

SONGMAMBA: EFFICIENT MUSIC STRUCTURE ANALYSIS VIA BIDIRECTIONAL STATE-SPACE MODELING

Yuan Wang¹, Zhiping Qiu², Shenyang Xu³, Yan Han¹

Yuanfeng Jiang¹, Xiaobing Li¹, Wei Li⁴

¹Central Conservatory of Music, Beijing, China; ²BNRist, Tsinghua University, Beijing, China

³Xinghai Conservatory of Music, Guangzhou, China; ⁴Fudan University, Shanghai, China

amateur@mail.ccom.edu.cn weili-fudan@fudan.edu.cn

ABSTRACT

Music structure analysis requires precise boundaries and extended context, while dense attention becomes costly for long recordings. SongMamba combines multi-resolution MusicFM and MuQ embeddings with bidirectional selective state-space encoding for boundary detection and section labeling. Ablations motivate a compact four-block encoder with linear sequence complexity. On SongFormBench-HX, it achieves state-of-the-art boundary detection with untrimmed F1 of 0.7412/0.8394 at 0.5/3 seconds, exceeding SongFormer’s reported 0.703/0.784. Downstream inference at 420 seconds is 1.84 times faster with 48.1% lower peak GPU allocation. A local-strength decoder preserves nearby transitions; fixed-checkpoint evidence shows that postprocessing can substantially change scores.

1. MOTIVATION

Music structure analysis supports navigation, retrieval, and summarization [1]. Interpreting ambiguous transitions requires local evidence and extended context. SongFormer [2] supplies global attention at quadratic frame-to-frame cost.

Mamba [3] instead propagates compact, input-dependent states. Its causal scan lacks following context; inspired by bidirectional beat tracking [4], we combine forward and backward scans while retaining SongFormer’s frontend and task.

2. MODEL AND DECODING

SongMamba combines multi-resolution inputs, bidirectional encoding, and local decoding for structural prediction.

Multi-resolution inputs. MusicFM [5] and MuQ [6] each provide 1024-dimensional streams with 30/420-

second context, following SongFormer. Concatenation yields 4096 dimensions at 25 Hz. Projection to 2048, stride-three downsampling to 1024, and projection to 512 yield 8.33 output frames per second (Figure 1).

Bidirectional encoding. Independent Mamba-1 modules process original and reversed sequences. With time reversal R and $U = \text{LN}(H)$, aligned outputs are

$$F = \text{Mamba}_{\rightarrow}(U), \quad B = R(\text{Mamba}_{\leftarrow}(R(U))). \quad (1)$$

State dimension is 16, convolution width four, and expansion factor two. Fusion uses $M = F + B$, projected concatenation $M = W_c[F; B] + b_c$, or a frame/channel gate:

$$G = \sigma(W_g[F; B] + b_g), \quad (2)$$
$$M = 2(G \odot F + (1 - G) \odot B).$$

Initialization at $G = 0.5$ or $W_c = [I \ I]$ reproduces summation at width 512. Pre-normalized residual mixer and feed-forward sublayers use dropout 0.1; the latter expands by four with GELU. Encoder complexity is linear, excluding embedding extraction.

Learning. Sigmoid and softmax heads predict boundaries and seven HX functions. Boundary BCE with total-variation regularization and function cross-entropy with focal loss are weighted 0.2/0.8. AdamW uses learning rate 10^{-4} , batch size four, warmup, and cosine decay; EMA validation selects checkpoints.

Decoding. Local-strength peak picking uses an approximately 1.4-second full window (11 frames), threshold 0.01, and one-second minimum spacing and endpoint exclusion. Majority voting assigns segment labels.

3. EXPERIMENTS

Fusion, capacity, and efficiency studies guide model selection alongside system and decoder comparisons.

Protocol. SongFormBench-HX uses 512/200/200 training/validation/test songs [2, 7]. Ablations average 203 valid chunks with annotation-assisted extents and shared decoding. $F_{.5}/F_3$ exclude outer boundaries at 0.5/3-second tolerances; PWF measures grouping [8]. System evaluation averages 200 songs using ACC at 10 Hz and untrimmed HR.5F/HR3F.

Directional fusion. Forward-only and bidirectional sum/gate/concat variants test following context and fusion choice over three runs.



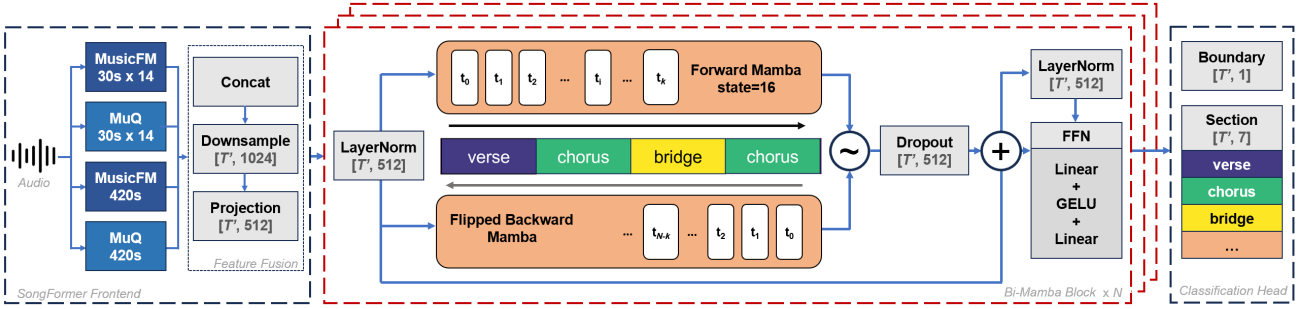


Figure 1. SongMamba: multi-resolution embeddings, bidirectional encoding, and joint structural prediction.

Model	$F_{.5}$	F_3	PWF
SongMamba, forward	.7045	.8014	.7557
SongMamba, sum	.7402	.8395	.7788
SongMamba, gate	.7419	.8470	.7832
SongMamba, concat	.7364	.8424	.7742

Table 1. HX fusion ablation (trimmed boundary F1).

Configuration	$F_{.5}$	F_3	PWF
Depth: 2 blocks	.7380	.8409	.7717
Depth: 3 blocks	.7366	.8385	.7851
Depth: 4 blocks	.7419	.8470	.7832
State: 16	.7412	.8411	.7748
State: 32	.7113	.8253	.7651
State: 64	.7324	.8401	.7782

Table 2. HX capacity ablations (best in bold).

Model capacity. Separate studies vary depth (2/3/4 blocks) and state size (16/32/64); the latter fixes four blocks with sum fusion.

Efficiency. Identical 420-second inputs on an H800 (batch one) compare median downstream latency over 60 alternating CUDA-event timings after five warmups. Peak allocation is measured separately; embedding extraction is excluded.

4. RESULTS AND DISCUSSION

The results guide model selection and show how efficiency and postprocessing shape system performance.

Directional fusion. Sum gains 3.57 $F_{.5}$ points over forward-only Mamba (Table 1), favoring bidirectionality. Gating adds 0.17 points; concatenation offers no consistent gain. Removing a direction reduces parameters; this study is not parameter-matched.

Model capacity. Four blocks lead on boundaries (Table 2). State 64 improves grouping but loses strict boundary accuracy; state 32 underperforms. We select four blocks and state 16; the MIREX checkpoint retains parameter-free sum fusion.

Efficiency. SongMamba takes 22.87 ms versus SongFormer’s 41.99 ms (1.84-fold faster), with 657 versus 1264 MiB peak allocation (48.1% lower).

System	ACC	HR.5F	HR3F
SongFormer, reported	.7930	.7030	.7840
SongMamba, sum	.7835	.7412	.8394

Table 3. HX untrimmed results; SongFormer reported [2].

Boundary detection. SongMamba exceeds reported SongFormer HR.5F/HR3F by 3.82/5.54 percentage points (Table 3); ACC is 0.95 points lower. Decoding differs, so these system gains do not isolate the encoder.

Postprocessing. On the Harmonix-derived HX test set, 444 of 2031 internal boundaries (21.86%) have a neighbor within 11.88 seconds. SongFormer’s approximately 12-second competition radius rejects the weaker of two nearby peaks even when both represent genuine transitions, reducing recall. We therefore use an approximately 1.4-second full window to retain close transitions. With a fixed SongFormer checkpoint, recall at 0.5 seconds on these 444 boundaries rises from 44.14% to 71.17% (+27.03 points). Postprocessing thus dramatically changes results without retraining. Multiple rules change, not only window size.

Trimming. The same outputs yield trimmed HR.5F/HR3F of .7372/.8302, down 0.40/0.92 points. Aggregation differs from chunk ablations. We advocate double-trim evaluation [9]: remove leading/trailing silent segments, then exclude active-span outer boundaries from reference and prediction. Explicit decoder and silence-handling rules prevent trivial endpoint matches from obscuring internal boundary quality.

5. CONCLUSION

SongMamba establishes a new state of the art in boundary detection on SongFormBench-HX, substantially outperforming MIREX 2025 systems and the reported SongFormer baseline. Faster inference, lower memory use, and linear encoder scaling offer potential for longer whole-song processing with less fragmentation. These results open new directions for efficient state-space MSA. Beyond model design, more robust post-processing and rigorous evaluation protocols—notably double trimming—remain essential for MSA. We hope our findings provide both a solid foundation and new momentum for future MSA research.

6. REFERENCES

- [1] O. Nieto, G. J. Mysore, C.-i. Wang, et al., “Audio-based music structure analysis: Current trends, open challenges, and applications,” *Trans. Int. Soc. Music Inf. Retrieval*, vol. 3, no. 1, pp. 246–263, 2020.
- [2] C. Hao, R. Yuan, J. Yao, et al., “SongFormer: Scaling music structure analysis with heterogeneous supervision,” in *Proc. IEEE ICME*, 2026.
- [3] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” in *Proc. COLM*, 2024.
- [4] G. Ru, Y. Yu, and W. Li, “BeatMamba: Bidirectional selective state-space modeling for efficient beat tracking,” in *Proc. IEEE ICASSP*, 2026.
- [5] M. Won, Y.-N. Hung, and D. Le, “A foundation model for music informatics,” in *Proc. IEEE ICASSP*, 2024, pp. 1226–1230.
- [6] H. Zhu, Y. Zhou, H. Chen, et al., “MuQ: Self-supervised music representation learning with mel residual vector quantization,” *arXiv:2501.01108*, 2025.
- [7] O. Nieto, M. C. McCallum, M. E. P. Davies, et al., “The Harmonix Set: Beats, downbeats, and functional segment annotations of Western popular music,” in *Proc. ISMIR*, 2019, pp. 565–572.
- [8] C. Raffel, B. McFee, E. J. Humphrey, et al., “mir_eval: A transparent implementation of common MIR metrics,” in *Proc. ISMIR*, 2014, pp. 367–372.
- [9] A. Marmoret, “Unsupervised evaluation of deep audio embeddings for music structure analysis,” in *Proc. Sound and Music Computing Conf.*, 2026.