SongCreator: Lyrics-based Universal Song Generation

Shun Lei[1], Yixuan Zhou[1], Boshi Tang[1], Max W. Y. Lam[2], Feng Liu[2], Hangyu Liu[2], Jingcheng Wu[2], Shiyin Kang[2], Zhiyong Wu[1,3], Helen Meng[3]

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

[2] Independent Researcher

[3] The Chinese University of Hong Kong, Hong Kong SAR, China

Content

THUHCSI
清华大学人机语音交互实验室

- **Song Generation**

  - **Generate songs with harmonious and pleasant vocals and accompaniment**
  - Previous works mostly explored specific aspects of song generation, such as **vocal composition**, **instrumental arrangement**, and **harmonious generation**, but none of them is able to combine these three for **song generation**
    - The **coordination** among various complex elements in vocals and accompaniment poses significant challenges for generating songs as an entity
    - The demands for song generation are **highly diverse**, not only lyrics-to-song generation but also independent vocal or instrumental music generation, song editing, and song generation from a given audio prompt or pre-determined track.

Table 1: A comparison of song generation with related tasks in the literature. We use **Composition** to denote whether the model can complete vocal composition, **Arrangement** to denote whether the model can arrange the instrumental accompaniment, and **Harmony** to denote whether vocals and accompaniment sound harmonious and pleasant together.

Tasks	**Inputs**	**Outputs**	**Composition**	**Arrangement**	**Harmony**
Singing Voice Synthesis [15-20]	Scores	Vocals	✗	✗	✗
SongComposer [21]	Lyrics	Vocals	✓	✗	✗
Text-to-Music [22-25]	Text Description	Music	✗	✓	✗
Accompaniment Generation [26-30]	Vocals	Music	✗	✓	✓
Song Generation	Lyrics	Song	✓	✓	✓

THUHCSI 清华大学人机语音交互实验室

- **Contribution**
  - Propose a novel **dual-sequence language model (DSLM)** for song generation, which not only **emphasizes the respective quality** of vocals and accompaniment but also learns their **mutual influences** to coordinate them into harmonious songs
  - Propose a **series of attention mask strategies**, which enables our model to complete song generation tasks of various forms, such as **editing**, **understanding**, and **generation**
  - Based on the above mechanism, we propose a versatile system for song generation named SongCreator
    - Support **universal conditioning and generation** for **eight tasks**
    - Achieving **state-of-the-art** or **competitive** performances on all tasks

Tasks	Conditions	Outputs
Lyrics-to-song*	Lyrics, [Vocal prompt], [Accompaniment prompt]	Song, Vocals
Lyrics-to-vocals*	Lyrics, [Vocal prompt]	Vocals
Accompaniment-to-song	Lyrics, Accompaniment, [Vocal prompt]	Song, Vocals
Vocals-to-song	Vocals, [Lyrics], [Accompaniment prompt]	Song, Music
Music continuation	Accompaniment prompt	Music
Song editing*	Lyrics, Vocals, Accompaniment	Song, Vocals
Vocals editing	Lyrics, Vocals	Vocals
Vocals editing in song*	Lyrics, Vocals, Accompaniment	Song, Vocals

Overall Model

- BEST-RQ: Self-supervised learning model to obtain the semantic tokens from audio, which encapsulates sufficient semantic and acoustic details that are necessary for reconstructing
- DSLM: The “brain” of our system for predicting the semantic token of songs from a variety of optional inputs, including lyrics, vocal prompt, accompaniment prompt, pre-determined vocal track, and pre-determined accompaniment track
- LDM: Decode the semantic tokens into high-quality song audio

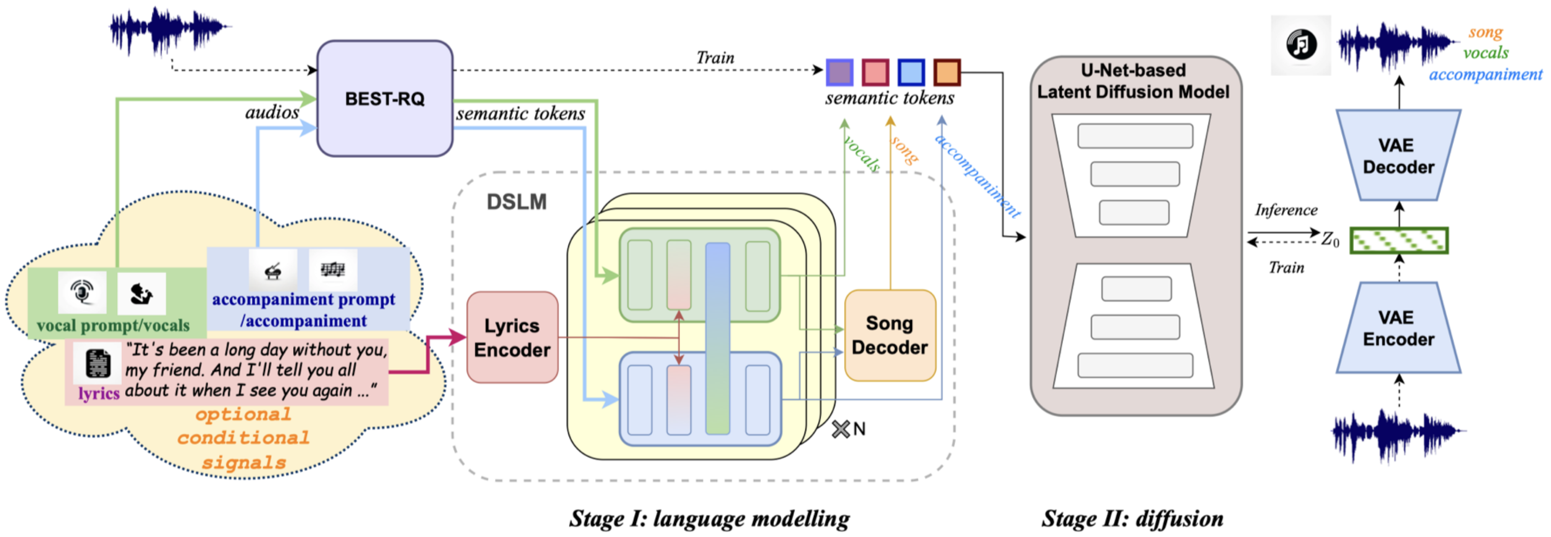
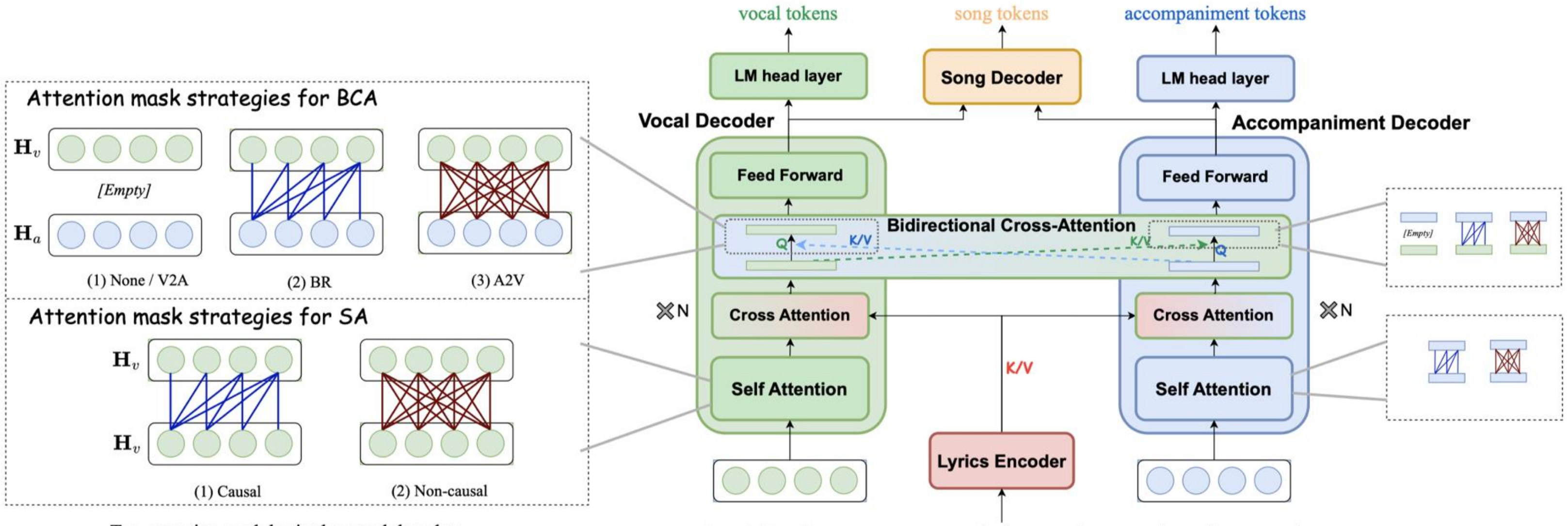


Figure 1: The overview of SongCreator. The BEST-RQ tokens is a proxy that bridges the DSLM and the latent diffusion model.

Methodology

SongCreator: Lyrics-based Universal Song Generation

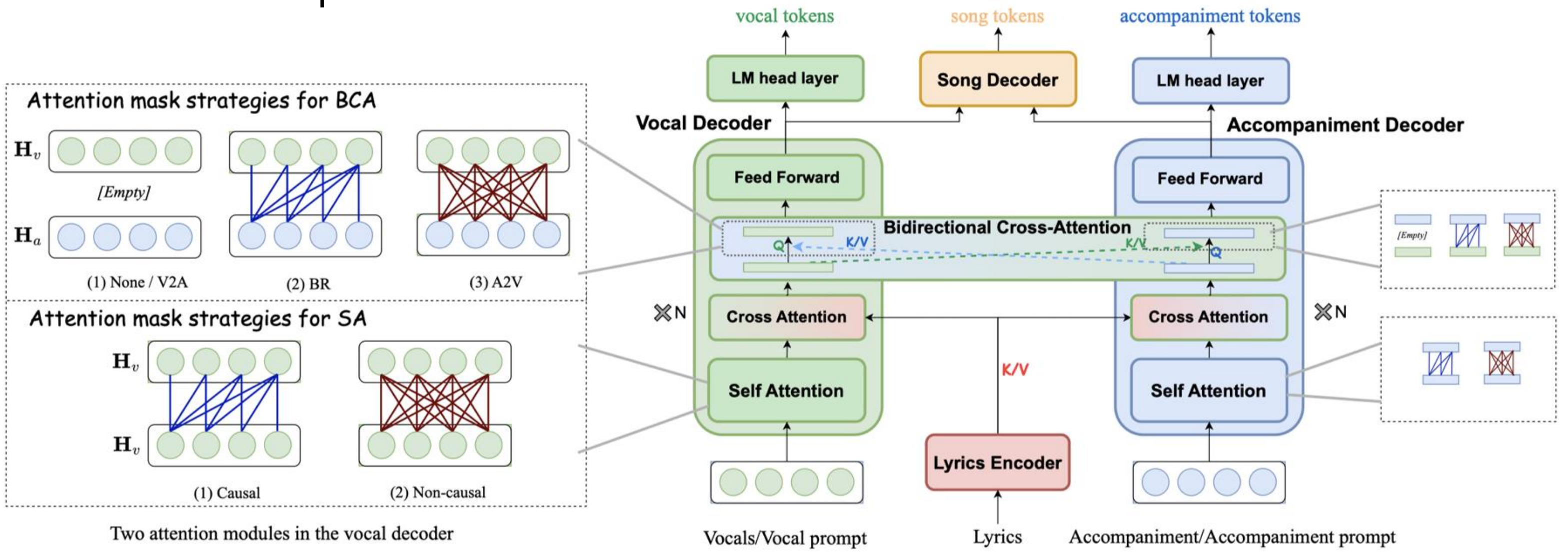
- **Dual-Sequence Language Model (DSLM)**
  - A lyrics encoder to extract critical information related to the pronunciation of the lyrics
  - Two decoders to autoregressively generate semantic tokens for the vocals and accompaniment, respectively
    - Utilize the cross-attention layer to attend the information from the lyrics encoder
    - Utilize the bidirectional cross-attention (BCA) layer to capture and model the complex interrelationship between vocals and accompaniment
  - A final song decoder to non-autoregressively generate semantic tokens for songs



Two attention modules in the vocal decoder

- ## Attention mask strategies for DSLM

  - For Self-Attention (SA)
    - Causal attention mask: each token can only access the leftward context tokens and itself
    - Non-causal attention mask: all tokens can attend to each other within the same sequence
  - For Bidirectional Cross-Attention (BCA)
    - Bidirectional (BR) mask: tokens in each sequence can attend to the tokens in the other sequence that occur at earlier positions
    - Accompaniment-to-Vocals (A2V) and Vocals-to-Accompaniment (V2A) mask: tokens in one sequence can attend to all tokens in the other sequence
    - None mask: neither sequence can attend to the other



SongCreator: Lyrics-based Universal Song Generation

- **The result of tasks**
  - SongCreator achieves **state-of-the-art** or **competitive** performances on all tasks
  - SongCreator can better maintain the acoustic conditions from prompt and can independently control the vocals and accompaniment in the generated songs
  - humans judge the edited song produced by SongCreator to be as natural as the original unedited song

Table 3: Lyrics-to-song evaluation without audio prompt.

Model	FAD ↓	Musicality ↑	Quality ↑
Ground Truth	-	4.3 ± 0.04	4.09 ± 0.05
MusicLM	6.47	3.21 ± 0.09	3.25 ± 0.07
MusicGen	2.31	3.08 ± 0.06	2.99 ± 0.06
GPT	8.18	3.32 ± 0.10	3.26 ± 0.08
GPT (Vocals & Song)	11.23	3.55 ± 0.09	3.64 ± 0.07
SongCreator	**2.14**	$\mathbf{4.25 \pm 0.05}$	$\mathbf{4.08 \pm 0.06}$
SongCreator (Single)	3.04	3.85 ± 0.06	3.75 ± 0.05

Table 4: Lyrics-to-vocals evaluation without audio prompt.

Model	Musicality ↑	Quality ↑
Ground Truth	3.89 ± 0.09	3.91 ± 0.07
MusicLM	3.31 ± 0.06	3.35 ± 0.06
VALL - E	3.15 ± 0.08	3.23 ± 0.06
GPT	3.64 ± 0.07	3.58 ± 0.07
SongCreator	$\mathbf{3.98 \pm 0.04}$	$\mathbf{3.79 \pm 0.05}$
SongCreator (Vocal Only)	3.68 ± 0.06	3.63 ± 0.05
SongCreator (Single)	3.53 ± 0.06	3.64 ± 0.05

Table 5: Prompt-based lyrics-to-song. We sample the prompt at random from a held-out set.

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**	$\mathbf{4.01 \pm 0.07}$	$\mathbf{3.82 \pm 0.08}$

Table 6: Prompt-based lyrics-to-vocals. We sample the prompt at random from a held-out set.

Model	SECS ↑	Musicality ↑	Similarity ↑
Ground Truth	0.62	3.63 ± 0.08	3.57 ± 0.08
VALL - E	0.66	3.34 ± 0.07	3.30 ± 0.08
SongCreator	**0.68**	$\mathbf{3.57 \pm 0.06}$	$\mathbf{3.55 \pm 0.07}$

Table 7: Vocals-to-song evaluation.

Model	FAD ↓	Musicality ↑	Harmony ↑
Ground Truth	-	4.12 ± 0.05	3.91 ± 0.08
SingSong	3.37	3.67 ± 0.10	3.63 ± 0.08
SingSong (Diffusion)	4.13	3.71 ± 0.08	3.67 ± 0.06
GPT	3.07	3.73 ± 0.07	3.69 ± 0.07
SongCreator	1.88	$\mathbf{3.77 \pm 0.08}$	$\mathbf{3.77 \pm 0.07}$
SongCreator (Single)	**1.46**	3.58 ± 0.08	3.65 ± 0.06

Table 8: Accompaniment-to-song evaluation.

Model	FAD ↓	Musicality ↑	Harmony ↑
Ground Truth	-	4.15 ± 0.07	4.11 ± 0.07
SingSong	1.82	3.36 ± 0.06	3.42 ± 0.07
SingSong (Diffusion)	2.98	3.66 ± 0.06	3.65 ± 0.05
GPT	1.64	3.53 ± 0.08	3.53 ± 0.09
SongCreator	1.24	$\mathbf{3.67 \pm 0.05}$	$\mathbf{3.78 \pm 0.06}$
SongCreator (Single)	**1.23**	3.60 ± 0.07	3.62 ± 0.06

Table 9: Music continuation evaluation.

Model	FAD ↓	Musicality ↑	Similarity ↑
Ground Truth	-	3.9 ± 0.11	3.70 ± 0.10
AudioLM	1.33	3.95 ± 0.10	3.78 ± 0.08
GPT	**1.28**	3.90 ± 0.10	3.73 ± 0.11
SongCreator	1.54	$\mathbf{3.97 \pm 0.08}$	$\mathbf{3.83 \pm 0.08}$

Table 10: Song editing evaluation.

Model	FAD ↓	MCD ↓	Musicality ↑	Naturalness ↑
Ground Truth	-	-	4.08 ± 0.07	3.99 ± 0.06
GPT	2.29	8.30	3.84 ± 0.07	3.72 ± 0.06
SongCreator	**1.81**	7.90	$\mathbf{4.01 \pm 0.06}$	$\mathbf{3.78 \pm 0.07}$
SongCreator (Single)	1.87	**7.85**	3.93 ± 0.08	3.75 ± 0.08

Table 11: Vocals editing evaluation.

Model	SECS ↑	Musicality ↑	Naturalness ↑
Ground Truth	-	3.65 ± 0.08	3.45 ± 0.07
GPT	0.87	3.64 ± 0.07	$\mathbf{3.43 \pm 0.07}$
SongCreator	0.87	$\mathbf{3.68 \pm 0.06}$	3.31 ± 0.06
SongCreator (Single)	0.87	3.63 ± 0.06	3.41 ± 0.06

Thanks!

Listen to Samples

Contact: leis21@mails.tsinghua.edu.cn