INFERENCE-TIME CONTROL OF TONAL TENSION IN SYMBOLIC MUSIC GENERATION

Anonymous Authors
Anonymous Affiliations
anonymous@ismir.net

ABSTRACT

Large language models (LLMs) and Transformer-based architectures have achieved remarkable progress in symbolic music generation, producing outputs with increasing coherence, stylistic richness, and expressive depth. Controllability in symbolic music generation is essential for aligning outputs with compositional intent and user-specified goals. Among high-level perceptual attributes, tonal tension remains underexplored for explicit control. In this work, we present a novel approach that integrates a computational model of tonal tension into a transformer generation framework through a dual-level beam search strategy. At the token level, candidate continuations are re-ranked for probability and diversity, while at the bar level, tension similarity measures ensure alignment with a target tension curve. Preliminary evaluations indicate that this method enables explicit control of tonal tension while maintaining overall musical quality and coherence. This contributes to the broader effort of aligning LLMs with creative control, and highlights tonal tension as an underexplored but musically salient axis of controllability.

1. BACKGROUND

Symbolic music generation has progressed rapidly with Transformer-based architectures, which now produce long, coherent sequences comparable to natural language outputs [1, 2]. Beyond fluency, however, musicians require the ability to control music generative systems toward specific stylistic [3] or expressive goals [4–6]. Although there are various controllable generation methods, explicit control over tonal tension in symbolic music models remains rare.

Several approaches to modeling tonal tension have been proposed. Lerdahl's cognitive model [7] is conceptually rich but difficult to implement due to manual hierarchies and parameter settings. Systems such as MorpheuS [8, 9] employ Spiral Array–based features like dissonance or tensile strain, yet these often overlook deeper harmonic structure. By contrast, the model based on Tonal Interval Vector (TIV) [10, 11] offers a perceptually grounded approach by projecting chroma features into an interval space where harmonic similarity and chordal relationships are well represented. However, previous applications of TIV to generative tasks have been limited to offline optimization [12], restricting their use for real-time or interactive composition.

Inference-time control techniques provide a promising alternative to training-based methods. Approaches such as beam search [13] or surprisal profile matching [14] demonstrate that external constraints can be dynamically enforced during decoding without retraining. Building on these ideas, our work integrates target tension curves into Transformer-based music generation by combining two complementary strategies: local sequence quality is maintained through token-level re-ranking based on model probability and diversity, while global expressive control is introduced by re-ranking complete bars against a specified tension trajectory. This framework illustrates how tonal tension can function as a perceptually meaningful and practically effective control signal for LLM-based symbolic music generation.

2. METHOD

We represent music using REMI+ [3], an extension of the REMI format [15] that encodes notes with position, pitch, velocity, duration, and adds tokens for chords, tempo, instruments, and time signatures. A Transformer is trained in a translation-style setup [3], where bar-level control tokens condition the decoder to generate REMI+ sequences. We include time signature, instrument list, and note density as control tokens, and introduce *tonal tension* as an additional conditioning feature.

To model tonal tension, we use the TIV framework introduced by Bernardes et al. [10], which projects chroma features into a Tonal Interval Space (TIS) where harmonic relations are represented more perceptually. Building on this representation, we employ the computational model proposed by Navarro et al. [11], which combines three components: (i) *tonal distance*, capturing relationships between chords, keys, and tonal functions; (ii) *dissonance*, reflecting internal chordal roughness; and (iii) *voice-leading*, measuring melodic stability across successive chords. A weighted combination of these components yields an efficient descriptor of harmonic tension, which we use both as a control signal during training and as a scoring function at inference time.

Model	Inference	Instrument F1	Note Density	Groove Sim	Tension Corr
Baseline	normal	0.82	0.88	0.52	0.16
Baseline + Tension	normal	0.83	0.62	0.54	0.18
Baseline + Tension	Dual Beam	0.86	0.85	0.56	0.50

Table 1. Objective evaluation metrics comparing baseline Transformer models (with and without tension conditioning) against our dual-level beam search inference strategy. Results for 8-bar samples represent performance on the entire test set for the first-best ranked candidates from our dual-level beam search inference

The core contribution is a dual-level beam search strategy. At the token level, candidate continuations are expanded and re-ranked using a balance of model probability and diversity, ensuring music quality. At the bar level, once a bar is completed, candidates are additionally evaluated against a target tonal tension curve computed with TIV. This allows us to control the generated music toward rising, falling, or arch-shaped patterns of tension and release.

Although we introduce tension as a control token during training, its effect is limited since tonal tension is non-differentiable and cannot be directly optimized with a loss. The primary mechanism of control is therefore applied at inference through our dual-level beam search. This design keeps the approach modular and largely model-agnostic: because re-ranking operates only during decoding, it can be applied to a variety of symbolic music LLMs without retraining. The separation between token-level control and bar-level control reflects musical fluency at the note and chord level, and global tension shaping at the bar level. Retaining multiple beams further provides composers with diverse musical generations under the same target curve.

3. PRELIMINARY EXPERIMENTS

3.1 Setup

We trained on the Lakh MIDI-Matched dataset [16], preprocessing files with Midi Miner [17] to extract chordal tracks, yielding 25,555 usable MIDI pieces. Data were split into 0.85, 0.10, and 0.05 proportions for training, validation, and test sets. Our Transformer model (512 dimensions, 12 heads, 4 encoder and 6 decoder layers, max length 256) was trained for 12 epochs with cross-entropy loss on an NVIDIA A40 GPU.

At inference, we combined nucleus sampling ($p = 0.9$) with our dual-level beam search. Each step expanded 8 candidates, re-ranked at the token level by probability and diversity ($\lambda = 0.7$). At the bar level, the top 3 beams were selected according to a tension weight of 4.0, with a temperature of 0.9, explicitly shaping the generated sequence toward the desired tension curve.

3.2 Objective Evaluation

We assess the impact of tension control via our dual-level beam search using four metrics. Instrument F1, Note Density, Groove Similarity, and Tension Correlation. Results for 8-bar samples are shown in Table 1.

The baseline Transformer achieves strong instrument accuracy (0.82) and relatively stable rhythmic patterns (0.52 groove similarity), but exhibits weak control over tonal tension, with a low correlation of 0.16 against the target curves. Adding tension tokens during training yields only a marginal improvement (0.18), confirming that token-based conditioning is insufficient when the underlying descriptor is non-differentiable and complex.

In contrast, our proposed dual-level beam search achieves a substantial leap in tension alignment, with correlation rising to 0.50. This represents more than a twofold increase compared to the baseline, indicating that explicit inference-time re-ranking is essential for shaping expressive trajectories. Importantly, this improvement is not achieved at the expense of other qualities: instrument accuracy improves to 0.86, groove similarity increases slightly to 0.56, and note density remains close to the baseline level (0.85 vs. 0.88). These results demonstrate that tension shaping can be integrated while preserving musical quality.

Taken together, the results suggest that inference-time control not only enhances expressive accuracy but also supports a flexible design that could generalize to other non-differentiable musical descriptors. Moreover, the preservation of note density and groove implies that the method does not compromise the naturalness of the generated music, a key requirement for composer-facing applications.

4. CONCLUSION

We introduced a dual-level beam search method that integrates a tonal tension model into Transformer-based symbolic music generation. The approach enables inference-time control of tonal tension while preserving music quality and diversity. Preliminary results demonstrate its promise, and ongoing efforts aim to improve scalability, perceptual accuracy, and cross-repertoire applicability. We view this as a step toward practical, controllable LLMs for music that empower human creators with nuanced expressive tools.

Next directions include: (i) scaling to longer pieces where maintaining global control is more challenging, (ii) developing interactive composer interfaces that visualize and edit tension curves in real time, (iii) extending control to other expressive dimensions such as rhythmic complexity or emotional trajectory, (iv) exploring multimodal prompting scenarios that combine text and symbolic constraints, and (v) evaluating performance across culturally diverse repertoires to ensure inclusivity in generative outcomes.

5. REFERENCES

[1] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 2, pp. 604–624, 2020.

[2] L. Deng and Y. Liu, *Deep learning in natural language processing*. Springer, 2018.

[3] D. von Rütte, L. Biggio, Y. Kilcher, and T. Hofmann, "Figaro: Controllable music generation using learned and expert features," in *The Eleventh International Conference on Learning Representations*, 2023.

[4] S. Tian, C. Zhang, W. Yuan, W. Tan, and W. Zhu, "Xmusic: Towards a generalized and controllable symbolic music generation framework," *arXiv preprint arXiv:2501.08809*, 2025.

[5] D. Makris, K. R. Agres, and D. Herremans, "Generating lead sheets with affect: A novel conditional seq2seq framework," in *2021 international joint conference on neural networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[6] L. N. Ferreira, L. Mou, J. Whitehead, and L. H. Lelis, "Controlling perceived emotion in symbolic music generation with monte carlo tree search," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 18, no. 1, 2022, pp. 163–170.

[7] F. Lerdahl and C. L. Krumhansl, "Modeling tonal tension," *Music perception*, vol. 24, no. 4, pp. 329–366, 2007.

[8] D. Herremans and E. Chew, "Towards emotion based music generation: A tonal tension modelbased on the spiral array," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 41, 2019.

[9] ——, "Tension ribbons: Quantifying and visualising tonal tension," in *International Conference on Technologies for Music Notation and Representation–TENOR'16*, 2016.

[10] G. Bernardes, D. Cocharro, M. Caetano, C. Guedes, and M. E. Davies, "A multi-level tonal interval space for modelling pitch relatedness and musical consonance," *Journal of New Music Research*, vol. 45, no. 4, pp. 281–294, 2016.

[11] M. Navarro-Cáceres, M. Caetano, G. Bernardes, M. Sánchez-Barba, and J. Merchán Sánchez-Jara, "A computational model of tonal tension profile of chord progressions in the tonal interval space," *Entropy*, vol. 22, no. 11, p. 1291, 2020.

[12] M. Navarro-Cáceres, J. F. Merchan Sanchez-Jara, V. Reis Quietinho Leithardt, and R. García-Ovejero, "Assistive model to generate chord progressions using genetic programming with artificial immune properties," *Applied Sciences*, vol. 10, no. 17, p. 6039, 2020.

[13] L. Ferreira, L. Lelis, and J. Whitehead, "Computer-generated music for tabletop role-playing games," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 16, no. 1, 2020, pp. 59–65.

[14] M. R. Bjare, S. Lattner, and G. Widmer, "Controlling surprisal in music generation via information content curve matching," in *ISMIR*, 2024, pp. 922–929.

[15] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1180–1188.

[16] C. Raffel, "Learning-based methods for comparing sequences, with ap- plications to audio-to-midi alignment and matching (unpublished doctoral dissertation). columbia university." 2016.

[17] R. Guo, D. Herremans, and T. Magnusson, "Midi miner–a python library for tonal tension and track classification," *arXiv preprint arXiv:1910.02049*, 2019.