Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38423

Research on Generative Artificial Intelligence Assisted Methods for Guzheng Composition

Jingyao Chen

Department of Chinese Instrumental Music, China Conservatory of Music, Beijing, China
wellington589125@gmail.com

Abstract. The creation of Guzheng music involves a coordinated management of modes, melody, phrasing, and performance techniques, whereas the generic symbolic music generator often represents pitch and rhythm better than specific Chinese instrumental characteristics. A framework for controlling the process of generative artificial intelligence creation based on the solo scores with the licensed or open access with REMI event encoding and the six-layer Transformer decoder is created. Modes, tempo, two-bar motif seeds, and tokens of sliding, bending, vibrato, and tremolo techniques are included in the list of generation conditions explicitly. Three configurations are tested on the basis of pitch distribution distance, rhythmic similarity, modal tone ratio, repetition frequency, novelty of n-grams, structural length validity, and evaluation by experts. Full model brings the value of pitch distribution distance to 0.098 ± 0.019, increases the modal tone ratio to 0.934 ± 0.024, and achieves the structural length validity of 0.946 ± 0.025. Expert evaluation scores become 4.47 ± 0.34 in stylistic consistency and 4.42 ± 0.37 in playability, while novelty score remains 4.16 ± 0.39. These findings show that the combination of symbolic control and conditioning with Guzheng technique brings an improvement of structural validity, stylistic appropriateness, and usability without reduction of novelty.

Keywords: Generative Artificial Intelligence, Guzheng Composition, Symbolic Music Generation, Transformer, Controllable Generation

1. Introduction

Modal balance, phrase coherence, ornamented pitch movement, and technique-dependent articulation are crucial factors in Guzheng composition, thus making its transition from general symbolic music generation quite challenging. However, recent studies have shown that deep learning methods are capable of modeling musical events and long-term dependencies in musical compositions, while the problem of representation construction, coherence, and evaluation of models is of primary importance [1]. Symbolic generation tools are used for creating reproducible pipelines of music representation, transformation, and evaluation in order to increase the comparability of generation processes [2]. Furthermore, theme-conditioned Transformer research proves that explicit musical conditions can direct the generation process in a much more reliable way than unconstrained continuation of music [3]. It is vital for Guzheng composition, as such features as slides, bends, vibrato, tremolo, and pentatonic scale organization create the style of the

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38423

piece itself. A controllable pipeline is designed using REMI encoding and Transformer decoder, with the mode, tempo, seed themes, and techniques as conditions.

2. Literature review

2.1. Symbolic music generation

The concept of symbolic music generation entails the representation of the music scores in discrete event sequences that are learnable, manipulable, and objectively assessable. Event-based models provide the possibility of explicit coding of features such as pitch, duration, position, tempo, and structures, thus providing the ground for sequence generation [4]. The recent study comparing Transformers to recurrent networks proved the effectiveness of the former approach for long-distance modeling and the latter one for short-distance melodic flow preservation [5]. However, for Guzheng composition, it becomes a very important issue since modal structures have to co-exist with ornamentations and phrasal continuity. Thus, a good representation model should be able to maintain the temporal structures and at the same time introduce specific events for the particular instrument.

2.2. Guzheng computational composition

Research into Computational Guzheng is moving beyond just sequence generation towards the creation of instrument-specific music. Han et al. prove that the data for the Guzheng and recurrent learning can model its transitional patterns, modality, and rhythm [6]. Zhao et al. further develop this line of research by creating diffusion-based cross-modal generation which can incorporate traditional characteristics of the Guzheng within the generative model. Recent REMI-based work using the TransformerXL framework generates music for the Guzheng by modeling rhythm, style, and contextual information [7]. These pieces of work have proved the technical viability but have left control distributed among various representations and goals. This means that the compositional approach needs to include modes, tempos, melodic seeds, and playing technique in one and the same symbolic sequence.

2.3. Controllable AI-assisted music creation

Controllable music generation considers musical parameters as concrete conditions that direct the decoding process and influence the output behavior. The recent studies that have incorporated chord conditions into Transformer models for music generation suggest that the high-level musical knowledge can be used along with autoregressive generation to enhance controllability and structure coherence [8]. In case of Guzheng composition, controllable generation will need to have a more specialized design based on mode, tempo, seed themes, and typical performing techniques. Mode and tempo are responsible for global structure, seed themes are associated with local thematic elements, and technique tokens limit idiomaticity. The interactions between them might improve the stylistic coherence but might influence novelty and repetition as well. So, controllability cannot be assessed by adherence alone [9].

DOI: 10.54254/2755-2721/2026.38423

3. Experimental methods

3.1. Dataset construction

Solo Guzheng songs' licensed or publicly accessible music sheets are gathered using MusicXML and MIDI formats, with their processing done using Python, music21, and pretty_midi. The pieces of music are standardized in terms of meter, tempo, note length, and event timing prior to extraction of the phrases. Duplicated notes, abnormal durations, invalid overlapping, and metrical inconsistencies are removed. Phrases are extracted from each song such that they have lengths between four and eight measures with preservation of any cadence and section points within the songs. Each of the phrases is annotated with the modes, tempo, and Guzheng techniques used in the phrase, such as slides, pitch bends, vibrato, and tremolo. The data augmentation process involves consistent mode-based transposition, slight rhythmic variation, and re-segmentation of phrases without changing their identity. Clean phrases are encoded as REMI sequences including Bar, Position, Pitch, Duration, Tempo, Mode, and Technique tokens. The data split involves partitioning the phrases on the level of compositions, ensuring that fragments of the same song do not appear in different subsets.

3.2. Conditional generation model

REMImotif events and a six-layer Transformer decoder with eight attention heads, having a 512 dimensional hidden vector representation is used for Conditional Generator. The attributes like Mode, Tempo, Two-Measure Seed Motif, and Guzheng techniques tokens are prefixed to the Phrase Events so that self-attention can understand their interaction with the attributes like Bar, Position, Pitch, and Duration. Multi-level attention studies show the importance of such modeling [10]. The conditional sequence probability is defined as shown in Equation (1).

$$p_\theta\left(x_{1:T}|c\right)=\prod_{t=1}^{T} p_\theta\left(x_t|x_{<t},c\right) \tag{1}$$

Here, $x_{1:T}$ denotes the generated event sequence, c is the condition set, and θ represents model parameters. Training minimizes conditional cross-entropy, as shown in Equation (2).

$$L_{CE}(\theta)=-\frac{1}{T}\sum_{t=1}^{T}\log p_\theta\left(x_t|x_{<t},c\right) \tag{2}$$

Here, L_{CE} is the mean token-generation loss and T is sequence length. Model is developed using PyTorch framework and trained using AdamW optimization with early stopping based on validation data. At inference time, two-measure seed motif starts decoding, followed by top-p sampling which generates multiple four to eight measure phrases. Conditions such as mode, tempo, and technique stay same for all candidates generated so that variations occur due to sampling of tokens only.

3.3. Experimental evaluation

Three models are analyzed: REMI standard baseline, a mode-tempo-conditional model, and a comprehensive model taking into account Guzheng technique tokens in addition to the previously

DOI: 10.54254/2755-2721/2026.38423

mentioned parameters. Objective analysis comprises pitch distribution, rhythm similarity, modal tone proportion, repetition rate, novelty of n-grams, and length validity. Human analysis is carried out following the recent trends of computational music modeling with human evaluation [11]. Modal conformity is calculated as shown in Equation (3).

$$MTR=\frac{1}{N}\sum_{i=1}^{N} I\left(p_i \in M\right) \tag{3}$$

Here, N is the number of notes, p_i is note i, MM is the target modal pitch-class set, and I equals one when p_i belongs to M. Configuration effects are tested with one-way ANOVA, as shown in Equation (4).

$$F=\frac{\frac{\sum_{g=1}^{K} n_g\left(\bar{y}_g-\bar{y}\right)^2}{K-1}}{\frac{\sum_{g=1}^{K}\sum_{i=1}^{n_g}\left(y_{gi}-\bar{y}_g\right)^2}{N-K}} \tag{4}$$

Here, K is the number of model configurations, n_g is group size, $\bar{y}_g$ is the group mean, $\bar{y}$ is the overall mean, and N is the total number of observations. Each configuration uses identical seed motifs and decoding settings. Guzheng performers or composers rate coherence, stylistic consistency, novelty, and playability on five-point scales. Means, standard deviations, significance tests, post-hoc comparisons, and correlations with objective indicators are then calculated.

4. Results

4.1. Objective generation performance

There were obvious distinctions between the three models regarding structural and stylistic control. It can be seen from Table 1 that the completed model had the minimum pitch-distribution distance of 0.098 ± 0.019 compared to 0.129 ± 0.024 for mode-and-tempo configuration and 0.184 ± 0.031 for the standard REMI configuration. Rhythmic similarity rose from 0.721 ± 0.052 to 0.794 ± 0.043 and 0.836 ± 0.038 correspondingly. The modal tone ratio of the completed model was 0.934 ± 0.024, and at the same time, the structural-length validity was 0.946 ± 0.025. Meanwhile, repetition rate dropped to 0.249 ± 0.036 from 0.286 ± 0.045. On the contrary, n-gram novelty rose to 0.718 ± 0.042 from 0.681 ± 0.049. According to one-way ANOVA test, there were significant differences between the three models for all six measures with p values below 0.

Table 1. Objective generation performance across model configurations

Metric	Standard REMI Baseline	Mode + Tempo	Complete + Technique	F	p
Pitch-distribution distance	0.184 ± 0.031	0.129 ± 0.024	0.098 ± 0.019	72.84	<0.001
Rhythmic similarity	0.721 ± 0.052	0.794 ± 0.043	0.836 ± 0.038	39.27	<0.001

DOI: 10.54254/2755-2721/2026.38423

Table 1. (continued)

Modal tone ratio	0.842 ± 0.041	0.917 ± 0.028	0.934 ± 0.024	51.63	<0.001
Repetition rate	0.286 ± 0.045	0.262 ± 0.038	0.249 ± 0.036	6.91	0.002
n-gram novelty	0.681 ± 0.049	0.706 ± 0.044	0.718 ± 0.042	5.47	0.006
Structural-length validity	0.853 ± 0.039	0.914 ± 0.031	0.946 ± 0.025	46.18	<0.001

4.2. Expert evaluation results

Expert ratings further differentiated the musical usability of generated phrases. As shown in Figure 1, the complete model obtained mean scores of 4.39 ± 0.36 for coherence, 4.47 ± 0.34 for stylistic consistency, and 4.42 ± 0.37 for playability, exceeding the standard REMI baseline scores of 3.62 ± 0.48, 3.48 ± 0.51, and 3.56 ± 0.54. The mode-and-tempo model produced intermediate scores of 4.08 ± 0.41, 4.03 ± 0.43, and 3.95 ± 0.45. Novelty remained stable across configurations at 4.05 ± 0.44, 4.11 ± 0.40, and 4.16 ± 0.39, with no significant difference detected ($F = 1.55$, $p = 0.216$). Significant differences occurred for coherence ($F = 18.62$, $p < 0.001$), stylistic consistency ($F = 31.44$, $p < 0.001$), and playability ($F = 24.87$, $p < 0.001$). Modal tone ratio correlated with stylistic consistency ($r = 0.71$, $p < 0.001$), while rhythmic similarity correlated with coherence ($r = 0.64$, $p < 0.001$).

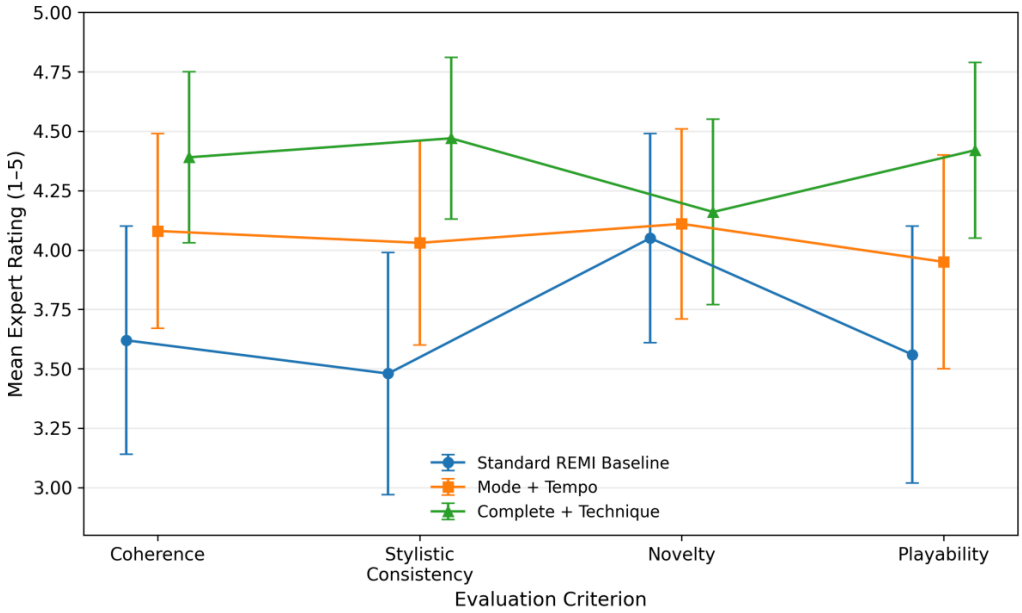


Figure 1. Expert evaluation of generated Guzheng phrases across model configurations

5. Discussion

As shown in the results, controllability seems to apply across several musical levels rather than a single improvement on sequence accuracy. While mode and tempo mainly affect global structure, this is why they are stronger in affecting modal tone ratio and rhythmic similarity, while technique tokens apply nearer to the phrase level, and therefore correspond better to stylistic consistency and performance playability. Stable novelty values are significant in the sense that stronger conditioning

DOI: 10.54254/2755-2721/2026.38423

can reduce the sampling space too much and result in formulaic outputs. In the current case, the presence of low repetition values alongside the novelty indicates that the technique conditioning did not lead the model to use the memorized patterns only. The correlation between modal tone ratio and stylistic consistency is also an evidence of the ability of objective symbolic criteria to reflect some aspects of expert judgments, although this relationship is still incomplete. This is because professional judgment includes gesture continuity, idiomatic fingering, phrase direction, and expressiveness, which are just approximated by discrete tokens.

6. Conclusion

Controlled composition of Guzheng music based on the REMI symbolic representation, Transformer decoder and conditions for mode, tempo, seed phrases and playing techniques was developed. Step-wise comparison of three model variants revealed that the addition of conditional information helps to achieve more stability, consistency with the mode, style and playability of the results. Objective metrics and subjective assessments help to form a more holistic picture of the quality of generated phrases. However, the limitations of this approach are that symbolic events do not capture continuous pitch gestures, changes in timbre and dynamics, as well as nuances of performers. Future work will need to include richer gesture representations, Guzheng corpora of different styles, and audio-based modeling.

References

[1] Ji, S., Yang, X., & Luo, J. (2023). A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Computing Surveys, 56*(1), 1–39.
[2] Hernandez-Olivan, C., & Beltran, J. R. (2023). Musicaiz: A python library for symbolic music generation, analysis and visualization. *SoftwareX, 22*, 101365.
[3] Shih, Y.-J., et al. (2022). Theme transformer: Symbolic music generation with theme-conditioned transformer. *IEEE Transactions on Multimedia, 25*, 3495–3508.
[4] Zhang, N. (2020). Learning adversarial transformer for symbolic music generation. *IEEE Transactions on Neural Networks and Learning Systems, 34*(4), 1754–1763.
[5] Ghoshal, S., et al. (2026). Fusing memory and attention: A study on LSTM, Transformer and hybrid architectures for symbolic music generation. *Expert Systems with Applications*, 131173.
[6] Han, M., et al. (2024). Firefly algorithm-based LSTM model for Guzheng tunes switching with big data analysis. *Heliyon, 10*(12).
[7] Zhao, E., Zheng, J., & Cao, M. (2024). CPTGZ: Generating Chinese Guzheng music from Chinese paintings based on diffusion model. *IEEE Access, 12*, 169247–169262.
[8] Li, Y., Xiong, Y., & Hu, X. (2026). Multi-layer rhythmic modeling and interactive system for guzheng music style generation. *Discover Artificial Intelligence*.
[9] Gao, Z. (2026). A chord-controlled transformer for controllable and coherent music generation. *Complex & Intelligent Systems, 12*(2), 89.
[10] Zhou, L., et al. (2025). MLAT: A multi-level attention transformer capturing multi-level information in compound word encoding for symbolic music generation. *EURASIP Journal on Audio, Speech, and Music Processing, 2025*(1), 17.
[11] Ferreira, P., Limongi, R., & Fávero, L. P. (2023). Generating music with data: Application of deep learning models for symbolic music composition. *Applied Sciences, 13*(7), 4543.