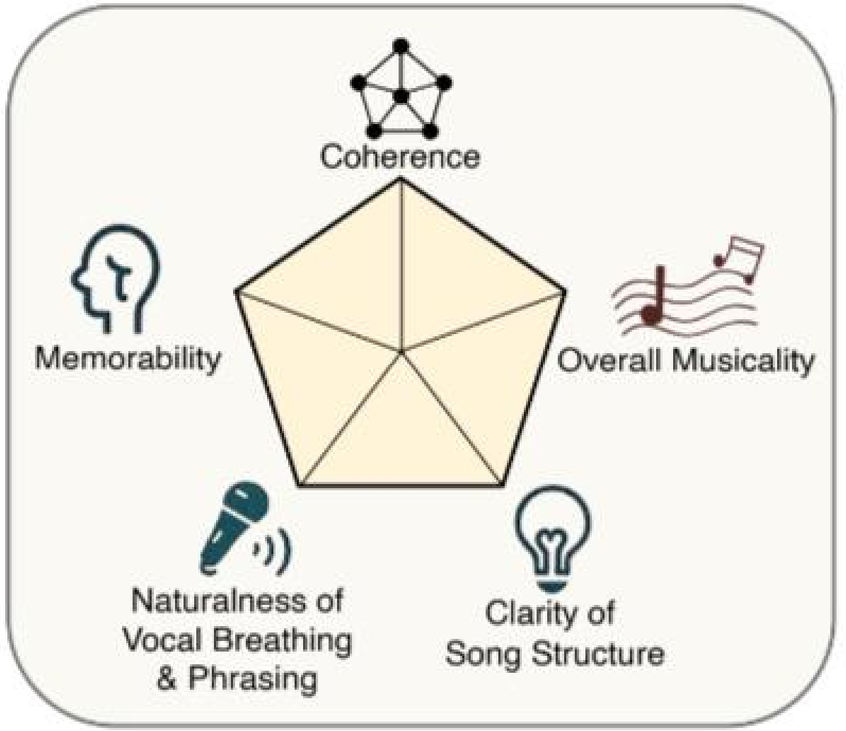


J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

SongEval

评估维度

- **整体连贯性:** 评估歌曲不同部分（主歌、副歌）在音乐和情感上的连续性
- **记忆点:** 衡量歌曲是否拥有易于记忆的旋律、节奏或歌词片段
- **人声自然度:** 评估演唱中的呼吸控制和乐句处理是否自然流畅。
- **结构清晰度:**衡量歌曲结构（如主歌-副歌-桥段）是否清晰、合乎逻辑
- **整体音乐性:** 对旋律、乐器以及人声与伴奏融合度的综合听感评价



《淬炼》

在风雨中不畏惧,我们从不低头
跨越崇山峻岭,心向着光明,前行不休
站在这片大地上，坚定的眼神
正义与梦想，照亮了明天
烈火与冰雪交织,心中燃烧信念
前路虽远，脚步不止
用坚韧撑起这片蓝天
我们共同抬头，面向未来的晨曦

重庆的山河，心中最自豪
无论风雨，依然不退缩脚步高
岁月磨砺，誓言铸成钢铁牢
携手共进，向着明天奔跑

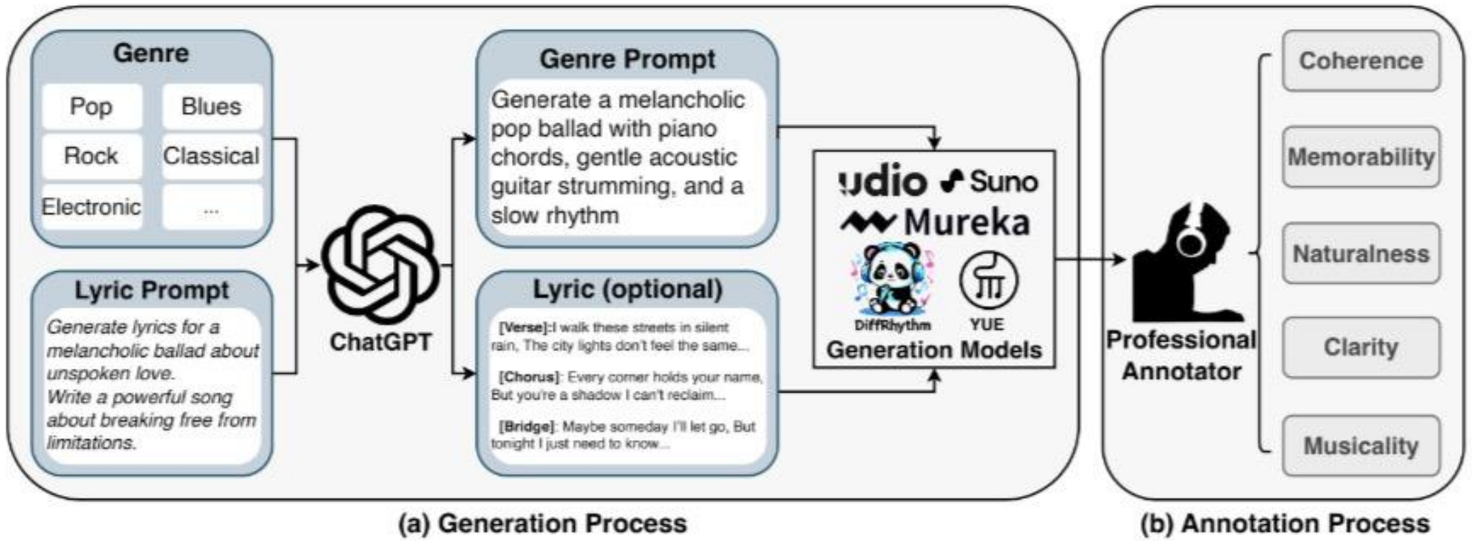
每一道霞光，照亮心头光耀
每一声呐喊，坚定步伐不放掉
让我们与时代并肩，追逐那份希望
重庆，永远在心中闪亮如星耀

J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

SongEval

数据收集

- **两阶段生成流程:**
  - **提示语生成:** 利用ChatGPT生成与九大音乐流派对齐的歌词及风格提示（描述情绪、风格和乐器）
  - **歌曲合成:** 将生成的提示输入到五个主流歌曲生成模型（如Suno, Udio, DiffRhythm等）中，合成风格多样的歌曲
- **数据标注**：16位专业音乐背景标注人员根据五个美学维度，在1-5分的范围内进行打分



J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

SongEval

- **实验对比**
  - **数据集:**
    - **训练集:** 2,199首生成歌曲
    - **评估集:** 200首未见过的生成歌曲 + 50首真实的非版权歌曲
  - **评估指标:**
    - Mean Squared Error (MSE)
    - Linear Correlation Coefficient (LCC)
    - Spearman Rank Correlation Coefficient (SRCC)
    - Kendall's Tau Rank Correlation (KTAU)

	System	Utterance-level				System-level			
		MSE↓	LCC↑	SRCC↑	KATU↑	MSE↓	LCC↑	SRCC↑	KATU↑
Coherence	MOSNet-based	0.339	0.882	0.854	0.679	0.187	0.923	0.904	0.751
	LDNet-based	0.421	0.882	0.860	0.684	0.238	0.948	0.934	0.793
	SSL-based	0.237	0.900	0.882	0.719	0.088	0.959	**0.962**	**0.860**
	UTMOS-based	**0.195**	**0.917**	**0.898**	**0.741**	**0.073**	**0.962**	0.954	0.844
Memorability	MOSNet-based	0.360	0.874	0.851	0.672	0.206	0.919	0.889	0.727
	LDNet-based	0.547	0.867	0.846	0.671	0.340	0.936	0.920	0.776
	SSL-based	0.276	0.897	0.891	0.723	0.104	0.951	0.945	0.810
	UTMOS-based	**0.241**	**0.910**	**0.901**	**0.739**	**0.096**	**0.955**	**0.958**	**0.849**
Naturalness	MOSNet-based	0.406	0.843	0.818	0.634	0.203	0.923	0.901	0.740
	LDNet-based	0.449	0.867	0.855	0.688	0.247	0.924	0.911	0.763
	SSL-based	0.243	0.896	0.885	0.718	**0.079**	0.955	**0.942**	**0.820**
	UTMOS-based	**0.219**	**0.909**	**0.896**	**0.734**	0.081	**0.957**	0.941	0.809
Clarity	MOSNet-based	0.354	0.876	0.855	0.675	0.186	0.925	0.919	0.757
	LDNet-based	0.450	0.862	0.853	0.677	0.249	0.925	0.916	0.773
	SSL-based	0.235	0.903	0.889	0.720	**0.085**	**0.952**	**0.951**	**0.824**
	UTMOS-based	**0.221**	**0.908**	**0.894**	**0.728**	0.091	0.951	0.939	0.804
Musicality	MOSNet-based	0.337	0.877	0.854	0.677	0.168	0.934	0.928	0.784
	LDNet-based	0.466	0.881	0.861	0.689	0.262	0.944	0.927	0.779
	SSL-based	0.220	0.908	0.893	0.733	**0.066**	0.965	**0.970**	**0.864**
	UTMOS-based	**0.203**	**0.916**	**0.901**	**0.745**	0.072	**0.966**	0.969	0.859

J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

SongEval

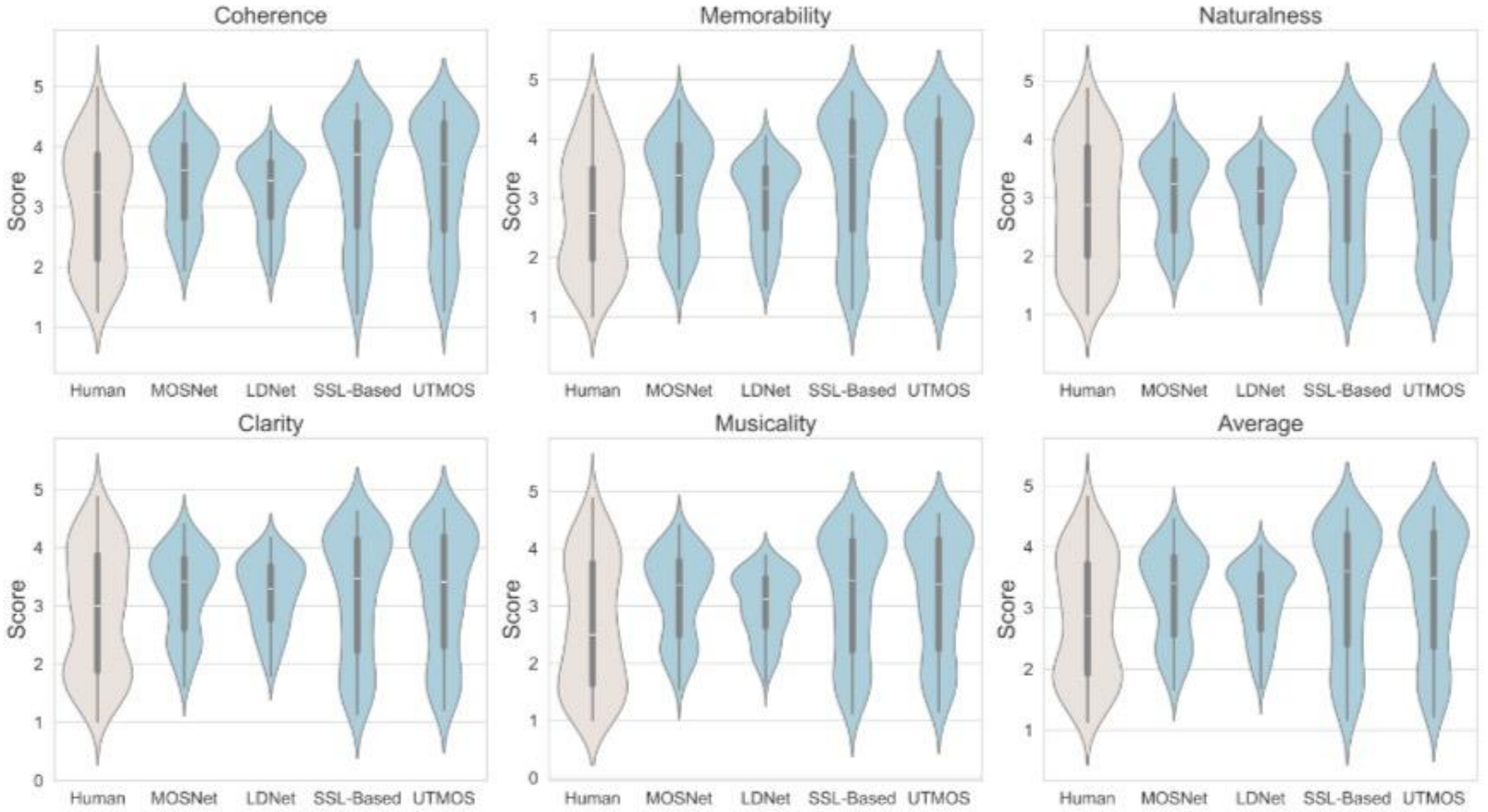
- **实验对比**
  - 所有在SongEval上训练的模型都能较好地预测多维度的美学分数。
  - 基于自监督学习（SSL-based）和UTMOS-based的模型表现更优

	System	Utterance-level				System-level			
		MSE↓	LCC↑	SRCC↑	KATU↑	MSE↓	LCC↑	SRCC↑	KATU↑
Coherence	MOSNet-based	0.339	0.882	0.854	0.679	0.187	0.923	0.904	0.751
	LDNet-based	0.421	0.882	0.860	0.684	0.238	0.948	0.934	0.793
	SSL-based	0.237	0.900	0.882	0.719	0.088	0.959	**0.962**	**0.860**
	UTMOS-based	**0.195**	**0.917**	**0.898**	**0.741**	**0.073**	**0.962**	0.954	0.844
Memorability	MOSNet-based	0.360	0.874	0.851	0.672	0.206	0.919	0.889	0.727
	LDNet-based	0.547	0.867	0.846	0.671	0.340	0.936	0.920	0.776
	SSL-based	0.276	0.897	0.891	0.723	0.104	0.951	0.945	0.810
	UTMOS-based	**0.241**	**0.910**	**0.901**	**0.739**	**0.096**	**0.955**	**0.958**	**0.849**
Naturalness	MOSNet-based	0.406	0.843	0.818	0.634	0.203	0.923	0.901	0.740
	LDNet-based	0.449	0.867	0.855	0.688	0.247	0.924	0.911	0.763
	SSL-based	0.243	0.896	0.885	0.718	**0.079**	0.955	**0.942**	**0.820**
	UTMOS-based	**0.219**	**0.909**	**0.896**	**0.734**	0.081	**0.957**	0.941	0.809
Clarity	MOSNet-based	0.354	0.876	0.855	0.675	0.186	0.925	0.919	0.757
	LDNet-based	0.450	0.862	0.853	0.677	0.249	0.925	0.916	0.773
	SSL-based	0.235	0.903	0.889	0.720	**0.085**	**0.952**	**0.951**	**0.824**
	UTMOS-based	**0.221**	**0.908**	**0.894**	**0.728**	0.091	0.951	0.939	0.804
Musicality	MOSNet-based	0.337	0.877	0.854	0.677	0.168	0.934	0.928	0.784
	LDNet-based	0.466	0.881	0.861	0.689	0.262	0.944	0.927	0.779
	SSL-based	0.220	0.908	0.893	0.733	**0.066**	0.965	**0.970**	**0.864**
	UTMOS-based	**0.203**	**0.916**	**0.901**	**0.745**	0.072	**0.966**	0.969	0.859

J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

SongEval

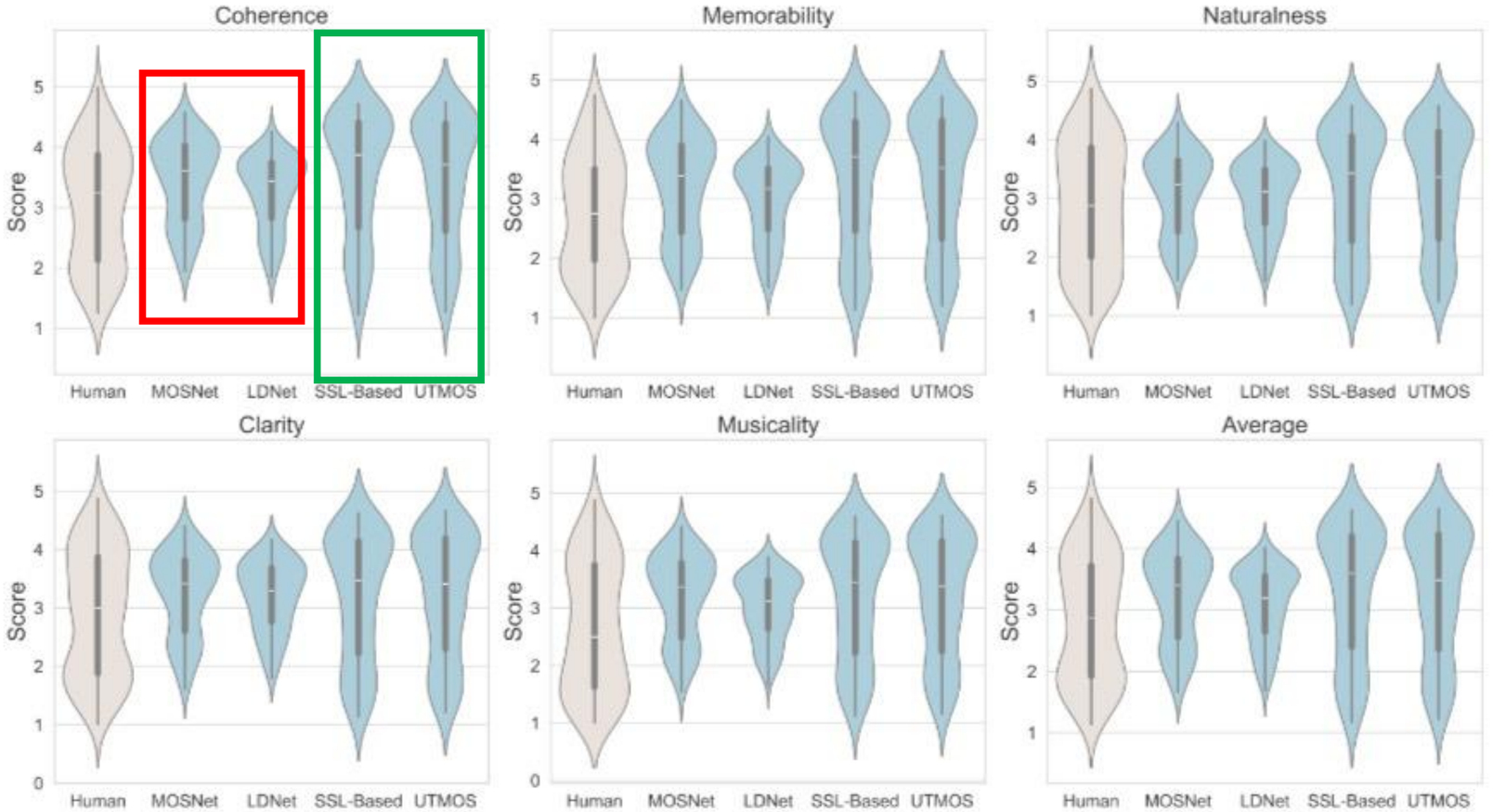
- 实验对比



J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

SongEval

- 实验对比



J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

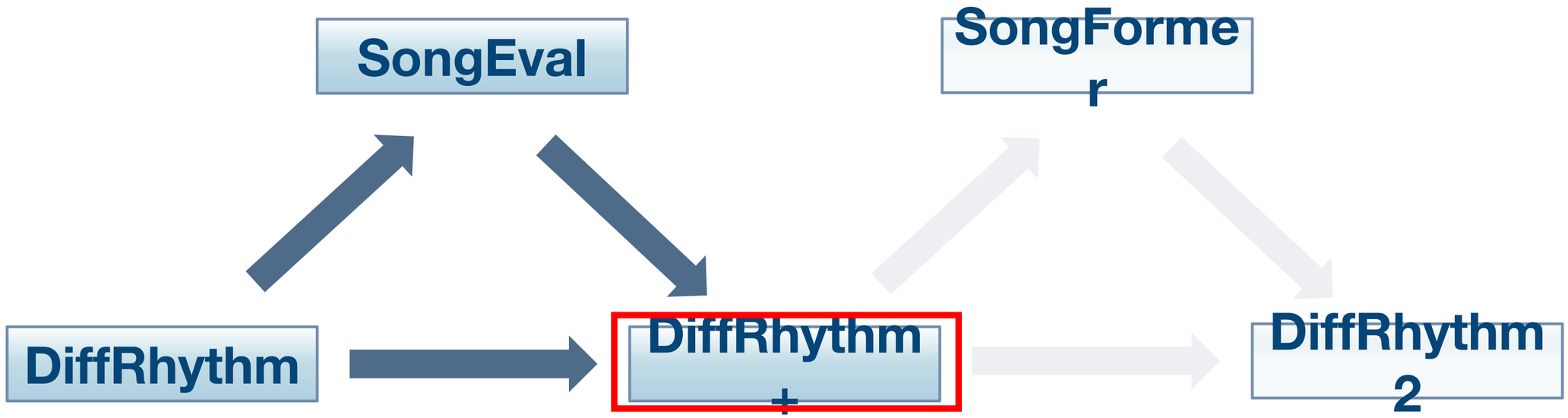
SongEval

- 对比其他客观指标
  - Audiobox-Aesthetic
    - Production Quality (PQ), Production Complexity (PC), Content Enjoyment (CE), Content Usefulness (CU)
  - Song-level vocal range

	CE	CU	PC	PQ	Vocal Range	Aesthetic (Ours)
Coherence	0.631	0.679	0.433	0.636	0.657	0.917
Memorability	0.605	0.654	0.400	0.625	0.667	0.910
Naturalness	0.602	0.645	0.396	0.616	0.739	0.909
Clarity	0.574	0.627	0.394	0.603	0.694	0.908
Musicality	0.608	0.653	0.388	0.622	0.751	0.916
Average	0.614	0.662	0.408	0.630	0.702	0.912

J Yao, G Ma, H Xue, H Chen, C Hao, Y Jiang, H Liu, R Yuan, J Xu, W Xue, H Liu, L Xie. SongEval: A Benchmark Dataset for Song Aesthetics Evaluation. https://arxiv.org/pdf/2503.01183

DiffRhythm+



DiffRhythm+

- **DiffRhythm局限性**
  - **数据不平衡**:
    - 中英文歌曲比例失衡（约为3:6），导致中文歌曲生成效果较差
    - 易出现歌词重复或遗漏
  - **风格控制受限**: 仅支持音频作为风格提示，控制的灵活性和精度不足
  - **整体音乐性**:生成的音乐在编曲复杂性、表现力上有所欠缺，且偶尔出现清晰度和连贯性问题

H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

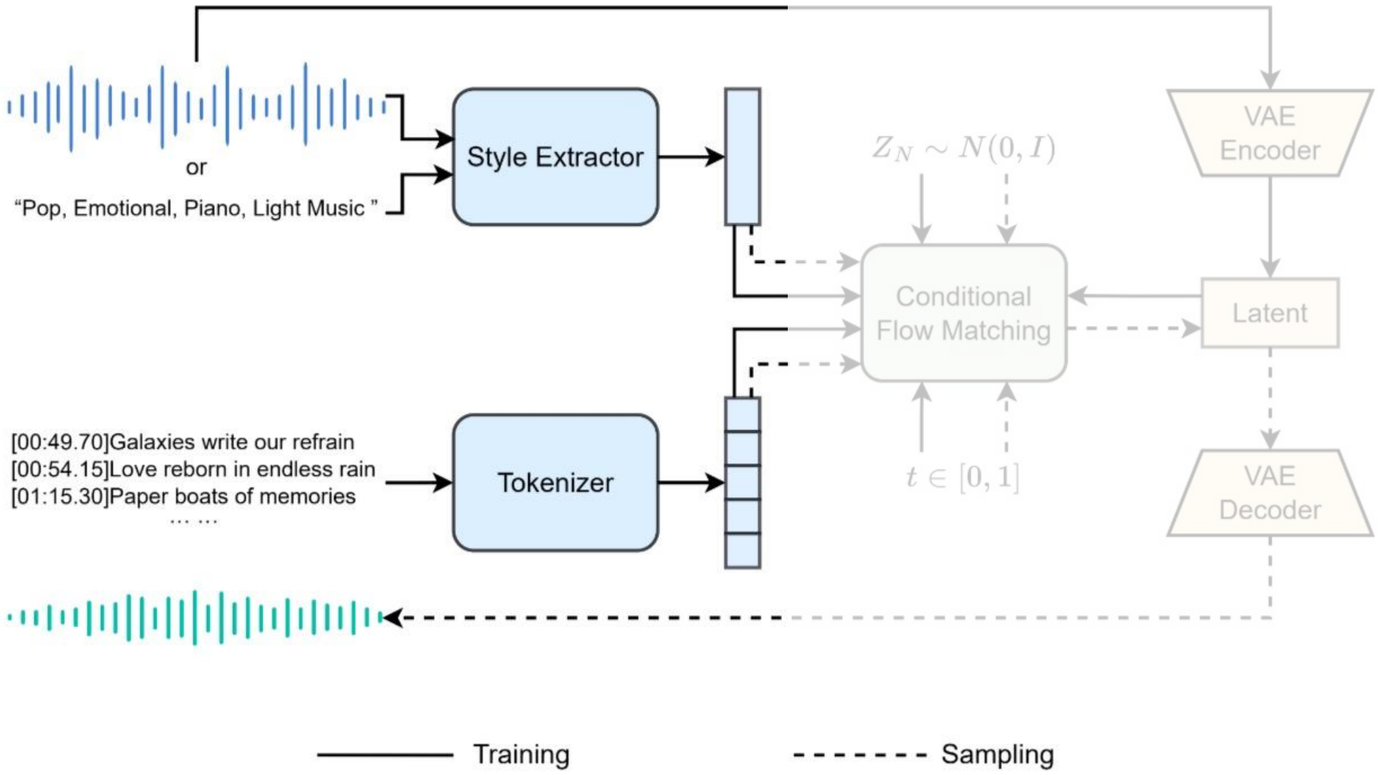
- **DiffRhythm局限性**
  - 数据不平衡:
    - 中英文歌曲比例失衡（约为3:6），导致中文歌曲生成效果较差
    - 易出现歌词重复或遗漏
  - **风格控制受限**: 仅支持音频作为风格提示，控制的灵活性和精度不足
  - **整体音乐性**:生成的音乐在编曲复杂性、表现力上有所欠缺，且偶尔出现清晰度和连贯性问题
- **DiffRhythm+贡献**
  - **大规模均衡数据**: 更大规模、更均衡的训练数据，提升音乐质量和歌词对齐准确性
  - **多模态风格控制**: 引入MuLan作为风格提取器，支持**文本**和**音频**作为风格输入
  - **强化学习优化**: 首次将DPO应用于歌曲生成，利用音乐美学评估模型，使生成结果更符合人类审美偏好

H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

模型结构

- **风格输入**: 文本或音频通过**MuLan**编码为风格表征
- **歌词输入**: 歌词编码为音素序列嵌入

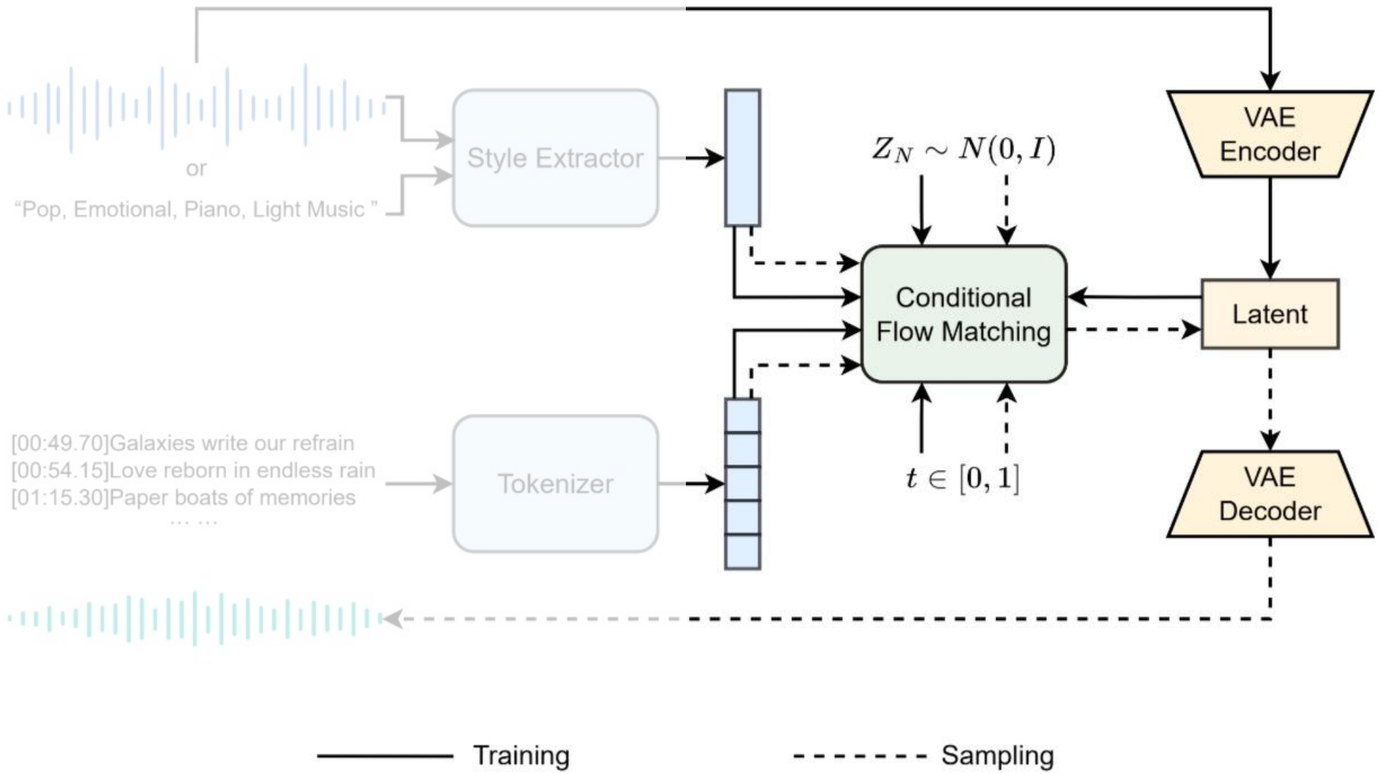


H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

模型结构

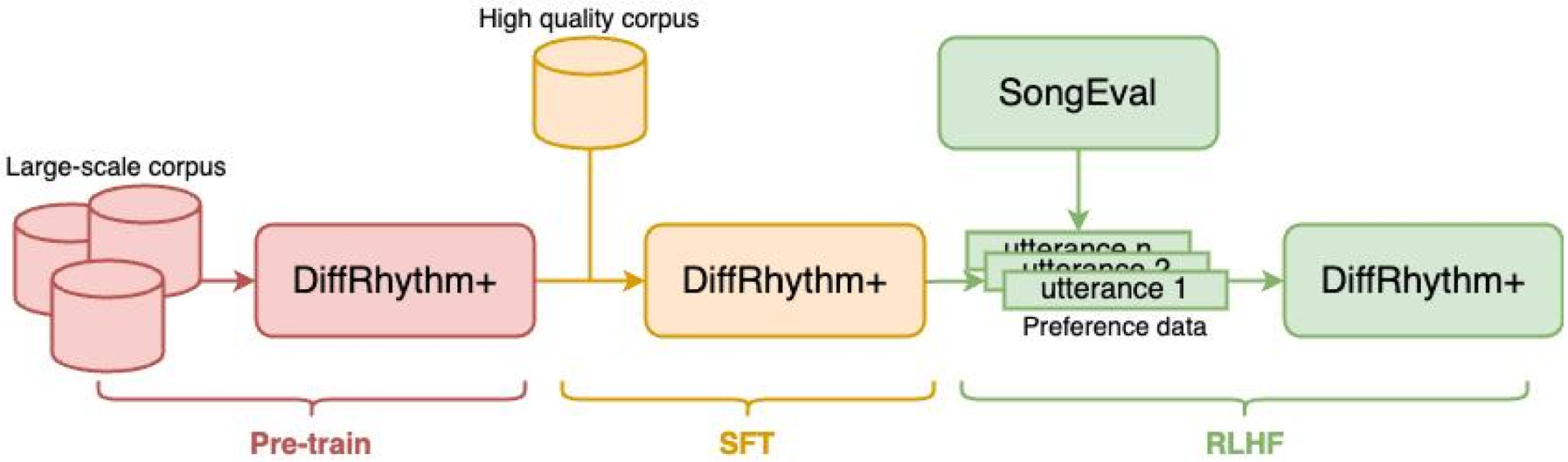
- **VAE模块**: 负责将音频波形编码为隐空间表示，以及从隐空间解码回音频
- **Flow Matching模块**:基于DiT结构，以非自回归的方式对音乐的隐空间分布进行建模



H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

- 如何让模型知道什么样的音乐更好听
  - **基于人类美学感知的偏好对齐:** 传统的训练目标（如L2 loss）并不直接等同于人类的“音乐审美”
    - **构建偏好数据集:** 使用SongEval为生成歌曲打分，挑选chosen-reject pair
    - **偏好对齐:** DPO算法在SFT模型基础上进行后训练



H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

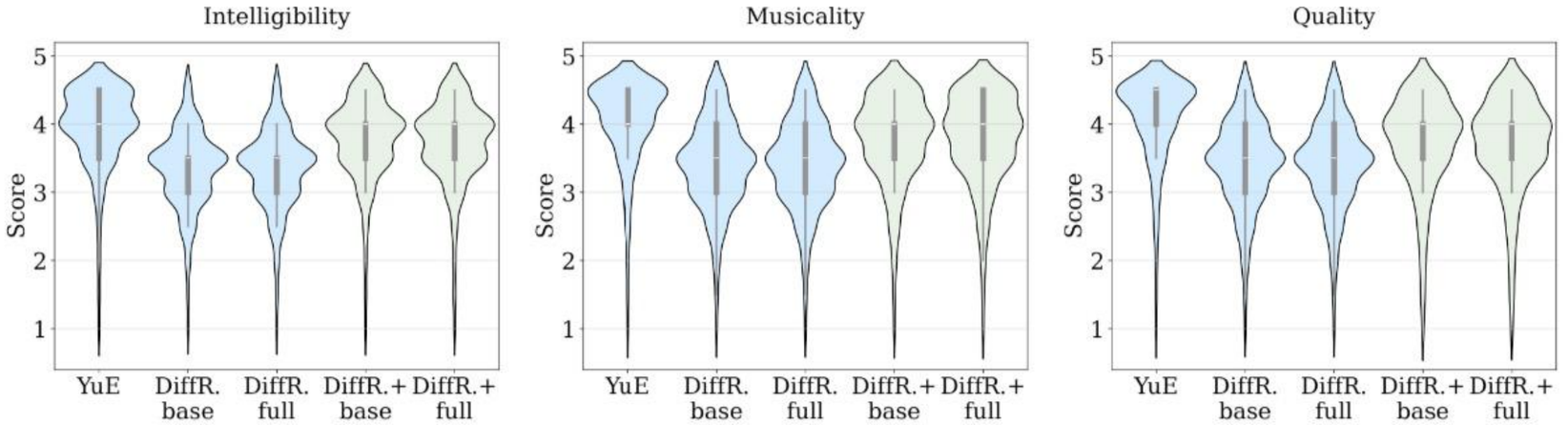
实验设置

- **训练数据**:
  - 12万小时歌曲数据，中/英/纯音乐比例均衡 (2:2:1)
  - 从预训练数据中筛选出2.5万小时高质量歌曲数据
- **基线系统**: SongLM, YuE, DiffRhythm
- **评估指标**:
  - **客观指标**: KL散度、FAD（分布匹配），PER（歌词清晰度），CLaMP 3（对齐度），RTF（计算效率）
  - **主观指标**: MOS人工听测（音乐性、音质、清晰度）
  - **美学指标**: SongEval, Audiobox-aesthetic

H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

- 实验对比
  - 客观&主观对比



Metric	Distri. Match		Align.	Intelli.	Comp. Eff.
	KL↓	FAD↓	CLaMP 3↑	PER↓	RTF↓
SongLM	0.736	1.921	0.101	21.35%	1.717
YuE	**0.372**	**1.624**	**0.240**	15.14%	10.385
DiffR.+ base	0.488	1.835	0.152	**14.85%**	0.036
DiffR. base	0.723	2.113	0.103	17.47%	**0.034**
DiffR.+ full	0.512	1.872	0.165	14.96%	0.039
DiffR. full	0.741	2.251	0.112	18.02%	0.037

Metric Model	AudioBox-aesthetic				SongEval				
	CE↑	CU↑	PC↑	PQ↑	Coh↑	Mem↑	NVBP↑	CSS↑	OM↑
GT	7.159	7.578	5.975	7.787	4.078	4.048	3.786	3.780	3.823
SongLM	6.538	7.232	5.614	7.412	2.588	2.474	2.278	2.233	2.412
YuE	7.115	7.543	6.280	7.894	3.722	**3.714**	**3.288**	3.418	**3.458**
DiffR.+ base	7.345	**7.769**	**6.283**	**7.978**	**3.768**	3.682	3.203	3.414	3.287
DiffR. base	6.224	7.163	5.351	7.421	2.634	2.417	2.215	2.261	2.281
DiffR.+ full	**7.443**	7.518	6.224	7.852	3.738	3.675	3.153	**3.421**	3.395
DiffR. full	6.481	6.912	5.441	7.441	2.612	2.531	2.334	2.185	2.331

H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

❖ 消融实验

数据规模与平衡
- 增加数据量能显著提升音乐质量
- 中英文数据平衡对性能的提升效果明显

Ablation Category	Setting	[1–2]↓	[2–3]↓	[3–4]↑	[4–5]↑	Mean↑	[3–5] %↑
Data Scale & Balance	60,000 hours (ZH:EN = 1:1)	33 / 32	32 / 28	35 / 40	0	2.31	37.5%
	120,000 hours (ZH:EN = 1:1)	2 / 1	58 / 51	40 / 48	0	**2.80**	44.0%
	120,000 hours (ZH:EN = 1:2)	4 / 1	58 / 40	38 / 59	0	2.75	48.5%
Style Conditioning	~~Audio Latent + LSTM~~	~~14~~	~~117~~	~~69~~	~~0~~	~~2.63~~	~~34.5%~~
	Audio MuLan (Audio Prompt)	4	87	107	2	2.85	54.5%
	Mixed MuLan (Audio Prompt)	4	88	107	1	**2.88**	54.0%
	T5 Embedding + LSTM	30	119	51	0	2.46	25.5%
	Audio MuLan (Text Prompt)	71	108	21	0	2.23	10.5%
	Mixed MuLan (Text Prompt)	11	106	83	0	**2.75**	41.5%
DPO & Training Stage	DiffR.	68	91	40	1	2.25	20.5%
	DiffR.$^+$ Pretrain	3	109	88	0	2.80	44.0%
	DiffR.$^+$ SFT	3	78	110	9	2.86	59.5%
	DiffR.$^+$ DPO	1	36	145	18	**3.19**	81.5%
DPO Winner	GT Winner	7	64	127	2	2.94	64.5%
	Generated Winner	1	36	145	18	**3.19**	81.5%
DPO Training Epochs	4 Epoch	4	47	139	10	3.11	74.5%
	8 Epoch	1	36	145	18	**3.19**	81.5%
	12 Epoch	2	35	138	25	3.16	81.5%
Win-Lose Score Gap	Gap = 0	4	38	138	20	3.15	79.0%
	Gap = 0.4	1	36	145	18	**3.19**	81.5%
	Gap = 0.8	4	62	121	13	3.04	67.0%

H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

消融实验

MuLan风格控制:
- 效果远超原始的音频latent+LSTM
- 同时使用文本和音频进行训练（Mixed MuLan）能达到最佳效果

Ablation Category	Setting	[1–2]↓	[2–3]↓	[3–4]↑	[4–5]↑	Mean↑	[3–5] %↑
Data Scale & Balance	60,000 hours (ZH:EN = 1:1)	33 / 32	32 / 28	35 / 40	0	2.31	37.5%
	120,000 hours (ZH:EN = 1:1)	2 / 1	58 / 51	40 / 48	0	**2.80**	44.0%
	120,000 hours (ZH:EN = 1:2)	4 / 1	58 / 40	38 / 59	0	2.75	48.5%
Style Conditioning	Audio Latent + LSTM	14	117	69	0	2.63	34.5%
	Audio MuLan (Audio Prompt)	4	87	107	2	2.85	54.5%
	Mixed MuLan (Audio Prompt)	4	88	107	1	**2.88**	54.0%
	T5 Embedding + LSTM	30	119	51	0	2.46	25.5%
	Audio MuLan (Text Prompt)	71	108	21	0	2.23	10.5%
	Mixed MuLan (Text Prompt)	11	106	83	0	**2.75**	41.5%
DPO & Training Stage	DiffR.	68	91	40	1	2.25	20.5%
	DiffR.$^+$ Pretrain	3	109	88	0	2.80	44.0%
	DiffR.$^+$ SFT	3	78	110	9	2.86	59.5%
	DiffR.$^+$ DPO	1	36	145	18	**3.19**	81.5%
DPO Winner	GT Winner	7	64	127	2	2.94	64.5%
	Generated Winner	1	36	145	18	**3.19**	81.5%
DPO Training Epochs	4 Epoch	4	47	139	10	3.11	74.5%
	8 Epoch	1	36	145	18	**3.19**	81.5%
	12 Epoch	2	35	138	25	3.16	81.5%
Win-Lose Score Gap	Gap = 0	4	38	138	20	3.15	79.0%
	Gap = 0.4	1	36	145	18	**3.19**	81.5%
	Gap = 0.8	4	62	121	13	3.04	67.0%

H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

消融实验

Ablation Category	Setting	[1–2]↓	[2–3]↓	[3–4]↑	[4–5]↑	Mean↑	[3–5] %↑
Data Scale & Balance	60,000 hours (ZH:EN = 1:1)	33 / 32	32 / 28	35 / 40	0	2.31	37.5%
	120,000 hours (ZH:EN = 1:1)	2 / 1	58 / 51	40 / 48	0	**2.80**	44.0%
	120,000 hours (ZH:EN = 1:2)	4 / 1	58 / 40	38 / 59	0	2.75	48.5%
Style Conditioning	Audio Latent + LSTM	14	117	69	0	2.63	34.5%
	Audio MuLan (Audio Prompt)	4	87	107	2	2.85	54.5%
	Mixed MuLan (Audio Prompt)	4	88	107	1	**2.88**	54.0%
	T5 Embedding + LSTM	30	119	51	0	2.46	25.5%
	Audio MuLan (Text Prompt)	71	108	21	0	2.23	10.5%
	Mixed MuLan (Text Prompt)	11	106	83	0	**2.75**	41.5%
DPO & Training Stage	DiffR.	68	91	40	1	2.25	20.5%
	DiffR.$^{+}$ Pretrain	3	109	88	0	2.80	44.0%
	DiffR.$^{+}$ SFT	3	78	110	9	2.86	59.5%
	DiffR.$^{+}$ DPO	1	36	145	18	**3.19**	81.5%
DPO Winner	GT Winner	7	64	127	2	2.94	64.5%
	Generated Winner	1	36	145	18	**3.19**	81.5%
DPO Training Epochs	4 Epoch	4	47	139	10	3.11	74.5%
	8 Epoch	1	36	145	18	**3.19**	81.5%
	12 Epoch	2	35	138	25	3.16	81.5%
Win-Lose Score Gap	Gap = 0	4	38	138	20	3.15	79.0%
	Gap = 0.4	1	36	145	18	**3.19**	81.5%
	Gap = 0.8	4	62	121	13	3.04	67.0%

DPO优势:

•经过SFT之后再进行DPO，能带来**最显著的质量提升**

•证明了对齐人类偏好是提升AI音乐美学质量的有效途径

H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

	歌词/风格	DiffRhythm	DiffRhythm +
重复、漏字减少	风格：country, acoustic guitar, rustic 歌词：And wreck on the edges of your deepest fears and join with shoreline sinking in		
音乐结构更加丰富、质量更高	风格：Chinese classical, cello, piano, dramatic	只有钢琴伴奏	伴奏除了风格中显式提到的钢琴、大提琴外，还有笛子来适配“中国传统”风格
	风格：electronic 歌词：Well you keep, well you keep, moving on without me		

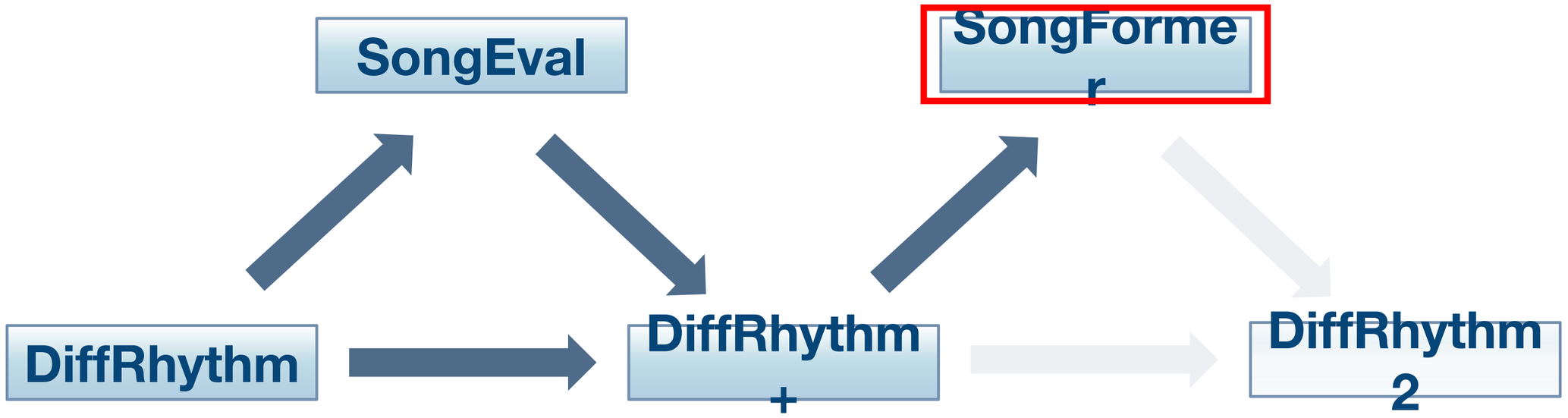
H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

DiffRhythm+

	模式/技巧	DiffRhythm	DiffRhythm+
演奏技巧涌现	转音，假声		
	自发人声和声		
	Rap solo		
	高音不破		
	美声唱法		
歌曲编辑/续写	编辑片段：[00:48 - 01:24]		
	续写片段：[00:25 - 01:35]		

H Chen, Y Jiang, G Ma, C Hao, S Wang, J Yao, Z Ning, M Meng, J Luan, L Xie. DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization. https://arxiv.org/pdf/2507.12890

SongFormer



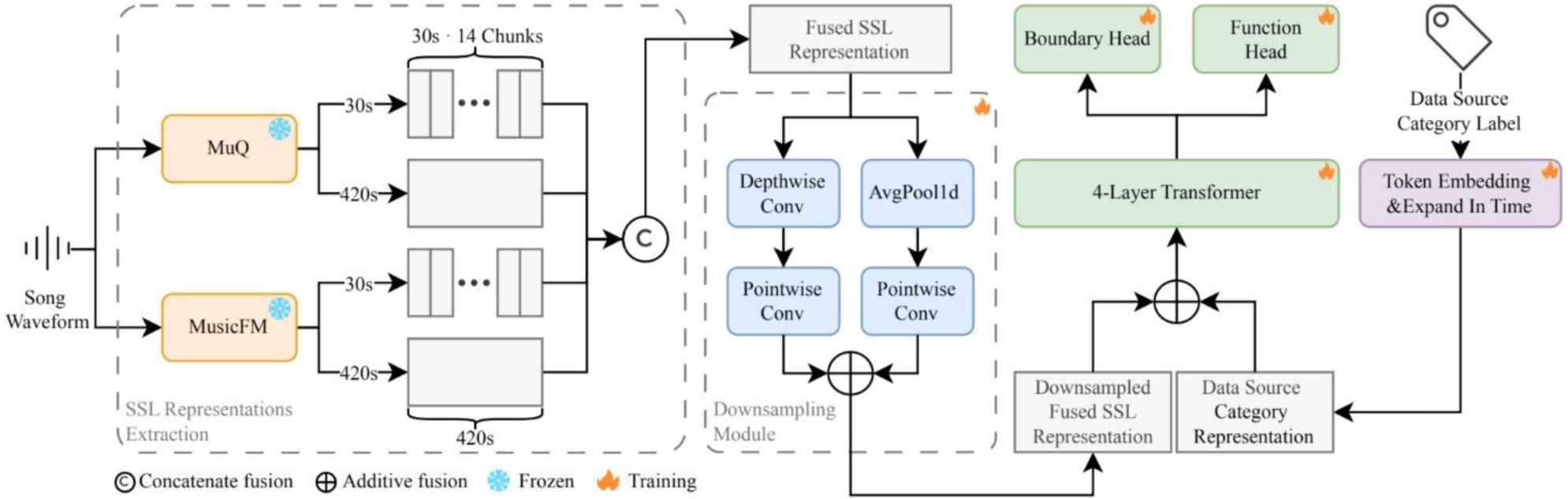
SongFormer

- **音乐结构分析 (MSA)**
  - MSA 是音乐理解和可控生成的基础
  - 当前进展受限于
    - 训练语料库规模小（HarmonixSet 仅 912 首歌曲）
    - 对预处理（如节拍跟踪、源分离）的依赖
  - 通用多模态 LLMs (如 Gemini 2.5 Pro)
    - 时间分辨率粗糙，可能引入对齐/格式问题
    - 非开源模型，成本较高

C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

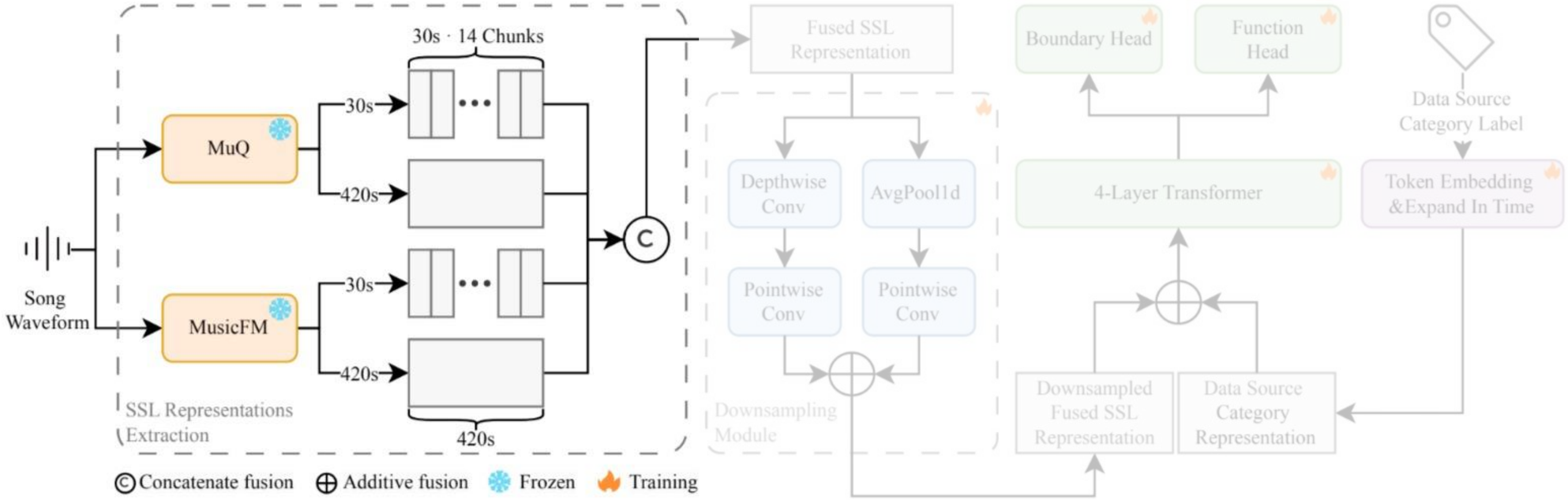
- **多尺度融合的异构框架**
  - **模型结构创新**：融合多分辨率自监督音频表示，捕捉细粒度和长序列依赖
  - **数据支持**：
    - 发布 SongFormDB (迄今为止最大的 MSA 语料库，超过 10k 歌曲)
    - 发布 SongFormBench (300 首专家验证基准)



C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

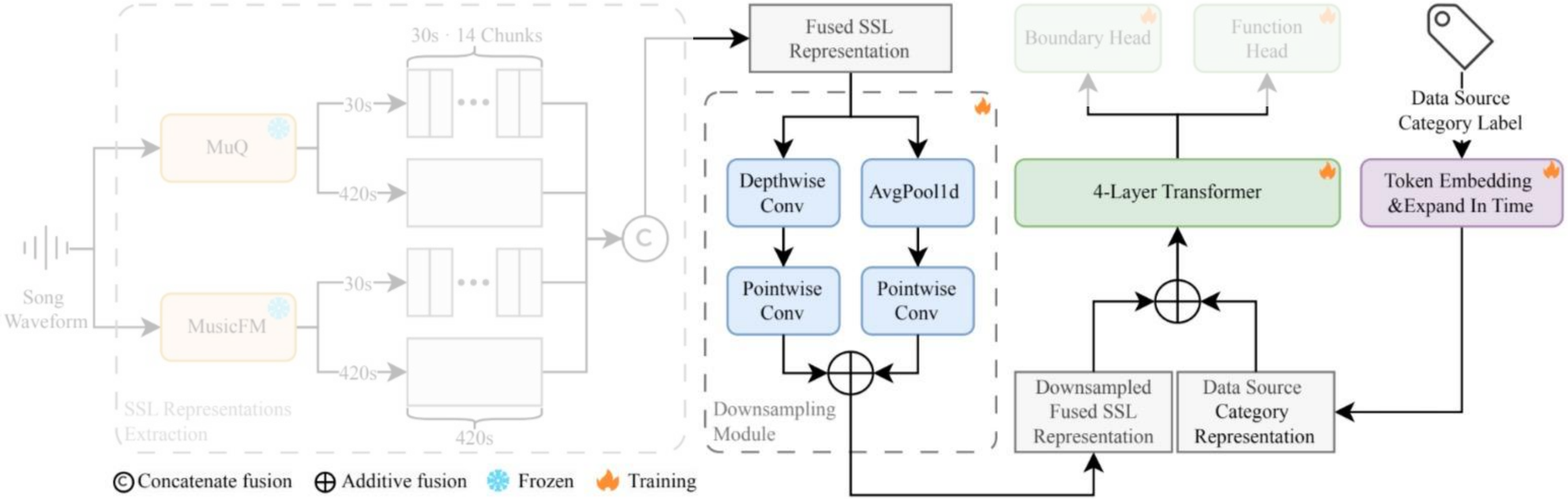
- 多尺度融合的异构框架
  - **多尺度SSL表征提取**：30s & 420s, MuQ & MusicFM
  - **特征融合**： 14 个非重叠的 30 秒块特征与 420 秒全局表示对齐并融合
    - 捕捉细粒度和长序列依赖，同时保持原始时间分辨率 (25 Hz)



C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

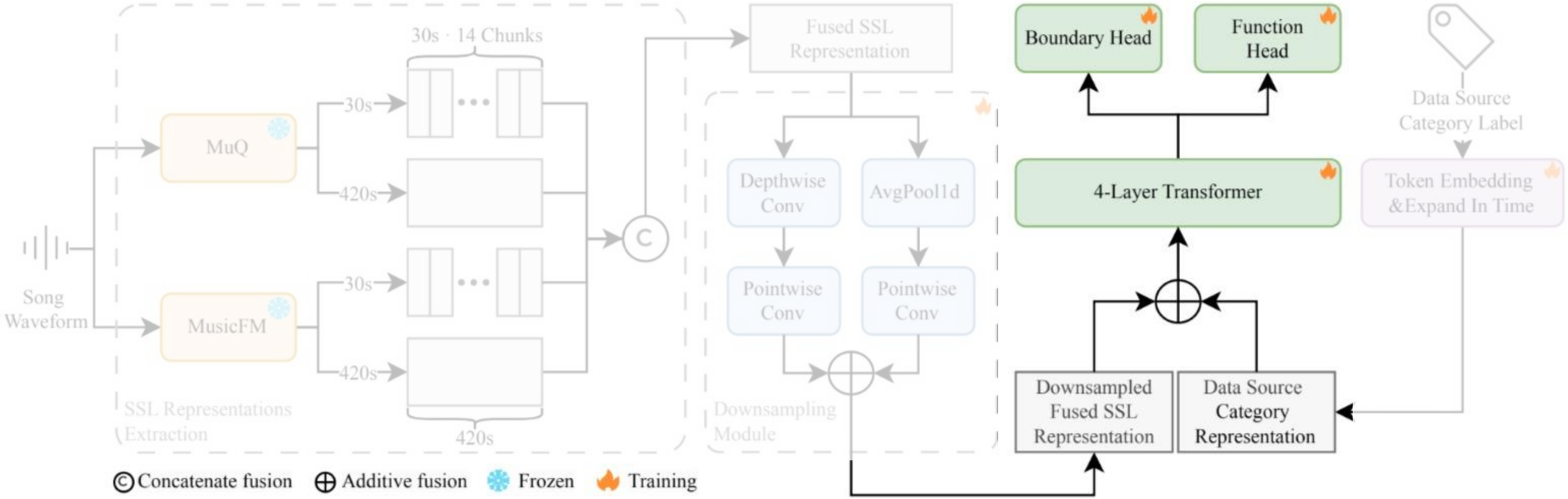
- 多尺度融合的异构框架
  - **残差下采样**：3倍下采样进一步提高计算效率，同时保留信息特征
  - **异构数据训练**：
    - 添加Label标注结构存在差异数据进行训练并指定tag，提升模型泛化性能
    - 推理时固定为HarmonixSet的tag



C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

- 多尺度融合的异构框架
  - **训练目标**：预测边界 &结构标签
    - 预测边界: BCE Loss + 平滑操作
    - 预测结构标签: 逐帧就按CE Loss



C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

- 大规模音乐结构分析语料库
  - 训练数据
    - **SongForm-HX:** 从 Mel 谱图重建 HarmonixSet 音频，并使用规则校正方法改进注释
    - **SongForm-Private**: 包含 4,314 首歌曲，使用 SOFA 修正时间戳，通过 SongFormer 伪注释
    - **SongForm-Hook**: 包含5,933 首歌曲，具有准确的结构注释
    - **SongForm-Gem**: 包含4,387 首歌曲 47 种语言，平衡节奏和流派比例，使用 Gemini 2.5 Pro 生成标注并进行后验校正
  - 评估数据
    - **SongFormBench-HarmonixSet (BHX):** 200 首 HarmonixSet 歌曲，经过专家标注者仔细修订
    - **SongFormBench-CN (BC):** 100 首中文歌曲，经过专家标注者修订

Dataset	Abbr.	Train	Eval	Test
SongForm-HX	HX	512	200	–
SongForm-Private	P	4,314	–	–
SongForm-Hook	H	5,933	–	–
SongForm-Gem	G	4,387	–	–
SongFormBench-HarmonixSet	BHX	–	–	200
SongFormBench-CN	BC	–	–	100

C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

- 实验结果
  - 评估指标
    - **HR.5F**： 0.5 秒内边界命中率的度量 (严格)
    - **HR3F**： 3 秒内边界命中率的度量 (宽松)
    - **ACC (Accuracy)**： 逐帧准确率，比较预测标签与真实值

Method	ACC	HR.5F	HR3F
SongFormBench-HarmonixSet			
Harmonic-CNN [25]	0.680⋆	0.559⋆	–
SpecTNT (24 s) [26]	0.701⋆	0.570⋆	–
SpecTNT (36 s) [26]	0.723⋆	0.558⋆	–
All-In-One [15]	0.740	0.596	0.730
MERT (5 s) [27]	0.574⋆	0.626⋆	–
MusicFM-Zhang et al. [20]	0.725⋆	0.640⋆	0.729⋆
MuQ_{iter} [18]	0.772⋆	–	–
LinkSeg-7Labels [16]	0.780	0.630	0.762
TA (Zhang et al., 2025) [20]	0.787⋆	0.610⋆	0.801⋆
Gemini 2.5 Pro [17]	0.748	0.423	**0.813**
SongFormer (HX)	0.795	**0.703**	0.784
SongFormer (HX+P+H)	0.806	0.697	0.780
SongFormer (HX+P+H+G)	**0.807**	0.696	0.780
SongFormBench-CN			
All-In-One [15]	0.834	0.563	0.771
LinkSeg-7Labels [16]	0.828	0.518	0.757
Gemini 2.5 Pro [17]	0.806	0.412	0.833
SongFormer (HX)	0.848	0.675	**0.856**
SongFormer (HX+P+H)	0.890	**0.690**	0.852
SongFormer (HX+P+H+G)	**0.891**	0.688	0.851

C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

- 实验结果
  - **HarmonixSet**
    - **整体对比:** 优于所有基线模型
    - **ACC**：SongFormer (HX+P+H+G) 达到最高 ACC (0.807)
    - **HR.5F**：SongFormer (HX) 达到最佳 HR.5F (0.703)
    - **HR3F**：SongFormer 依旧保持竞争力(0.784)
  - **SongFormBench-CN**
    - 在中文bench上效果优势更加明显

Method	ACC	HR.5F	HR3F
SongFormBench-HarmonixSet			
Harmonic-CNN [25]	0.680⋆	0.559⋆	–
SpecTNT (24 s) [26]	0.701⋆	0.570⋆	–
SpecTNT (36 s) [26]	0.723⋆	0.558⋆	–
All-In-One [15]	0.740	0.596	0.730
MERT (5 s) [27]	0.574⋆	0.626⋆	–
MusicFM-Zhang et al. [20]	0.725⋆	0.640⋆	0.729⋆
MuQ_{iter} [18]	0.772⋆	–	–
LinkSeg-7Labels [16]	0.780	0.630	0.762
TA (Zhang et al., 2025) [20]	0.787⋆	0.610⋆	0.801⋆
Gemini 2.5 Pro [17]	0.748	0.423	**0.813**
SongFormer (HX)	0.795	**0.703**	0.784
SongFormer (HX+P+H)	0.806	0.697	0.780
SongFormer (HX+P+H+G)	**0.807**	0.696	0.780
SongFormBench-CN			
All-In-One [15]	0.834	0.563	0.771
LinkSeg-7Labels [16]	0.828	0.518	0.757
Gemini 2.5 Pro [17]	0.806	0.412	0.833
SongFormer (HX)	0.848	0.675	**0.856**
SongFormer (HX+P+H)	0.890	**0.690**	0.852
SongFormer (HX+P+H+G)	**0.891**	0.688	0.851

C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

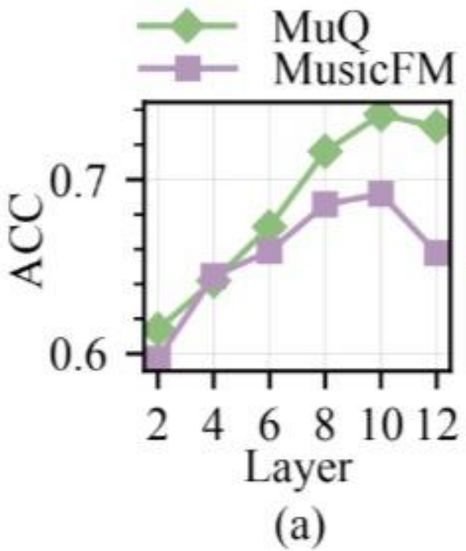
- 实验结果
  - **HarmonixSet**
    - **整体对比:** 优于所有基线模型
    - **ACC**：SongFormer (HX+P+H+G) 达到最高 ACC (0.807)
    - **HR.5F**：SongFormer (HX) 达到最佳 HR.5F (0.703)
    - **HR3F**：SongFormer 依旧保持竞争力(0.784)
  - **SongFormBench-CN**
    - 在中文bench上效果优势更加明显

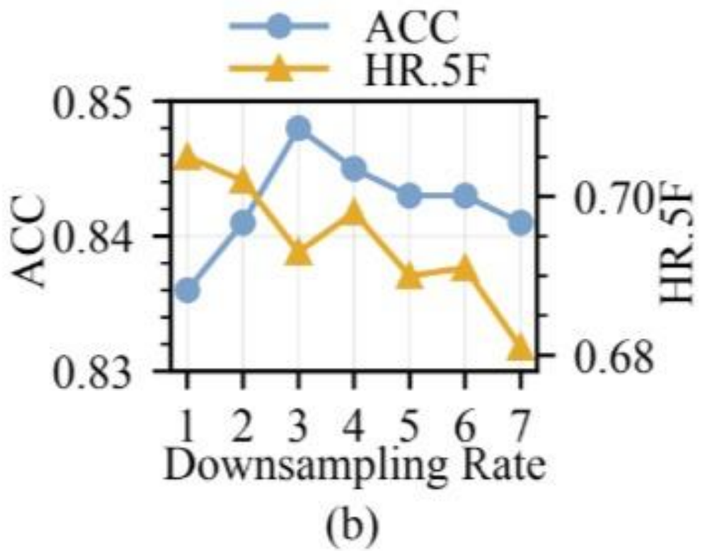
Method	ACC	HR.5F	HR3F
SongFormBench-HarmonixSet			
Harmonic-CNN [25]	0.680⋆	0.559⋆	–
SpecTNT (24 s) [26]	0.701⋆	0.570⋆	–
SpecTNT (36 s) [26]	0.723⋆	0.558⋆	–
All-In-One [15]	0.740	0.596	0.730
MERT (5 s) [27]	0.574⋆	0.626⋆	–
MusicFM-Zhang et al. [20]	0.725⋆	0.640⋆	0.729⋆
MuQ_{iter} [18]	0.772⋆	–	–
LinkSeg-7Labels [16]	0.780	0.630	0.762
TA (Zhang et al., 2025) [20]	0.787⋆	0.610⋆	0.801⋆
Gemini 2.5 Pro [17]	0.748	0.423	**0.813**
SongFormer (HX)	0.795	**0.703**	0.784
SongFormer (HX+P+H)	0.806	0.697	0.780
SongFormer (HX+P+H+G)	**0.807**	0.696	0.780
SongFormBench-CN			
All-In-One [15]	0.834	0.563	0.771
LinkSeg-7Labels [16]	0.828	0.518	0.757
Gemini 2.5 Pro [17]	0.806	0.412	0.833
SongFormer (HX)	0.848	0.675	**0.856**
SongFormer (HX+P+H)	0.890	**0.690**	0.852
SongFormer (HX+P+H+G)	**0.891**	0.688	0.851

C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

- 消融实验
  - SSL层数 & 下采样倍率
    - 使用第 10 层两种 SSL 表征结合效果最好
    - 增加下采样率会降低 HR (边界命中率)，而 ACC (准确率) 会先上升后下降
    - 适度的下采样在效率和准确性之间提供了最佳权衡
  - 多分辨率表示、下采样和异构监督策略显著提高了性能





C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

SongFormer

消融实验

多分辨率 SSL 表征的影响

- 30 秒 SSL 表征ACC 最低，短窗口无法捕捉全曲上下文
- SSL 窗口扩展到 420 提高了 ACC，但降低了 HR
  - 上下文不匹配
- 结合 420 秒SSL与30秒SSL实现了最佳性能

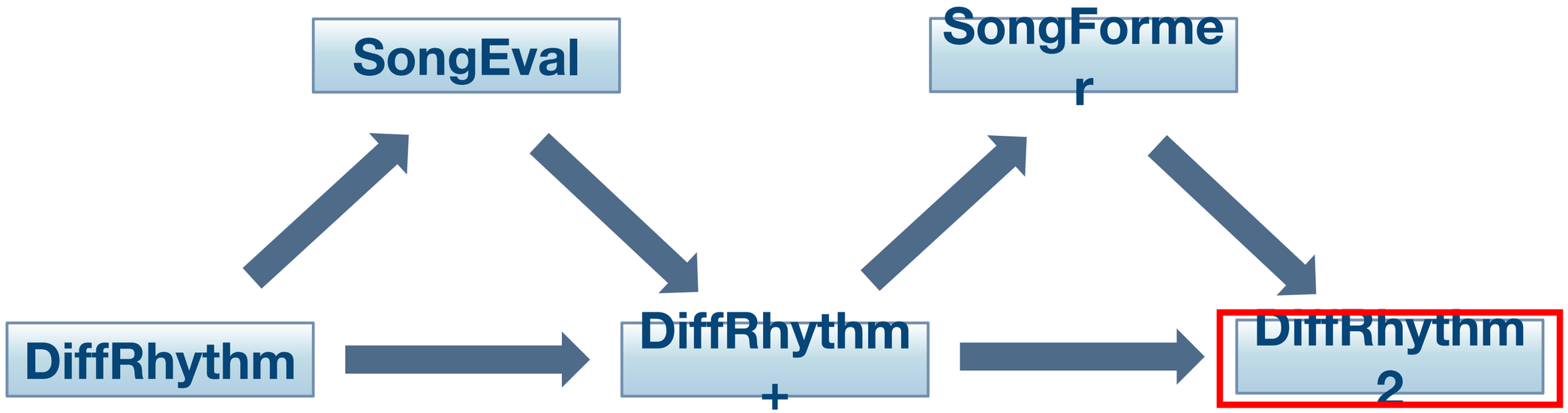
No.	30 s	420 s	Duration	ACC	HR.5F	HR3F
1	✓	✓	420 s	**0.848**	**0.693**	**0.816**
2	✓		30 s	0.782	0.689	0.817
3		✓	420 s	0.834	0.677	0.802
4	✓		420 s	0.835	0.693	0.812

M0	M1	M2	D	B	S	ACC	HR.5F	HR3F
✓	✓	✓	✓	T	✓	**0.848**	**0.693**	**0.816**
✓	✓	✓	✓	T	–	0.825	0.685	0.801
✓	✓	✓	✓	M	–	0.797	0.688	0.803
✓	✓	✓	–	M	–	0.789	0.690	0.802
✓	✓	–	–	M	–	0.754	0.688	0.802
✓	–	–	–	M	–	0.749	0.686	0.802
–	✓	–	–	M	–	0.718	0.669	0.786

M0/M1: MuQ and MusicFM M2: multi-resolution SSL
D: down sampling strategy B: architecture S: data source embedding

C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

DiffRhythm2



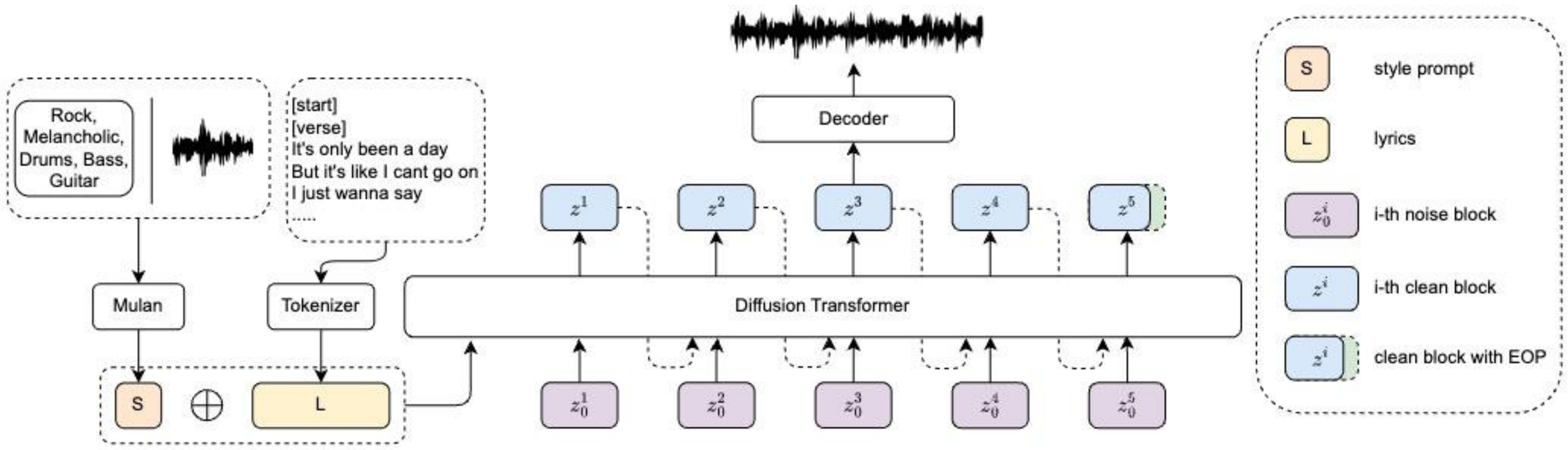
DiffRhythm2

- **研究现状**
  - **非自回归模型** (Diffrhythm，ACE-Step)：
    - NAR模型在伴奏的表现上比较容易取得较好的效果
    - 但存在对齐困难，需要额外输入或外部约束（如时间戳、特征约束损失）
  - **自回归模型** (Yue，SongGeneration)
    - AR模型在长序列的对齐稳定性上表现良好
    - 但生成速度慢，且Token的信息量有限，大多需要将人声伴奏分成独立的轨道进行生成

Y Jiang, H Chen, J Yao, Z Ning, Z Han, D Wu, M Meng, J Luan, Z Fu, L Xie. DiffRhythm 2: Efficient and High Fidelity Song Generation via Block Flow Matching.

DiffRhythm2

- **全新的端到端歌曲生成模型**
  - 利用Block Flow Matching，结合AR特性与Flow Matching，解决长序列对齐以及推理速度问题
  - 使用 MuQ-based REPA Loss，增强音乐性与整体结构
  - 提出 CrossPair DPO，解决多个目标分别DPO后再进行模型合并性能明显下降的问题
  - 实现 5Hz Music VAE，实现低帧率下高质量音乐压缩与重建



Y Jiang, H Chen, J Yao, Z Ning, Z Han, D Wu, M Meng, J Luan, Z Fu, L Xie. DiffRhythm 2: Efficient and High Fidelity Song Generation via Block Flow Matching.

参考资料

- **DiffRhythm & DiffRhythm+ & DiffRhythm2**
  - https://github.com/ASLP-lab/DiffRhythm
  - https://huggingface.co/spaces/ASLP-lab/DiffRhythm
  - https://aslp-lab.github.io/DiffRhythm.github.io
  - https://longwaytog0.github.io/DiffRhythmPlus
- **SongEval**
  - https://github.com/ASLP-lab/SongEval
  - https://huggingface.co/datasets/ASLP-lab/SongEval
- **SongFormer**
  - https://github.com/ASLP-lab/SongFormer
  - https://huggingface.co/datasets/ASLP-lab/SongFormer

C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

总结展望

- **精品歌曲生成**
  - 媲美大师的艺术作品，不追求即时生成，可以scaling inference time，生成+反思
- **理解生成相互促进**
  - 理解和生成的统一
  - 理解模型帮助打标、reward model，on-policy RL
  - 生成模型可控生成各类数据
- **Speech & Audio & Song 统一**
  - 统一的建模表征
  - 一个LLM模型生成所有，灵活控制Audio & Song出现的时机

C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

ICASSP 2026 Automatic Song Aesthetics Evaluation Challenge

- **首个针对生成歌曲美学评估竞赛**
  - Track1: 综合音乐性打分
  - Track2: 细粒度评估，五个维度打分 – 结构连贯性、记忆点、气口合理性、结构清晰度、整体音乐性
  - 基于SongEval 140小时专业音乐背景人员标注歌曲，涵盖9大流派和中英两种语言
- **日程设置**

2025.08	注册开放
2025.08	训练数据发布
2025.09	基线系统发布
2025.11	测试数据发布
2025.12	竞赛结果公布，论文提交
2026.01	论文录用通知
2026.05	研讨会召开

组委会

- 谢磊，西北工业大学
- 刘濑，上海音乐学院
- 王文武，萨里大学
- 雪巍，香港科技大学
- Yui, Sudo，SB Intuitions
- 党婷，莫纳什大学
- 刘濠赫，萨里大学
- 刘和鑫，南洋理工大学
- 吴婧瑶，麻省理工大学
- 史昊， SB Intuitions
- 姚继珣，西北工业大学
- 薛蕙心，上海音乐学院
- 袁锐斌，香港科技大学
- 马国斌，西北工业大学

C Hao, R Yuan, J Yao, Q Deng, X Bai, W Xue, L Xie. SongFormer: Scaling Music Structure Analysis with Heterogeneous Supervision.

Contributor

宁子谦

姜月鹏

马国斌

陈华康

郝春博

Audio, Speech & Language Processing Group

(ASLP@NPU)

www.npu-aslp.org

Thank You!