Education Reform and Development, 2026, Volume 8, Issue 10
https://bbwpublisher.com/journal/erd
Online ISSN: 2652-5372
Print ISSN: 2652-5364

Adapting Anticipatory Music Transformer for TurnBased MIDI CallandResponse

Zitong Wang, Sitong Hu

Beijing Middle School, Beijing 100028, China

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: This work adapts a frozen Anticipatory Music Transformer to turnbased MIDI callandresponse without retraining model weights. Centered on external engineeringlevel adaptation, the integrated solution incorporates MIDI phrase boundary detection, phraselevel musical constraints, asynchronous multithreaded scheduling, and speculative preloading, with no modification to underlying model parameters. We conduct comparative evaluation of candidate schemes, A0A6 module ablation studies, and scheduler playback tests based on fixed playback data. A total of 27 000 candidate records and 63 000 ablation records are obtained on the Call100 dataset, alongside 18 000 records from scheduler tests. The mean composite objective score of the controlled AMT rises from 0.561 to 0.733, though the rulebased motive baseline still achieves the highest score at 0.753. Within the scheduler model, the mean latency from confirmed phrase end to readiness of the first generated event decreases from 101.5 ms to 86.5 ms, and the target readinessmiss rate drops from 20.7 % to 5.2 %. These results support inferences drawn from structural proxy metrics and schedulerlevel measurements; they do not equate to validation of actual MIDI transmission, audio onset, or human performer evaluation. No humanparticipant data are analysed or reported in this paper.

Keywords: Symbolic music generation; MIDI; Call-and-response; Transformer; Speculative pre-loading

Online publication: November 25, 2026

1. Introduction

Callandresponse is a ubiquitous performance idiom within improvised music, widely observed in jazz, blues, and diverse folk improvisational excerpts. To computationally realise this turntaking humancomputer musical interaction, systems cannot rely merely on static offline melodic continuation. Generation modules must respond dynamically to realworld tempo fluctuations and pauses produced by human performers.

Thanks to selfattention mechanisms, autoregressive Transformers capture longrange structural dependencies within symbolic music and deliver stable performance for offline melody generation tasks. Nevertheless, their tokenwise serial decoding paradigm is poorly suited for live collaborative humancomputer performance. Direct integration of offline models into performance pipelines introduces practical challenges: misalignment of humanmachine performance timelines, unreliable phraseend localisation, prominent inferencelatency spikes, and structural collapse of generated excerpts under highentropy input conditions.

Such difficulties become more pronounced under shifting performance tempos and extensive freeform pauses typical of improvisational scenarios. Against this background, the present study mitigates these limitations via an external control architecture and builds a verifiable prototype for turnbased MIDI callandresponse without finetuning the underlying neural network.

The main contribution of this paper is an external adaptation framework built upon the offline Anticipatory Music Transformer (AMT). Drawing on pointprocessinformed engineering timers to characterise turn transitions in humancomputer performance, we design a closedloop virtual MIDI pipeline for signal routing, asynchronous decoding, buffered playback, and speculative preloading. Reproducible evidence is gathered via fixedplayback datasets covering overall candidatescheme performance, permodule contribution, and scheduler behaviour.

2. Related work

Early symbolic music generation largely adopted pianoroll or fixed timestep grid representations for music. Such quantisation produces sparse matrices and erases subtle temporal characteristics of live performance such as rubato and swing rhythm [1]. Eventstream MIDI representations alleviate these drawbacks, and the Anticipatory Music Transformer (AMT) further extends support for controllable timedevent generation and anticipation [2]. Even so, representational improvements alone cannot resolve phrasesegmentation challenges arising from frequent chord stacking and adhoc rests in improvisational performance.

Musicgeneration architectures have evolved from LSTM recurrent networks to Transformer models [3]. While AMT enables controllable timedevent generation and anticipation, its released frozen checkpoints lack native support for endpoint detection, undo semantics, turntaking scheduling, or MIDI orchestration required for this project.

Numerous humancomputer collaborative music studies combine Max/MSP interactive interfaces with Python backends for musical interaction [4], validating the practical utility of visual interactive controls [5]. Most existing research, however, seldom addresses how to encapsulate endpointdetection logic, playbackbuffering strategies, and outputfailure fallback mechanisms for offline autoregressive networks without retraining. The current work fills this engineering gap using virtual MIDI routing and asynchronous multithreaded scheduling.

Objective evaluation metrics are suitable for largescale reproducible structural musical analysis yet cannot fully substitute perceptual human listening assessments [6-7]. Within turnbased interactive settings, evaluation must consider not only generated melodic quality but also endpoint judgement and generationready latency. This paper uses structural proxy metrics, ablation studies, and fixedplayback results as complementary forms of evidence.

3. Pointprocessinformed timing and autoregressive generation

This work adopts a continuoustime MIDI event representation. Each musical event record encodes onset time, note duration, and MIDI pitch. Nonstrictlyincreasing time constraints are enforced to accommodate polyphonic chord performance scenarios.

In callandresponse interaction, the system receives humanperformed passages as conditioning context and generates corresponding AI replies. Causallymasked Transformers sequentially predict event onset time, duration, and pitch. Causal masking enables sequential continuation from model outputs alone; unassisted

generation, however, is prone to melodic repetition, tonal drift, or empty output segments, mandating supplementary external control constraints.

We adopt pointprocess survivalfunction heuristics for phraseend timing [8]. Adaptive phraseending thresholds are dynamically computed based on silence duration following the most recent note onset together with locallywindowed performance tempo. These thresholds are engineeringoriented timers driven by local onset density. They neither fit pointprocess likelihood functions nor constitute definitive proof that a phrase has concluded [9].

4. System design for turnbased MIDI callandresponse

4.1. Crossplatform virtual MIDI architecture

A modular virtualMIDIbus design decouples MIDI signal acquisition, audio rendering, neuralnetwork inference, and MIDI output scheduling. Virtual MIDI and digitalaudioworkstation (DAW) pipelines can be configured under Windows and macOS platforms. The overall system is split into a frontend acquisitionrendering module and a backend Python inferencecontrol module. Four sequential functional phases are defined: MIDI acquisition, dynamic endpoint detection, asynchronous generation, and microbuffered playback. Loose coupling among modules facilitates future extensions including endpoint error correction, streaming decoding, and speculative preloading.

4.2. MIDI phraseendpoint detection

Stoppingtime rules derived from pointprocess heuristics are encapsulated within a MIDI voiceactivitydetection module implementing four operational states: listening, candidate, revocation, and confirmation. Realworld performance of chords and rapid arpeggios produces bursts of onset events that inflate estimated performance rates and trigger premature phrase truncation. A 0.08second clustering window merges temporallyadjacent onsets into composite musical events to correct onsetdensity statistics [10-11].

Twostage revocable decision logic is implemented: upon silencecondition triggering, phraseend confirmation is deferred by opening a 0.15second validation window instead of immediate phraseend locking. New incoming note events detected within this window revoke the candidateend state. Only absence of fresh input throughout the validation window formally commits the human call segment. This parameter is an engineering setting balancing premature confirmation and waiting latency in proxy playback experiments; it is not a perceptuallycalibrated threshold and lacks groundtruth human annotation [12-13].

4.3. Asynchronous decoding and speculative preloading

Conventional generation workflows wait for complete responsephrase generation before feeding output to playback pipelines, yielding perceptible latency. Our dualthread architecture with threadsafe shared queues separates inference and playback responsibilities: the inference thread writes decoded musical events into shared queues as they are produced; the playback thread continuously consumes queue entries for MIDI output, parallelising computation and audio playback. An 80ms microbuffer mitigates playback dropout risk.

While confirmation windows reduce misclassification, they introduce additional response latency. Speculative preloading addresses this tradeoff: model inference commences in the background the moment a candidate phraseend is flagged. If phrase termination is subsequently confirmed, precomputed leading tokens are reused. If the candidate state merely corresponds to a transient performance pause, intermediate preloaded

outputs are discarded. This engineering technique trades computational overhead for reduced latency.

4.4. Sampling control and phraselevel constraints

Absolute timestamps and absolute pitch values are converted into relative offsets at input time to reduce dependency on fixed metrical grids. Kernel sampling filters lowprobability anomalous outputs during outputside decoding, and temperature parameters modulate generation exploration [14].

The phrasecontrol layer extracts callsegment features including duration, pitch range, and rhythmic density to formulate responsegeneration constraints. Input preprocessing cleans prompt segments by compressing trailing repetitive notes. Outputside postprocessing validates consecutive pitch repetitions and applies eventlevel repairs when thresholds are violated. Fullphrase motivetransform fallback is activated only upon empty neuralnetwork outputs; fallback operations include motive inversion, transposition, and rhythmic scaling. Generated responses are temporally scaled or compressed to maintain reasonable phrase lengths. The system defaults to chromatic freeimprovisation mode; optional postprocessing style constraints enforce pentatonic scales or regular 4/4time twobar melodic structures. These modules operate purely as postprocessors and do not alter model weights.

5. Experimental design and objective evaluation

Fixedplayback evidence chains are used for candidatescheme comparison, moduleablation attribution, and scheduler playback assessment. Objective metrics adopted herein serve only as structural proxies for musical quality and are not direct substitutes for human perceptual evaluation.

The Call100 dataset comprises 100 humancall phrases combining manuallydesigned stresstest material and public POP909 MIDI excerpts, covering clear motivic material, syncopated rhythms, free tempo, and sparse short fragments. Three approaches are compared: vanilla AMT, controlled AMT, and rulebased motivetransformation baseline. Six parameter presets are defined. Each presetsystemcall combination is repeated 15 times, yielding 27 000 trial records with statistical inference performed across the 100 call instances. Six evaluation dimensions are defined: tonality, rhythm, melody, repetitiveness, pitch diversity, and structural redundancy; dimension scores are weighted to compute composite objective scores.

Experimental results show that the rulebased motive baseline achieves the highest overall composite score, while controlled AMT outperforms vanilla AMT. Controlled AMT exhibits strong performance on duration matching; the motive baseline excels in tonal stability and cadence handling. Bootstrapped 95 % confidence intervals for the compositescore difference between controlled AMT and vanilla AMT across the 100call set do not cross zero.

Ablation experiments A0A6 produce 63 000 candidatelevel records, with external control modules incrementally enabled to observe score variations. Full controlled configuration yields a 0.172 compositescore improvement relative to the bare model. Styleconstraint modules deliver the largest performance gain; repetition suppression and duration matching also contribute positively. Fullphrase motivetransform fallback is never triggered across 9 000 generation runs and therefore contributes no measurable gain.

Scheduler playback tests are executed on a Windows 11 laptop. Measured as latency from confirmed phrase end to firstgeneratedevent readiness within the scheduler model, mean latency reads 101.543 ms without speculative preloading and 86.545 ms with preloading enabled, representing an average reduction of 14.999 ms. Target readinessmiss rate decreases from 20.744 % to 5.244 %. These figures do not represent

measured timings for MIDI transmission, hostDAW reception, audio onset, or stream underrun [15]. Speculative preloading optimises scheduler orchestration but does not accelerate Transformer intrinsic inference speed.

Certain inputs with clear motivic material and tonal questionlike phrasing attain high composite scores, whereas freetempo and notesparse excerpts tend to achieve lower scores. Some categories contain very few call instances; accordingly such observations cannot be generalised to categorywide conclusions. Eventlevel repetition repair interventions fire 3 712 out of 9 000 runs (41.24 %) for controlled AMT; fullphrase motive fallback occurs zero times out of 9 000. Eventlevel intervention indicates controllayer adjustment and is not equivalent to generation failure. No humansubject data are analysed or reported; preliminary exploratory listening trials are excluded from the formal evidence base.

6. Conclusion

This paper presents an external adaptation pipeline that enables a frozen Anticipatory Music Transformer to perform turnbased MIDI callandresponse without model retraining. The solution integrates adaptive phraseboundary detection, virtual MIDI routing, phraselevel musical controls, and a speculative pre-loading scheduler. Fixedplayback experiments demonstrate that external controls improve structural proxy metrics, and speculative preloading reduces schedulermodel firsteventready latency. Nevertheless, the proposed approach does not outperform the rulebased motive baseline across all evaluation dimensions. The presented evidence does not cover realworld MIDI transmission, DAWlevel audio onset, or human collaborative performance; such validations remain for future work.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Boulanger-Lewandowski N, Bengio Y, Vincent P, 2012, Modeling High Dimensional Sequences With Temporal Dependencies: Application to Polyphonic Music Generation and Transcription. Proceedings of the 29th International Conference on Machine Learning: 1159–1166.

[2] Thickstun J, Hall D, Donahue C, et al., 2023, Anticipatory Music Transformer. arXiv preprint arXiv:2306.08620.

[3] Eck D, Schmidhuber J, 2002, Learning the Temporal Structure of Music With LSTM Recurrent Networks. Neural Computation, 14(11): 2607–2625.

[4] Kalkhake T, Dubnov S, 2026, Towards Real Time Human AI Collaborative Music Performance: Accompaniment Generation Using Latent Diffusion Models and Max/MSP. arXiv preprint arXiv:2604.07612.

[5] Privato N, Lampado O, Novello A, 2022, Scramble Live: Real Time Musical Interaction Combining LSTM and Markov Chains. Proceedings of the 19th Sound and Music Computing Conference.

[6] Ji SL, Yang XY, Luo J, 2023, A Survey of Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluation and Challenges. ACM Computing Surveys, 56(1): 1–39.

[7] Lerch A, Arthur C, Kinns B, et al., 2025, A Survey of Evaluation for Music Generation Models. ACM Computing Surveys.

[8] Daley DJ, Vere-Jones D, 2003, An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and

Methods. 2nd ed. New York: Springer.

[9] Dannenberg RB, 1985, Online Algorithms for Real Time Accompaniment. Proceedings of the 1984 International Computer Music Conference: 193–198.

[10] Madsen ST, Widmer G, 2006, MIDI Voice Separation. Proceedings of the 7th International Conference on Music Information Retrieval: 57–60.

[11] Ura S, Simon I, Dillmann S, et al., 2020, Full of Feeling: Learning Expressive Music Performance. Neural Computing and Applications, 32: 9559–9567.

[12] Lago NP, Kong F, 2004, A Quest for Low Latency. Proceedings of the International Computer Music Conference: 33–36.

[13] Rotondi C, Shafi C, Aloisio C, et al., 2016, A Survey of Networked Music Performance Techniques. IEEE Access, 4: 8823–8843.

[14] Holtzman A, Buys J, Du L, et al., 2020, The Curious Case of Neural Text Degeneration. Proceedings of the International Conference on Learning Representations.

[15] Nelson M, Thom B, 2004, A Survey of Real Time MIDI Performance. Proceedings of the International Conference on New Interfaces for Musical Expression: 35–38.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.