DAMN: DELTA-ABSOLUTE MODELING WITH NESTED TRANSFORMERS FOR SYMBOLIC MUSIC GENERATION

Langchu Huang **Yang Xu**
Southern University of Science and Technology
langchuhuang@gmail.com xuyang@sustech.edu.cn

ABSTRACT

We describe DAMN (Delta-Absolute Modeling with Nested Transformers), a compact hierarchical autoregressive model for symbolic piano continuation in the MIREX 2026 Symbolic Music Generation task. Notes are represented as triplet tokens and decoded by two nested Transformers: a score-level decoder models dependencies among notes, while a note-level decoder predicts the three attributes sequentially. To expose metrical distance without sacrificing local rhythmic patterns, the score-level rotary position indices use absolute onset positions, whereas token embeddings and targets retain delta-onset values. The submitted configuration has 144M parameters and is trained from scratch on 8.2M 16-bar piano windows.

1. INTRODUCTION

Hierarchical sequence models provide a natural inductive bias for music, whose structure spans multiple interacting levels and internal dimensions [1, 2]. DAMN follows this idea with a nested autoregressive architecture that separates inter-note dependency modeling from intra-note attribute prediction. We further combine delta-onset tokens [3] with absolute-onset rotary position indices [4] to preserve local rhythmic patterns while exposing longer-range metrical distances.

The MIREX 2026 Symbolic Music Generation task considered here focuses on 4/4 symbolic piano continuation. Given a preceding 4-bar piano context with an optional pickup measure, the system generates a 12-bar continuation. The following sections describe our representation, model architecture, and training setup.

2. METHOD

2.1 Representation

After sorting notes by onset and then by pitch, note i is encoded as a triplet $x_i = (\Delta o_i, p_i, d_i)$, where

$$\Delta o_i = \min(o_i - o_{i-1}, 127), \quad d_i = \min(D_i, 127). \quad (1)$$

© Langchu Huang and Yang Xu. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Langchu Huang and Yang Xu, "DAMN: Delta-Absolute Modeling with Nested Transformers for Symbolic Music Generation", submitted to *ISMIR LBD*, 2026.

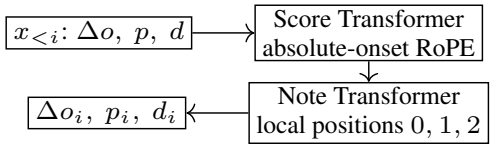

Figure 1. Overview of the DAMN architecture.

Here D_i denotes the original quantized duration. All three attribute vocabularies therefore contain 128 values. Notes at the same onset are ordered by increasing pitch, giving each chord a deterministic serialization.

2.2 Nested Decoders

Figure 1 summarizes the model architecture. Separate embeddings for delta-onset, pitch, and duration are summed to form each note representation. A learned BOS vector right-shifts the note sequence before it is passed through a causal, randomly initialized Qwen3-style decoder [5].

For each score-level input position i, the RoPE position index is defined as

$$r_i = \begin{cases} 0, & i = 0 \quad \text{(BOS)}, \\ o_{i-1} + 1, & i > 0, \end{cases}$$

so the onset of the current target note is never exposed. Notes sharing the same onset therefore share the same RoPE position index.

The score-level decoder produces a context vector

$$h_i = f_{\theta_s}(x_{<i}, r_{\le i}),$$

which conditions a three-position note-level decoder with the factorization

$$\begin{aligned} p_{\theta_n}(x_i \mid h_i) = {} & p_{\theta_n}(\Delta o_i \mid h_i) \\ & \cdot p_{\theta_n}(p_i \mid h_i, \Delta o_i) \\ & \cdot p_{\theta_n}(d_i \mid h_i, \Delta o_i, p_i). \end{aligned}$$

The embeddings are shared across the two decoders, and all attribute output weights are tied to their corresponding embeddings.

The training objective is the equally weighted average of the three attribute-wise cross-entropy losses:

$$\mathcal{L} = \frac{1}{3}\left(\mathcal{L}_{\Delta o} + \mathcal{L}_p + \mathcal{L}_d\right).$$

System	Params.	Similarity ↑
DAMN	144M	**0.744** [0.728, 0.760]
AMT [8]	780M	0.725 [0.708, 0.740]
DAMN-Mini	20M	0.719 [0.702, 0.736]
DAMN-Mini*	20M	0.713 [0.695, 0.730]
RWKV [9]	20M	0.706 [0.690, 0.721]
MuseCoco [10]	1.2B	0.672 [0.659, 0.686]

Table 1. MetaScore-SFT-100 CLaMP 3 similarity.
*Uses sequential RoPE positions.

3. DATA AND TRAINING

We construct a preprocessing pool of approximately 2.4 million MIDI files and select piano material before segmentation. Each piece is quantized to a sixteenth-note grid and partitioned into complete, non-overlapping 16-bar windows. We retain windows containing 32–512 notes, resulting in a training corpus of 8.2 million windows.

We deliberately do not restrict the training corpus to 4/4 meter. Although the model receives no explicit time-signature token, the prompt provides rhythmic evidence. We expect exposure to varied meters to improve rhythmic diversity and robustness.

We train the model from scratch for five epochs using AdamW with BF16 precision. The score-level decoder has 14 layers, while the note-level decoder has 2 layers; both share the same hidden size of 768. The learning rate follows a warmup–constant–cosine schedule with a peak value of 3×10^{-4}.

4. EVALUATION

We construct MetaScore-SFT-100 from a genre-stratified sample of MetaScore-Genre [6]. Automatic filtering retains 106 single-part piano scores in 4/4 with at least 16 measures, no lyrics, and reliable user ratings. After removing six clear quality-control failures by manual listening, 100 pieces remain across multiple genres. All benchmark pieces are excluded from the training set. Each system generates eight continuations per prompt.

We compute CLaMP 3 embeddings [7], compare each continuation with its reference by cosine similarity, and average over all candidates and prompts. We report 95% prompt-level bootstrap intervals. CLaMP 3 similarity serves only as a development proxy.

5. CONCLUSION

We presented DAMN, a nested autoregressive model for symbolic piano continuation. The model combines score-level note modeling with note-level attribute prediction and uses absolute-onset rotary positions alongside delta-onset tokens. Among the evaluated systems, DAMN achieves the highest CLaMP 3 similarity on our MetaScore-SFT-100 development benchmark. We submit DAMN to the MIREX 2026 Symbolic Music Generation task for official listening-based evaluation.

6. REFERENCES

[1] J. Ryu, H.-W. Dong, J. Jung, and D. Jeong, "Nested music transformer: Sequentially decoding compound tokens in symbolic music and audio generation," *arXiv preprint arXiv:2408.01180*, 2024.

[2] Y. Wang, S. Wu, J. Hu, X. Du, Y. Peng, Y. Huang, S. Fan, X. Li, F. Yu, and M. Sun, "NotaGen: Advancing musicality in symbolic music generation with large language model training paradigms," *arXiv preprint arXiv:2502.18008*, 2025.

[3] Z. Guo and S. Dixon, "Moonbeam: A MIDI foundation model using both absolute and relative music attributes," *arXiv preprint arXiv:2505.15559*, 2025. [Online]. Available: https://arxiv.org/abs/2505.15559

[4] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[5] A. Yang, A. Li, B. Yang *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[6] W. Xu, J. McAuley, T. Berg-Kirkpatrick, S. Dubnov, and H.-W. Dong, "Generating symbolic music from natural language prompts using an LLM-enhanced dataset," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2025, pp. 215–222. [Online]. Available: https://arxiv.org/abs/2410.02084

[7] S. Wu, Z. Guo, R. Yuan, J. Jiang, S. Doh, G. Xia, J. Nam, X. Li, F. Yu, and M. Sun, "CLaMP 3: Universal music information retrieval across unaligned modalities and unseen languages," in *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 2025, pp. 2605–2625. [Online]. Available: https://aclanthology.org/2025.findings-acl.133/

[8] J. Thickstun, D. Hall, C. Donahue, and P. Liang, "Anticipatory music transformer," *arXiv preprint arXiv:2306.08620*, 2023.

[9] C. Zhou-Zheng, J. Backsund, D. L. Chan, A. Coventry, A. Eslami, J. Goel, X. Han, D. Soomro, and G. Wei, "A traditional approach to symbolic piano continuation," in *Extended Abstracts for the Late-Breaking Demo Session of the 26th Int. Society for Music Information Retrieval Conf.*, 2025. [Online]. Available: https://futuremirex.com/portal/wp-content/uploads/2025/symbolic-music-generation/RWKV.pdf

[10] P. Lu, X. Xu, C. Kang, B. Yu, C. Xing, X. Tan, and J. Bian, "MuseCoco: Generating symbolic music from text," *arXiv preprint arXiv:2306.00110*, 2023.