

BEAT: Tokenizing and Generating Symbolic Music by Uniform Temporal Steps

ICML 2026 · Seoul

Lekai Qian¹ Haoyu Gu¹ Jingwei Zhao² Ziyu Wang^{3,4}

¹South China University of Technology ²National University of Singapore ³MBZUAI ⁴New York University

1 Motivation

Symbolic music can be viewed as **events**, **notation**, or a **grid**. Language models almost always consume the event view.

The piano-roll is just as natural, but treating it as an **image** for diffusion struggles with length, sparsity, and multi-track meter markers, so the grid route stayed unpopular.

Vision met the same problem with two routes: model the raw grid, or **tokenize then autoregress** like VQ-VAE.

So we ask: can the piano-roll be tokenized the same way, with a single beat as the basic token?

2 BEAT Encoding

Music becomes a 3-state piano-roll: **0** silence, **1** onset, **2** sustain; a beat spans τ steps.

Step 1 · Pattern. Each pitch's τ -step state becomes one base-3 **PAT** token, plus a mean **VEL**.

Step 2 · Beat. Only **active pitches** are kept; **PIT** gives the first pitch absolutely, the rest as intervals.

Step 3 · Sequence. BEAT and BAR markers, per-track **INS**, tempo and time signature.

Beats are **localized**, **compact**, and **approximately invariant** to transposition and time-shift.

3 Results

Task: symbolic music continuation

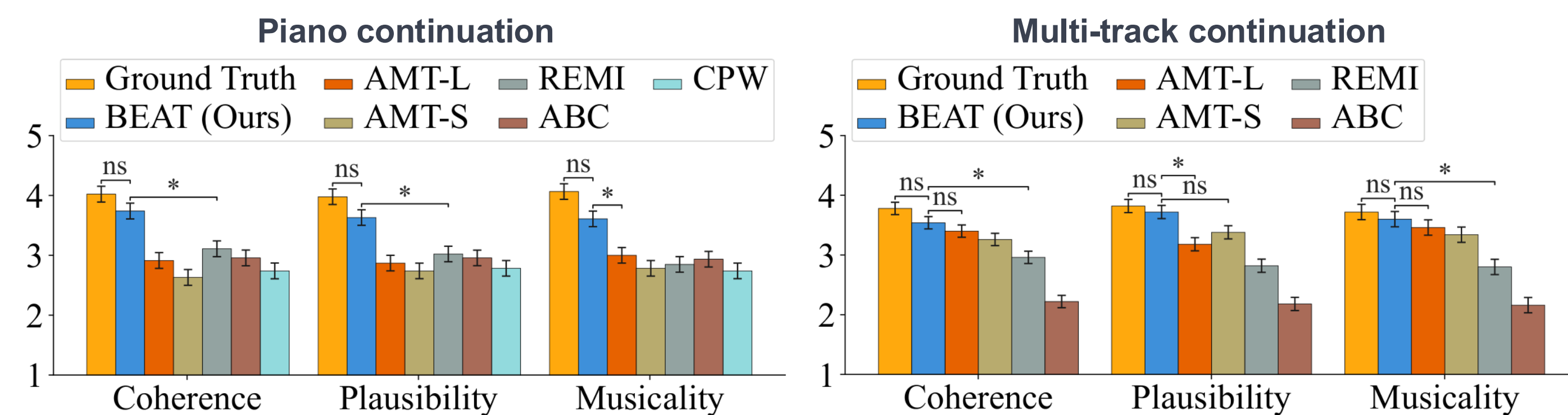
Objective · lower is better

BEAT is best on JS-GC and FMD for both piano and multi-track.

Method	Piano			Multi-track		
	GC	SC	FMD	GC	SC	FMD
Int. ABC	.677	.023	522	.594	.036	580
REMI(+)	.552	.038	551	.313	.008	463
CPW	.634	.024	587	—	—	—
AMT-S	.603	.055	447	.358	.078	449
AMT-L	.625	.053	445	.353	.029	442
Naive PR	.051	.106	582	—	—	—
BEAT (Ours)	.039	.021	437	.043	.009	421

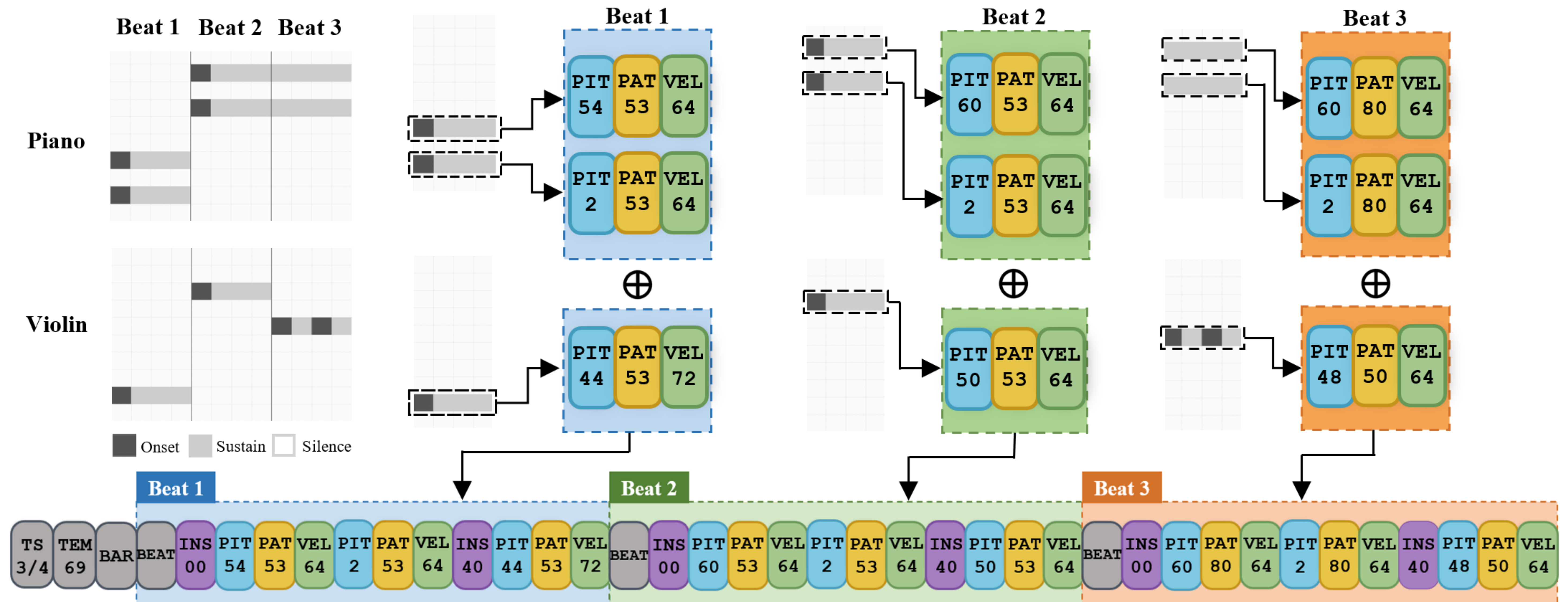
Subjective study · 32 listeners

On piano BEAT significantly outperforms every baseline; on multi-track it edges out AMT. The gap to ground truth is not significant.



Takeaway: A uniform beat is a simple but effective inductive bias. BEAT keeps the compactness of event tokens, adds the temporal regularity of the grid, and supports real-time control, all in one standard autoregressive framework.

BEAT Encoding · Framework



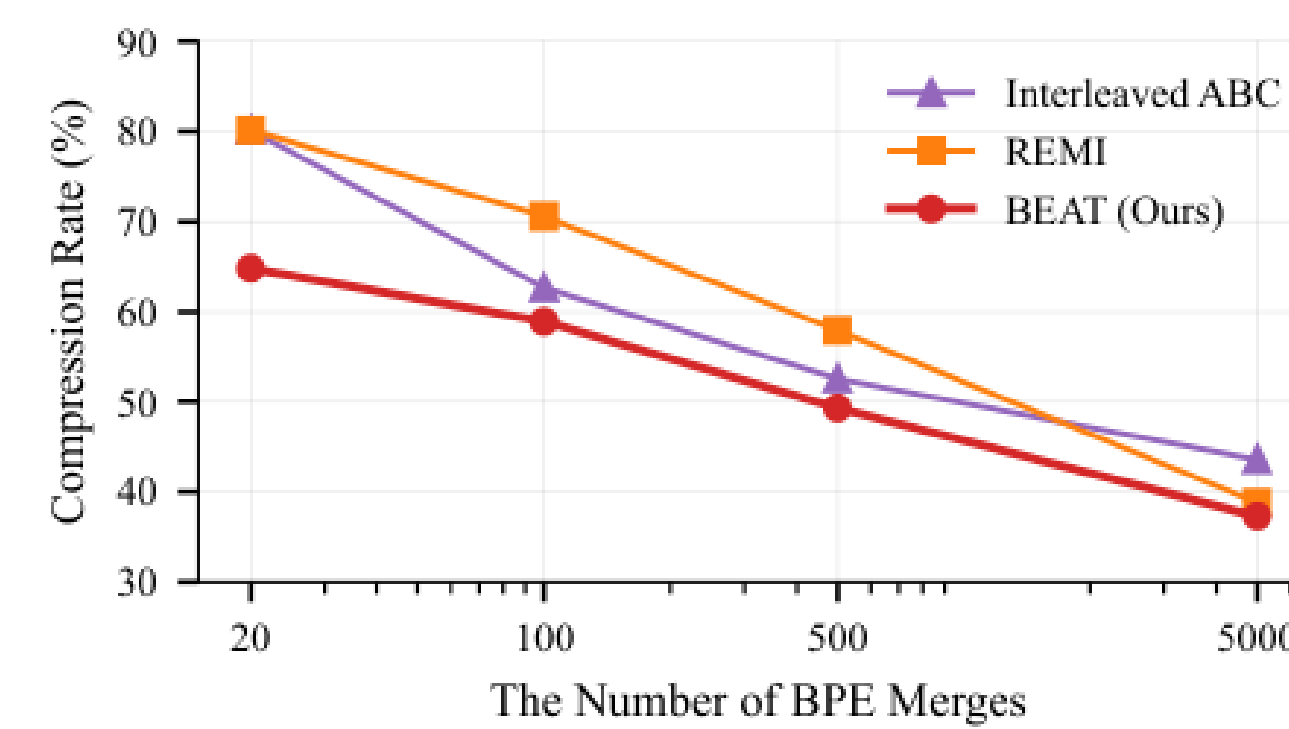
Each beat's active pitches become (PIT, PAT, VEL) triples. PIT gives the first pitch absolutely then relative intervals, PAT is the base-3 onset/sustain pattern over τ steps, and VEL is the mean velocity. Beats are then concatenated with BEAT, BAR, INS and tempo/time-signature markers into a single sequence.

4 Analysis

Compact, yet regular.

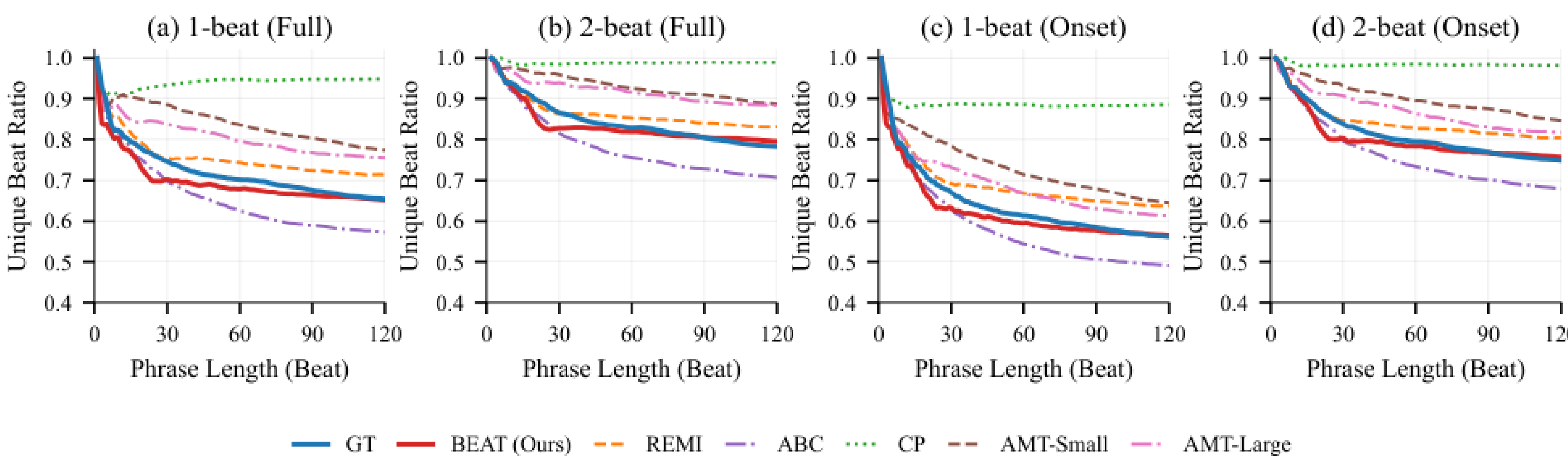
BEAT yields the shortest sequences, 16× shorter than a naive piano-roll. Its lower BPE rate reveals more reusable substructure.

Method	Avg. length
Int. ABC	3450.4
REMI	1902.6
AMT	1950.6
Naive PR	29340.9
BEAT (Ours)	1825.6



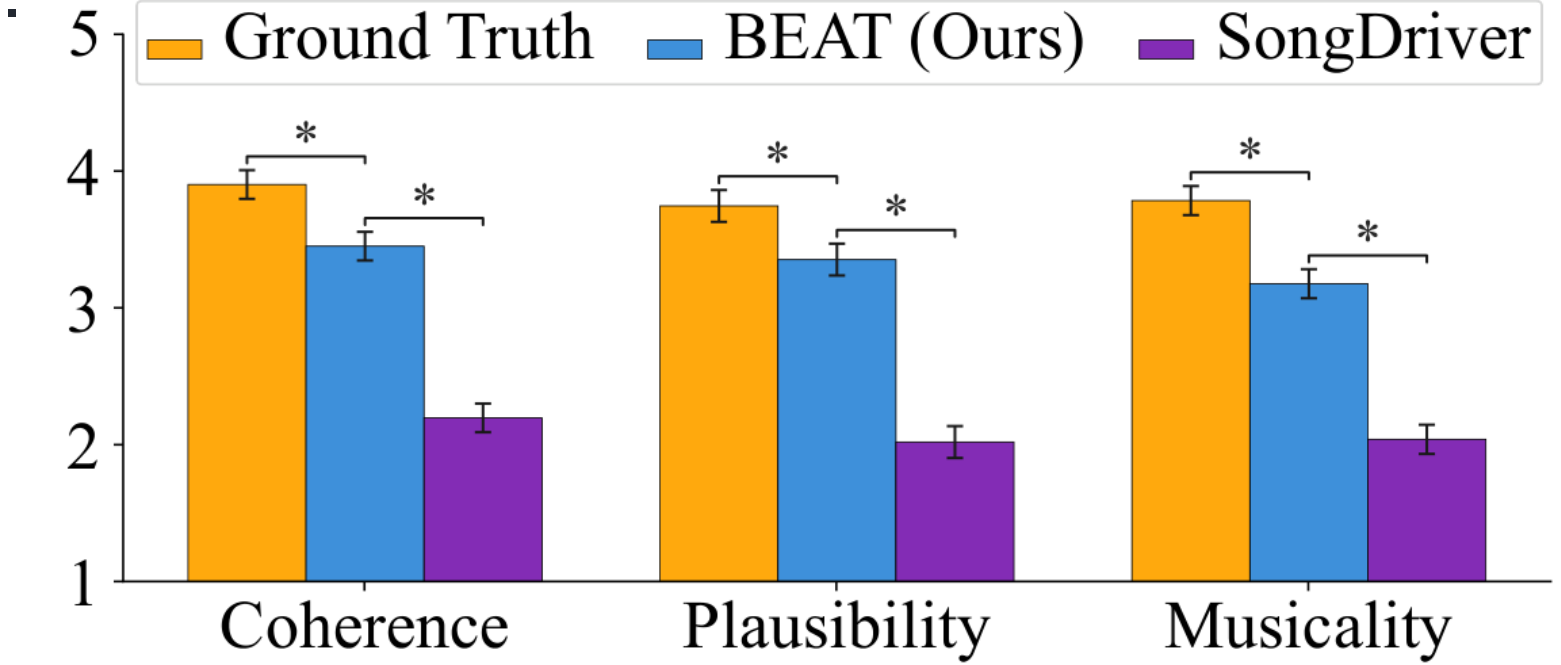
Long-term coherence · unique-beat growth

BEAT tracks ground truth within about 1%; other methods drift toward over-repetition or over-novelty.



Real-time by construction.

Strictly causal, beat-synchronized generation follows the melody in real time, which variable-duration event tokens struggle to do. BEAT significantly outperforms SongDriver.



Conclusions

- **Beat as the token:** event-level compactness with the piano-roll's explicit temporal regularity.
- **Musical priors built in:** invariance to transposition and time-shift lives in the tokens, not in weights.
- **Real-time for free:** the uniform temporal structure natively supports streaming accompaniment.
- **Perceptually validated:** listening study and repetition-diversity analysis confirm coherent structure.

Scan for demo

