LeVo: High-Quality Song Generation with Multi-Preference Alignment

Shun Lei[1], Yaoxun Xu[1], Zhiwei Lin[1], Huaicheng Zhang[3], Wei Tan[2], Hangting Chen[2], Yixuan Zhang[2], Chenyu Yang[4], Haina Zhu[5], Shuai Wang[6], Zhiyong Wu[1], Dong Yu[2]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

[2] Tencent AI Lab [3] Wuhan University

[4] The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, China

[5] X-LANCE Lab, Shanghai Jiao Tong University, Shanghai

[6] School of Intelligence Science and Technology, Nanjing University, Suzhou, China

Introduction

LeVo: High-Quality Song Generation with Multi-Preference Alignment

- **Problem**
  - **Token Representation Trade-offs**
    - **Mixed Tokens (Vocal + Accompaniment Combined)**
      - Pros: High vocal-instrument harmony & musicality
      - Cons: Degraded audio quality (quantization loss) & intelligibility (accompaniment masks vocals)
    - **Dual-Track Tokens (Separate Vocal & Accompaniment Sequences):**
      - Pros: Better sound quality & lyric alignment
      - Cons: Weaker musicality (increased sequence complexity) & vocal–instrument harmony (isolated prediction of two tracks)
  - **Data Scarcity and Preference Misalignment**
    - Uneven Quality → model becomes unstable
    - Lack of musicality annotation → model cannot learn prior about musicality → generated songs do not match human preferences
    - Unreliable automatic annotations → weak instruction following (lyrics / text prompt)

Introduction

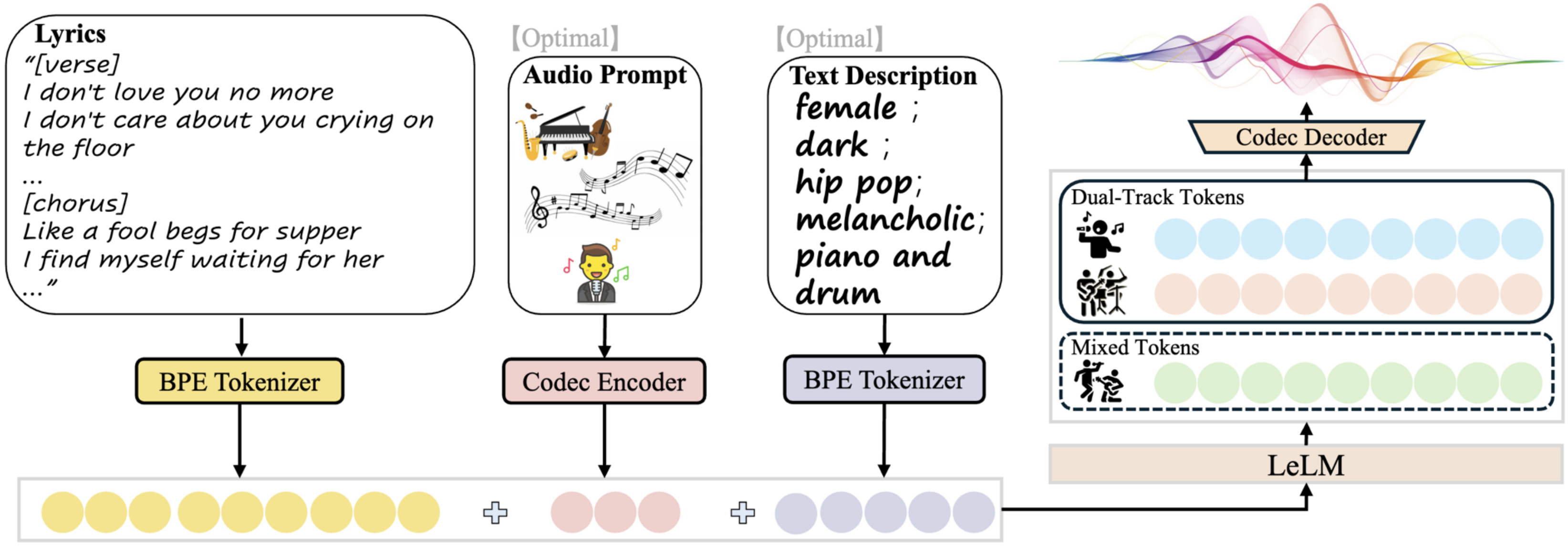
LeVo: High-Quality Song Generation with Multi-Preference Alignment

- **Contribution**
  - **Propose LeLM for parallel modeling of Mixed Tokens and Dual-Track Tokens**
    - Mixed Tokens: Capture high-level semantic information like melody and structure
    - Dual-Track Tokens: Capture fine-grained details for high fidelity vocals & accompaniment
  - **Introduce Multi-Preference Alignment Strategy for music generation**
    - Jointly optimizes: Lyric Alignment, Prompt Consistency, Musicality
  - **Establish a Three-Stage Training Paradigm: Pre-training → Modular Extension Training → Multi-Preference Alignment**
    - Pre-training → diversity & vocal–instrument harmony
    - Modular Extension Training → enhance sound quality & musicality w/o breaking pre-train knowledge
    - Multi-Preference Alignment → further improve instruction following & musicality

Methodology

LeVo: High-Quality Song Generation with Multi-Preference Alignment

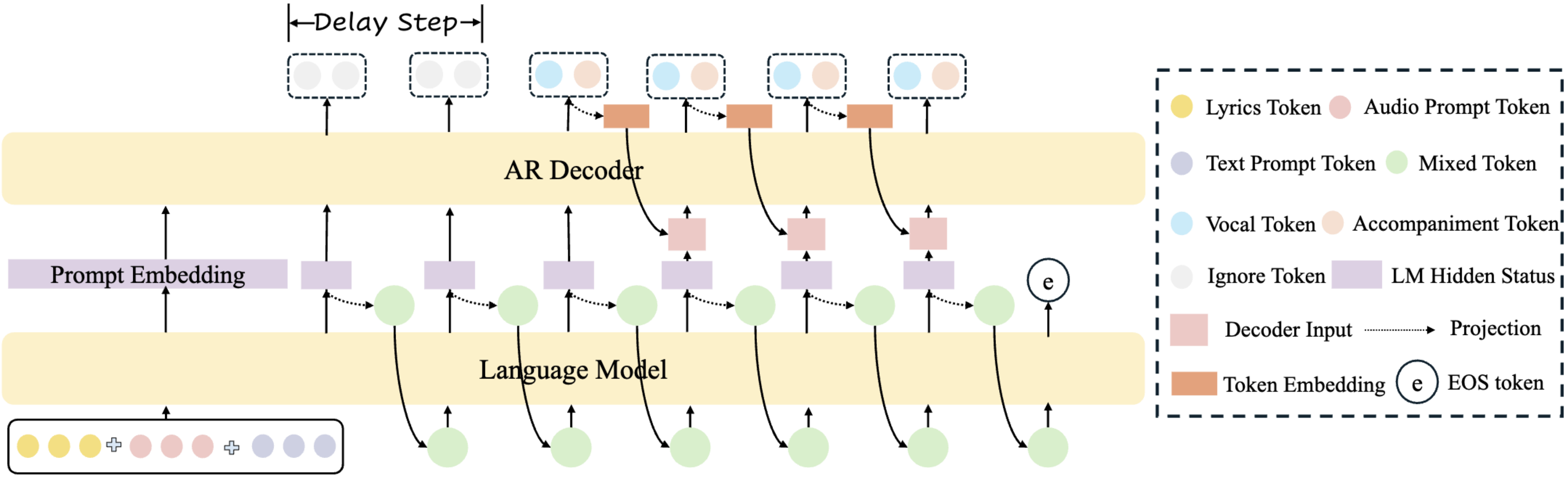
- **Overall Model**
  - **Music Encoder**: Obtains the Mixed Tokens and Dual-Track Tokens from audio, which encapsulates sufficient semantic and acoustic details that are necessary for reconstructing
  - **LeLM**: Servers as the "brain" of the system, modeling both Mixed Tokens and Dual-Track Tokens conditioned on diverse inputs (lyrics, text descriptions, audio prompts)
  - **Music Decoder**: Uses a latent diffusion model to generate high-quality music waveforms from the tokens



Methodology

LeVo: High-Quality Song Generation with Multi-Preference Alignment

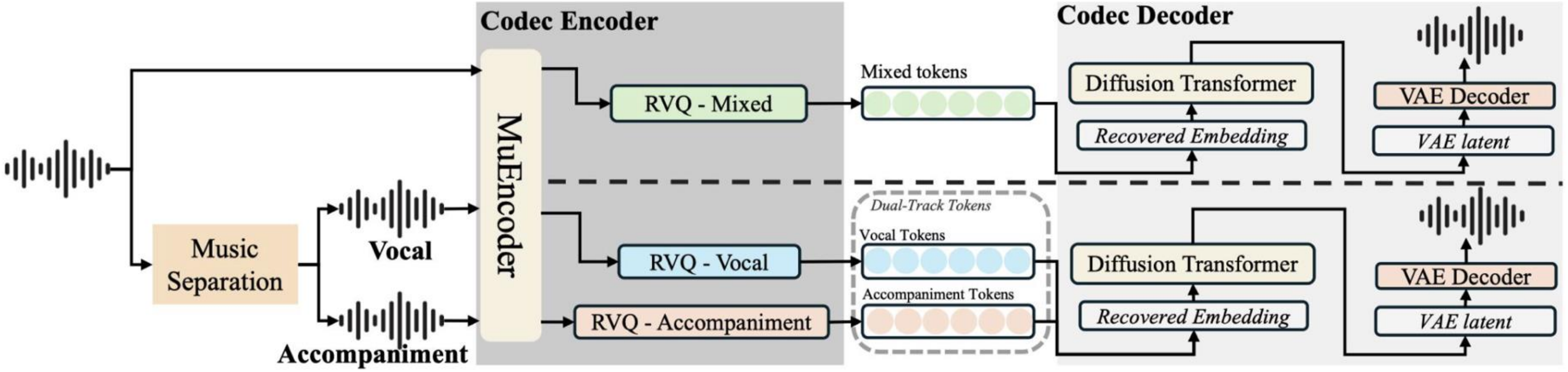
- **LeLM:** Model Mixed Tokens & Dual-Track Tokens without mutual interference by hierarchical modeling
  - **Language Model — "music structure creation"**
    - Predicts Mixed Tokens to capture high-level musical structure (melody, rhythm, arrangement)
  - **AR Decoder — "musical detail refinement"**
    - Predicts Dual-Track Tokens on top of language model outputs to refine fine-grained details of vocals & accompaniment
  - **Delay Step:** Provides additional context for local detail modeling without significantly increasing the sequence length.



Methodology

LeVo: High-Quality Song Generation with Multi-Preference Alignment

- **Music Codec:** High-compression, high-fidelity music coding and decoding
  - **Music Encoder — separate encoding of vocals & accompaniment**
    - Use MuEncoder to extract representations from each track independently
    - Then discretize into tokens (Dual-Track Tokens) by RVQ
  - **Music Decoder — joint decoding of vocals & accompaniment**
    - Use a Latent Diffusion Model (LDM) conditioned on Dual-Track Tokens to recover high-quality waveform efficiently
  - **Achieves high-fidelity reconstruction of 48kHz stereo music at an ultra-low bitrate of 0.7 kbps (operating at 25Hz frame rate)**



Methodology

LeVo: High-Quality Song Generation with Multi-Preference Alignment

- **Multi-Preference Alignment: align multiple preference dimensions into one model via DPO + interpolation**
  - **Semi-automatic construction of preference data**
    - Lyric Alignment → phoneme error number via ASR model
    - Prompt Consistency → audio–text similarity via MuLan model
    - Musicality → score from a pre-trained reward model
  - **Training & Alignment**
    - Run DPO individually on 3 preference dimensions
    - Then interpolate parameters → achieve a balanced trade-off across all preferences
- **Three-Stage Training Paradigm: reduce the mutual influence between different types of tokens & align multi-dimensional human preference**
  - **Stage 1 – Pre-training**
    - Only train Language Model → global structure, diversity and harmony
  - **Stage 2 – Modular Extension Training**
    - Only train AR Decoder → fine-grained local details, sound quality, musicality (keep Stage-1 knowledge untouched)
  - **Stage 3 – Multi-Preference Alignment**
    - Train entire LeLM with DPO → align all human preferences into final model

Experiments

LeVo: High-Quality Song Generation with Multi-Preference Alignment

- **Objective Evaluation**
  - Lowest PER (phoneme error rate)
  - Highest text–audio similarity (MuQ-T)
  - Highest perceived musical aesthetics (Audiobox-Aesthetics scores)

Table 1: Objective results of comparison and ablation systems for song generation. The asterisk (*) denotes that we reproduce SongGen using our training data. The overall first and second results are marked with **bold** and underline, respectively.

Models	FAD ↓	MuQ-T ↑	MuQ-A ↑	PER ↓	Content Scores ↑			
					CE	CU	PC	PQ
Suno-V4.5	2.59	**0.34**	0.84	21.6	7.65	7.86	5.94	8.35
Haimian	2.97	0.22	–	11.8	7.56	7.85	5.89	8.27
Mureka-O1	**2.50**	0.33	**0.87**	**7.2**	7.71	7.83	**6.39**	8.44
YuE	2.65	0.27	0.74	36.4	7.13	7.39	5.90	7.77
DiffRhythm	4.86	0.26	0.51	12.3	6.65	7.32	5.71	7.77
ACE-Step	2.69	0.28	–	37.1	7.37	7.52	6.26	7.85
SongGen*	2.68	0.25	0.80	27.5	7.63	7.79	5.94	8.37
LeVo	2.68	**0.34**	0.83	**7.2**	**7.78**	**7.90**	6.03	**8.46**
w/o Train stage 2	2.71	0.28	0.82	17.5	7.76	7.81	5.69	8.44
w/o AR decoder	2.83	0.27	0.80	26.0	7.54	7.71	5.61	8.32
w/o Dual-track	2.83	0.33	0.83	11.0	7.72	7.88	5.82	8.43
w/o DPO	2.60	0.31	0.82	10.6	7.70	7.86	5.89	8.39

Experiments

LeVo: High-Quality Song Generation with Multi-Preference Alignment

- **Subjective Evaluation**
  - Highest lyric accuracy among all open-source and commercial models
  - Overall quality (OVL), vocal melodic attractiveness (MEL), vocal–instrument harmony (HAM) and audio quality (AQ) surpassing all open-source models and several commercial systems, second only to Suno

Table 2: Subjective results of comparison and ablation systems for song generation. The asterisk (*) denotes that we reproduce SongGen using our training data. The overall first and second results are marked with **bold** and underline, respectively.

Models	MOS ↑					
	OVL	MEL	HAM	SSC	AQ	LYC
Suno-V4.5	**3.59**	**4.10**	**3.93**	**4.19**	**4.00**	3.17
Haimian	3.05	3.51	3.55	3.62	3.87	<u>3.32</u>
Mureka-O1	<u>3.42</u>	3.88	3.89	<u>4.14</u>	3.87	<u>3.32</u>
YuE	2.45	3.04	2.94	3.53	3.08	2.41
DiffRhythm	2.60	3.18	3.22	3.55	3.09	2.69
ACE-Step	2.26	3.02	3.30	3.21	2.36	2.22
SongGen*	2.91	3.43	3.44	3.66	3.69	2.84
LeVo	<u>3.42</u>	<u>3.93</u>	<u>3.90</u>	4.09	<u>3.96</u>	**3.38**
w/o Train stage 2	3.29	3.76	3.77	3.80	<u>3.96</u>	2.91
w/o AR decoder	2.93	3.44	3.34	3.59	3.71	2.74
w/o Dual-track	3.25	3.82	3.84	3.96	3.86	3.18
w/o DPO	3.18	3.71	3.76	3.97	3.93	3.18

Metric Notation
- OVL: Overall Quality
- MEL: Vocal Melodic Attractiveness
- HAM: Vocal–Instrument Harmony
- SSC: Song Structure Clarity
- AQ: Audio Quality
- LYC: Lyric Accuracy

Experiments

LeVo: High-Quality Song Generation with Multi-Preference Alignment

- **Music Codec Performance**
  - Mixed Tokens: best reconstruction at **0.35 kbps**, comparable to **2-layer XCodec (1 kbps)**
  - Dual-Track Tokens: best reconstruction at **0.7 kbps**, surpassing **4-layer XCodec (2 kbps)**

Method	CodeBook	Tokenrate (tps)	Bitrate (kbps)	VISQOL ↑	SPK_SIM ↑	WER (%)
Original music	—	—	—	—	—	10.92
SemantiCodec	1 x 32768	25	0.375	1.92/1.92	0.52	120.17
	1 x 16384	100	1.40	1.96/1.96	0.68	55.17
WavTokenizer	1 x 4096	40	0.48	2.93/2.93	0.49	101.49
	1 x 4096	75	0.90	3.05/3.05	0.56	86.19
XCodec	1 x 1024	50	0.50	3.04/3.04	0.53	85.10
	2 x 1024	50	2.00	3.30/3.30	0.79	55.37
	4 x 1024	50	2.00	3.38/3.38	0.63	36.32
	8 x 1024	50	4.00	3.58/3.58	0.72	26.42
MuCodec	1 x 16384	25	0.35	3.17/3.18	0.75	36.21
	4 x 10000	25	1.33	3.45/3.46	0.87	24.26
Music Codec (Mixed)	1 x 16384	25	0.35	3.27/3.27	0.78	38.22
	2 x 16384	25	0.70	3.34/3.34	0.82	33.43
	4 x 16384	25	1.40	3.52/3.53	0.84	28.92
Music Codec (Dual-Track)	16384+16384	25	0.70	3.43/3.44	0.82	31.54

Experiments

LeVo: High-Quality Song Generation with Multi-Preference Alignment

- **Ablation study — Framework**
  - Train Stage 2 + AR Decoder prevent interference between Mixed & Dual-Track Tokens
    - w/o Stage 2: all metrics drop, esp. lyric accuracy
    - w/o AR Decoder: further overall degradation
  - Parallel prediction balances sound quality, intelligibility, and harmony
    - w/o Dual-Track: sharp decline in sound quality, intelligibility, and musicality

Models	FAD ↓	MuQ-T ↑	MuQ-A ↑	PER ↓	Content Scores ↑			
					CE	CU	PC	PQ
Suno-V4.5	2.59	**0.34**	0.84	21.6	7.65	7.86	5.94	8.35
Haimian	2.97	0.22	–	11.8	7.56	7.85	5.89	8.27
Mureka-O1	**2.50**	0.33	**0.87**	**7.2**	7.71	7.83	**6.39**	8.44
YuE	2.65	0.27	0.74	36.4	7.13	7.39	5.90	7.77
DiffRhythm	4.86	0.26	0.51	12.3	6.65	7.32	5.71	7.77
ACE-Step	2.69	0.28	–	37.1	7.37	7.52	6.26	7.85
SongGen*	2.68	0.25	0.80	27.5	7.63	7.79	5.94	8.37
LeVo	2.68	**0.34**	0.83	**7.2**	**7.78**	**7.90**	6.03	**8.46**
w/o Train stage 2	2.71	0.28	0.82	17.5	7.76	7.81	5.69	8.44
w/o AR decoder	2.83	0.27	0.80	26.0	7.54	7.71	5.61	8.32
w/o Dual-track	2.83	0.33	0.83	11.0	7.72	7.88	5.82	8.43
w/o DPO	2.60	0.31	0.82	10.6	7.70	7.86	5.89	8.39

Models	MOS ↑					
	OVL	MEL	HAM	SSC	AQ	LYC
Suno-V4.5	**3.59**	**4.10**	**3.93**	**4.19**	**4.00**	3.17
Haimian	3.05	3.51	3.55	3.62	3.87	3.32
Mureka-O1	3.42	3.88	3.89	4.14	3.87	3.32
YuE	2.45	3.04	2.94	3.53	3.08	2.41
DiffRhythm	2.60	3.18	3.22	3.55	3.09	2.69
ACE-Step	2.26	3.02	3.30	3.21	2.36	2.22
SongGen*	2.91	3.43	3.44	3.66	3.69	2.84
LeVo	3.42	3.93	3.90	4.09	3.96	**3.38**
w/o Train stage 2	3.29	3.76	3.77	3.80	3.96	2.91
w/o AR decoder	2.93	3.44	3.34	3.59	3.71	2.74
w/o Dual-track	3.25	3.82	3.84	3.96	3.86	3.18
w/o DPO	3.18	3.71	3.76	3.97	3.93	3.18

Experiments

LeVo: High-Quality Song Generation with Multi-Preference Alignment

- **Ablation study — DPO strategy**
  - Proposed multi-preference alignment improves
    - Musicality (Content Scores, OVL, MEL)
    - Lyric Alignment (PER, LYC)
    - Instruction Following (MuQ-T, MuQ-A)
  - Single-preference DPO: targeted enhancement
    - Strategy 1 → PER
    - Strategy 2 → MuQ-T/A
    - Strategy 3 → Content Scores
  - Mixed training improves multiple aspects
  - Interpolation model best balances all dimensions

Models	FAD ↓	MuQ-T ↑	MuQ-A ↑	PER ↓	Content Scores ↑			
					CE	CU	PC	PQs
w/o DPO	**2.60**	0.31	0.82	10.6	7.70	7.86	5.89	8.39
with Strategy 1	2.85	0.30	0.81	**6.5**	7.72	7.86	5.97	8.42
with Strategy 2	2.89	**0.34**	**0.83**	10.3	7.75	7.87	5.96	8.43
with Strategy 3	<u>2.63</u>	0.32	0.82	11.2	**7.78**	**7.93**	**6.16**	<u>8.45</u>
Mixed Training	2.75	0.33	**0.83**	7.5	7.76	7.89	<u>6.04</u>	8.43
LeVo (Interpolation)	2.68	**0.34**	**0.83**	<u>7.2</u>	**7.78**	<u>7.90</u>	6.03	**8.46**

Models	MOS ↑					
	OVL	MEL	HAM	SSC	AQ	LYC
Suno-V4.5	**3.59**	**4.10**	**3.93**	**4.19**	**4.00**	3.17
Haimian	3.05	3.51	3.55	3.62	3.87	<u>3.32</u>
Mureka-O1	<u>3.42</u>	3.88	3.89	<u>4.14</u>	3.87	<u>3.32</u>
YuE	2.45	3.04	2.94	3.53	3.08	2.41
DiffRhythm	2.60	3.18	3.22	3.55	3.09	2.69
ACE-Step	2.26	3.02	3.30	3.21	2.36	2.22
SongGen*	2.91	3.43	3.44	3.66	3.69	2.84
LeVo	<u>3.42</u>	<u>3.93</u>	<u>3.90</u>	4.09	<u>3.96</u>	**3.38**
w/o Train stage 2	3.29	3.76	3.77	3.80	<u>3.96</u>	2.91
w/o AR decoder	2.93	3.44	3.34	3.59	3.71	2.74
w/o Dual-track	3.25	3.82	3.84	3.96	3.86	3.18
w/o DPO	3.18	3.71	3.76	3.97	3.93	3.18

Experiments

LeVo: High-Quality Song Generation with Multi-Preference Alignment

Thanks!

Source code: https://github.com/tencent-ailab/songgeneration
Hugging Face Space: https://huggingface.co/spaces/waytan22/SongGeneration-LeVo
Contact: leis21@mails.tsinghua.edu.cn

Listen to Samples