Segment-Factorized Full-Song Generation on Symbolic Piano Music

Ping-Yi Chen[1], Chih-Pin Tan[2], Yi-Hsuan Yang[2]

[1] National Cheng Kung University [2] National Taiwan University

Project Page

Paper

Motivation

Challenges for **full-song generation**

- Maintain coherence across the overall song structure
- Generate long sequences efficiently

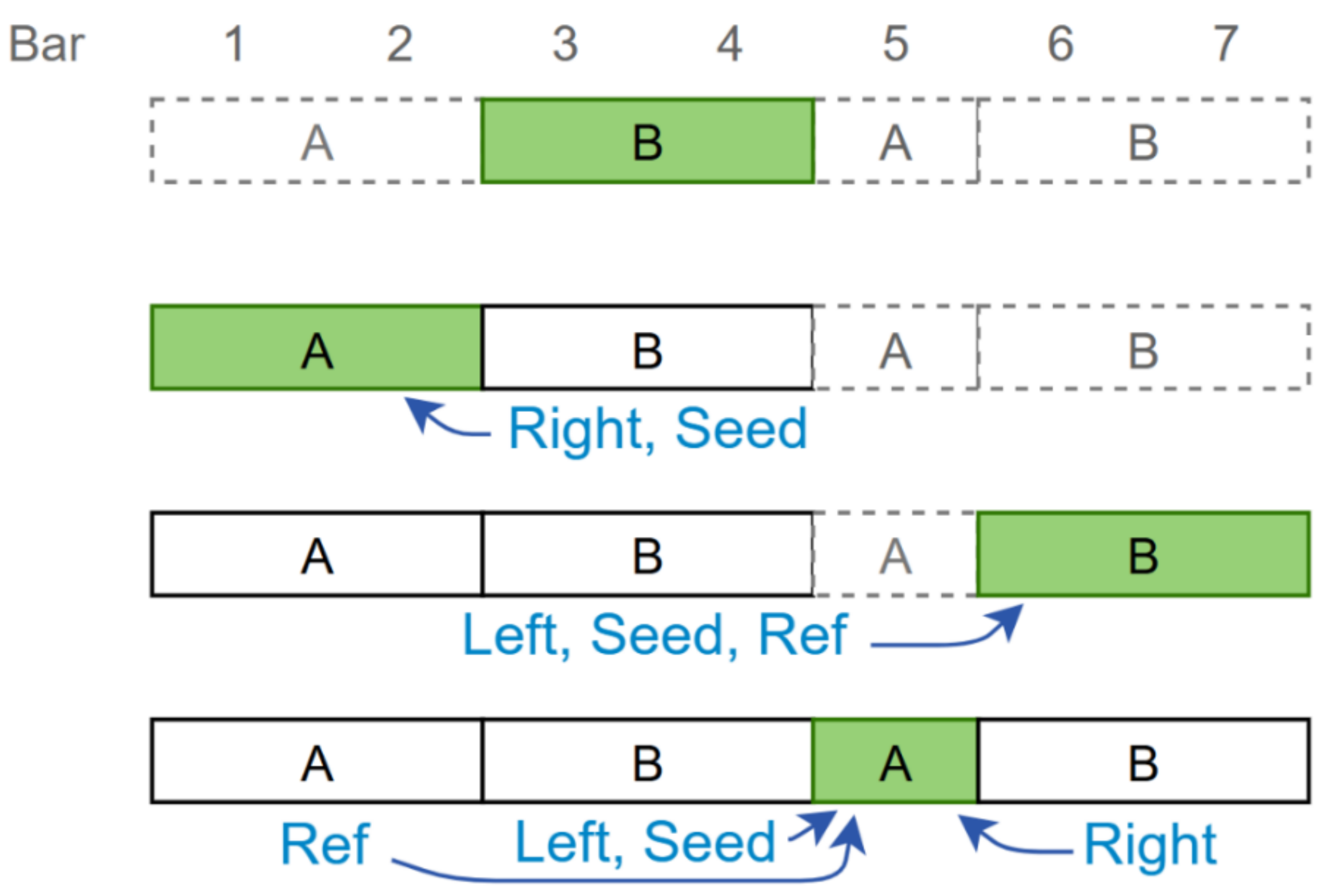
We ask: how do human create music without hitting these challenges?

- Begin with a theme and the song structure
- Selective attention to relevant context

inspires

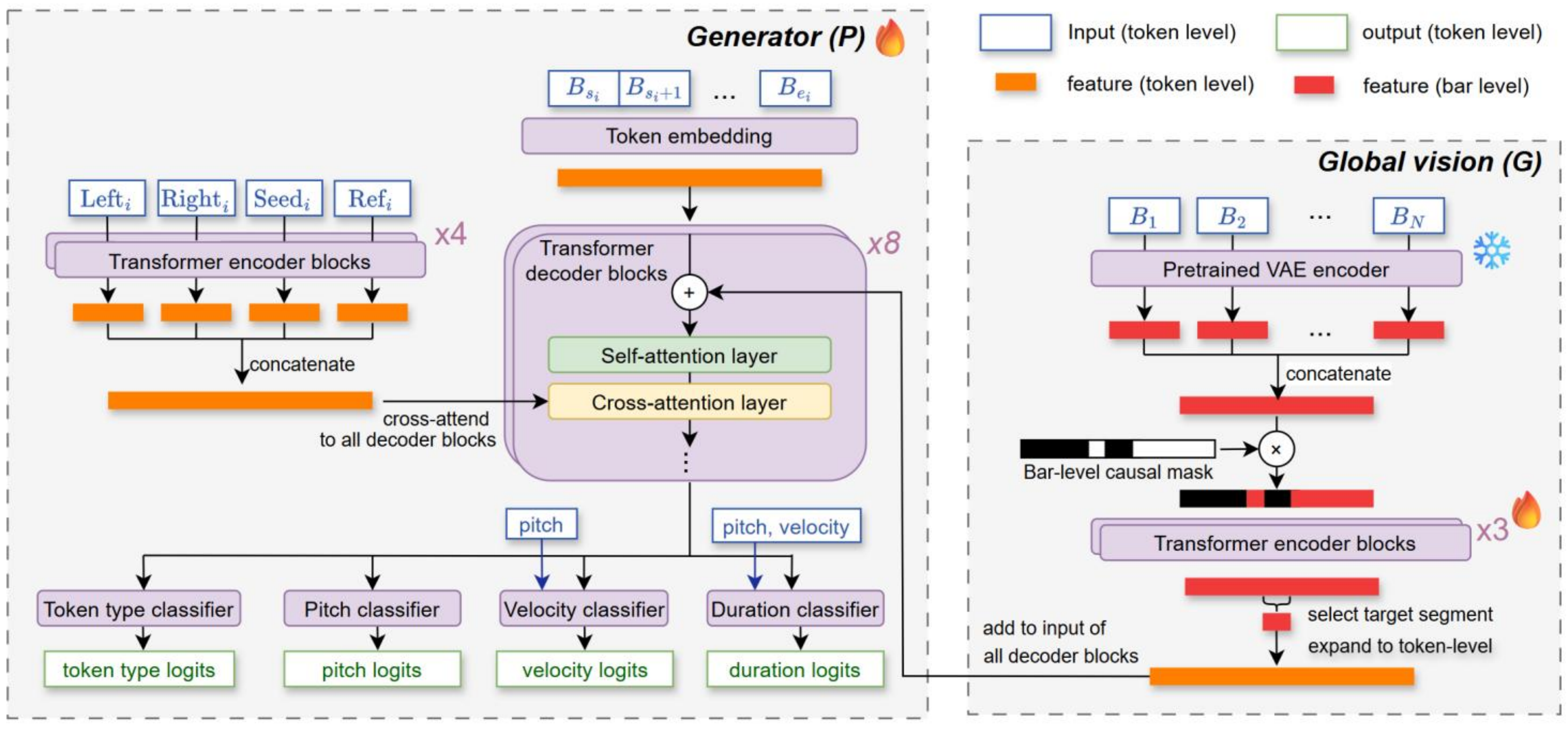
Formulation

- Training data are songs with segmentation labels
- The model learns to autoregressively generate segments in random orders
- Selected context for attention
  - **Left**: The left neighbor among already-generated segments
  - **Right**: The right neighbor among already-generated segments
  - **Seed**: The first generated segment, considered as the song's theme
  - **Reference**: An already-generated segment with the same label



Model Implementation

- Full Transformer
- Context segments cross-attend at token-level
- Cross and self-attention use RoPE based on position in song (not in token sequence)



Evaluation

		SI			User Study	
Model	**Inference Speed**	$\mathbf{SI}_{2-8}$	$\mathbf{SI}_{8-16}$	$\mathbf{SI}_{16+}$	**O**	**A**
SFS (Ours)	2.03 beat/sec.	0.3286	**0.2264**	**0.1109**	3.14	**3.59**
WholeSong	0.197 beat/sec.	0.3234	0.2262	0.0860	3.02	3.16
Flat	5.68 beat/sec.	**0.3426**	0.1990	0.0409	**3.36**	2.34
Datset	-	0.4398	0.3827	0.3300	4.00	4.07

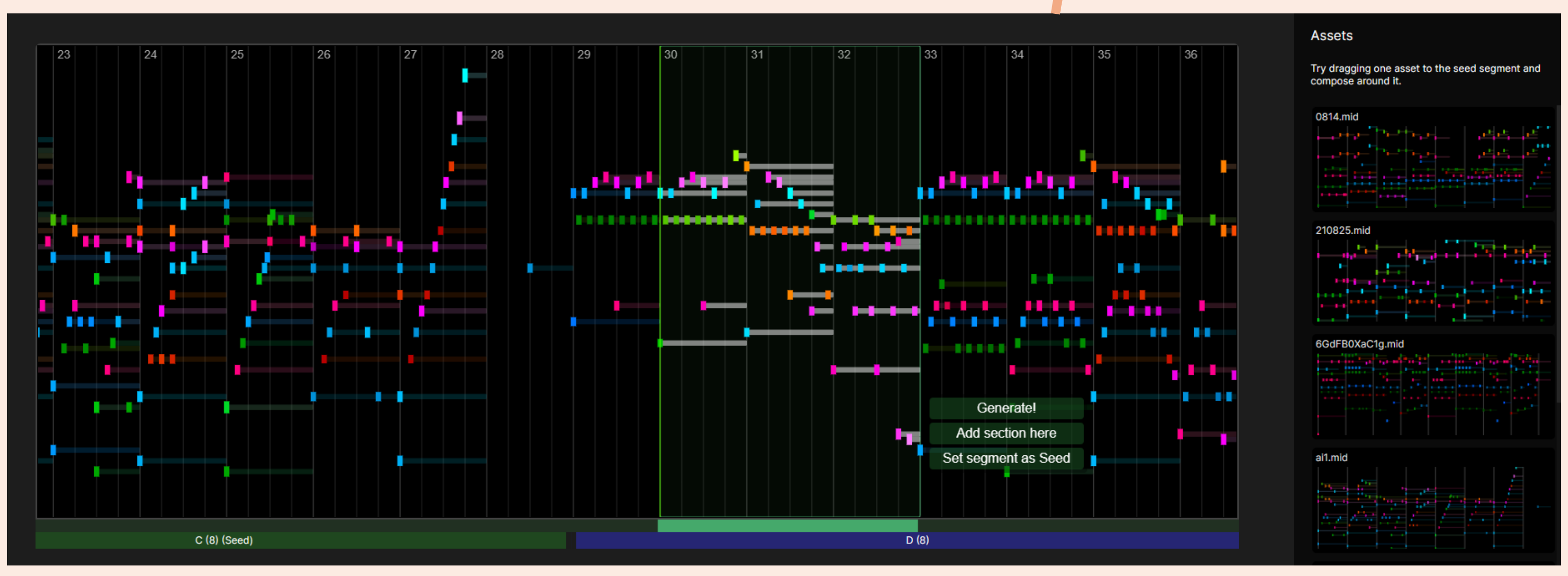
Baselines:

- *WholeSong* (Wang et al., 2024)
- *Flat* (GPT-like, no structure and seed condition)

- Inference speed measured on an RTX4090
- Structureness Indicator (**SI**) from Wu and Yang (2020)
- User study
  - 44 participants (21 amateur, 19 experienced, 4 professional)
  - 5-point scale for Adherence to Seed (**A**) and Overall Quality (**O**)

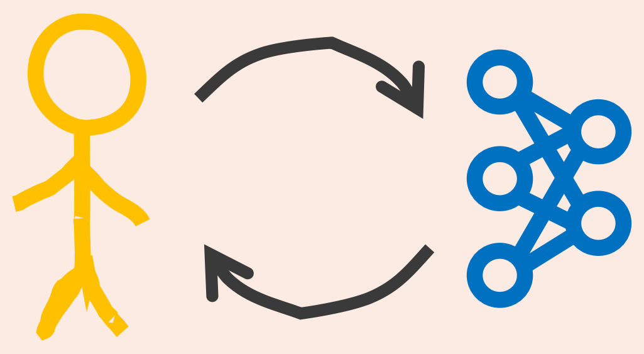
Interactive Interface

Available on GitHub



Collaborate on a piano roll

- Determine structure and Seed
- Compose a music fragment manually
- Edit AI-generated content



Generate music fragments on request

Flexible generation order → Revise previous content at any time

Fast enough → Real-time generation streaming at ~120 bpm