Journal of Innovation and Development
ISSN: 2958-4752 | Vol. 15, No. 3, 2026

HarmonyRFG: A Rule-Guided Spiking Transformer Framework for Real-Time Chord Progression Generation

Zimo Dong

Xianda College of Economics and Humanities, Shanghai International Studies University, 999 Dongtan Avenue, Chongming District, Shanghai, China

Abstract: This paper presents HarmonyRFG, a rule-guided chord generation framework that connects established harmonic practice with deep-learning-based sequence modeling. The framework couples the long-range representation capacity of the Transformer with the event-driven temporal modeling of Spiking Neural Networks, so that chord identity, harmonic function, and duration can be learned as interdependent musical variables rather than isolated symbols. Markov transition constraints and harmony-based scoring are further introduced to regulate local chord movement while retaining generative flexibility. The proposed approach therefore supports chord sequences that are musically coherent, responsive to real-time control, and suitable for creative contexts such as interactive installations, ambient composition, human-computer co-creation, and therapeutic sound environments.

Keywords: Music generation; chord progression; Spiking Neural Network; Transformer; Markov chain; generative art; interactive music.

1. Introduction

Music theory offers a structured vocabulary for organizing pitch relations, harmonic direction, and stylistic expectation. From common-practice harmony to the chord syntax of popular music, these principles help composers produce progressions that are intelligible to listeners. Many recent data-driven systems, however, model music mainly as sequences of discrete tokens. Without explicit harmonic constraints, such systems may produce progressions that are locally plausible but tonally unstable or structurally incoherent. Conversely, purely rule-based systems can enforce harmonic correctness, yet they often lack the adaptive and exploratory capacity required in open-ended creative environments.

This study introduces HarmonyRFG (Harmony Rule-based Forward Generation), a framework developed for real-time chord generation and ambient music production. The model is designed to reduce inference burden while maintaining musical continuity, which makes it appropriate for new media installations and live interaction. HarmonyRFG combines the temporal sensitivity of Spiking Neural Networks (SNNs) with the global contextual modeling of the Transformer, and it uses Markov-chain information together with harmony rules to guide chord continuity and formal organization. In contrast to MusicRL [1], where preference alignment may require human intervention, and DeepBach [2], whose stylistic range is deliberately narrow, HarmonyRFG exposes adjustable controls for emotion, mode, rhythmic density, and harmonic strictness. These controls allow artists to modify musical behavior in performance instead of treating the model as a fixed generator.

2. Background

Deep learning has substantially advanced computational music generation, but generated music is still frequently criticized for limited musical depth and weak affective nuance. A central reason is that music unfolds through time: its emotional and formal meaning depends not only on which events occur, but also on how they are prolonged, delayed, repeated, and resolved. Earlier neural approaches generally relied on sequence models. Mozer [3] explored melody generation with RNNs by incorporating pitch, duration, harmonic context, and psychoacoustic factors, providing an early example of combining musical knowledge with neural modeling. Eck and Schmidhuber [4] later used LSTM networks to generate and improvise blues-style progressions, while Boulanger-Lewandowski et al. [5] integrated RNNs with Restricted Boltzmann Machines for piano-roll-based polyphonic modeling. These models were important, but RNN and LSTM architectures often struggle with long-range musical form; their outputs can become repetitive, static, and insufficiently varied for higher-level artistic use. More recent self-attention models, including Pop Music Transformer [6], improve the capture of long-term structure, yet they still have difficulty producing rhythmically vivid and emotionally responsive music in real-time interactive settings.

DeepBach, by comparison, demonstrates the strength of rule-constrained neural generation. Its use of constraint satisfaction allows the system to produce four-part chorales that follow Bach-style conventions. This rigor is valuable, but it also ties the method to a specific idiom and limits stylistic extension. MusicRL applies reinforcement learning and reward design to increase diversity and creativity, but its workflow may involve considerable human steering and is not naturally suited to immediate interaction in installation or performance environments.

Against these limitations, HarmonyRFG is proposed as a framework for producing chord sequences that are smooth, theoretically grounded, and controllable in real time. The goal is not simply to improve prediction accuracy, but to construct a model that supports style switching, emotional adjustment, and live parameter modulation. Such a system is especially relevant to ambient music, improvisational accompaniment, interactive installation, music therapy, and other scenarios in

which the generator must respond quickly while still preserving musical coherence.

HarmonyRFG starts from the compositional premise that the rhythm of harmonic change is as significant as the chord labels themselves. We therefore combine Transformer and SNN components [7] so that temporal dynamics can be represented directly. By training the model on both chord sequences and duration sequences, rhythmic behavior is learned together with harmonic movement. In this setting, chords are no longer treated merely as isolated symbolic categories; they become events within a temporally extended musical process. Markov transition probabilities and chord-scoring rules are then added as constraints, allowing the output to remain compatible with music theory while preserving flexibility and artistic variation.

3. Methodlogy of HarmonyRFG

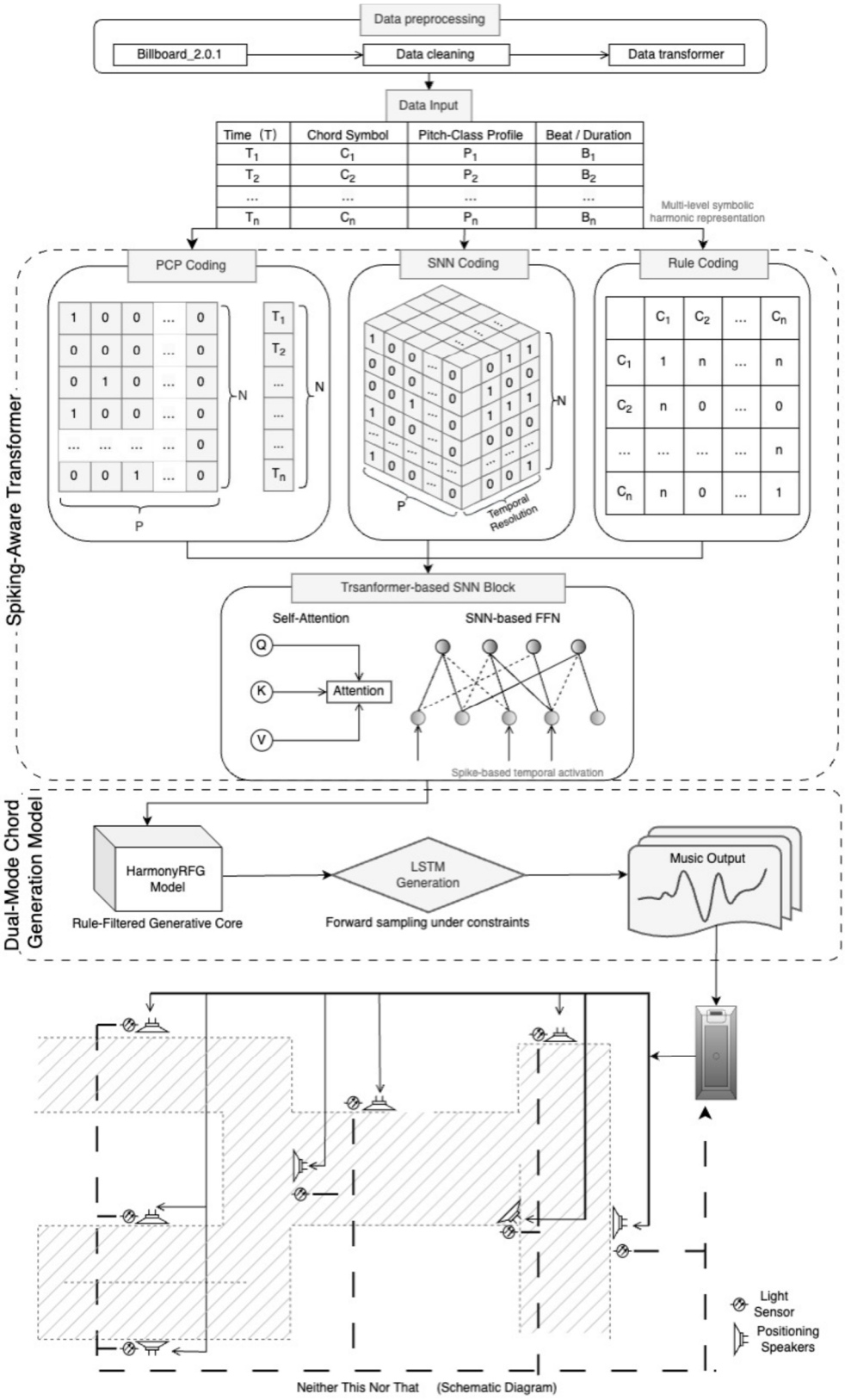

Figure 1. Operating Mechanism of HarmonyRFG

The proposed HarmonyRFG framework contains two main modules (Fig.1):

<1> Spiking-Aware Transformer (SAT)

<2> Dual-Mode Chord Generation Model (DMCG)

The model is trained with the billboard_2.0.1 dataset from The McGill Billboard Project. Chord annotations are encoded according to the proposed syntax for text-based chord representation [8,9].

The first module, the Spiking-Aware Transformer (SAT), modifies the standard Transformer architecture by replacing the Feed-Forward Network (FFN) component with an SNN-based block. In this design, the Transformer captures broad harmonic dependencies across the sequence, while the SNN block represents temporal dynamics associated with chord duration and transition timing.

The second module, the Dual-Mode Chord Generation Model (DMCG), uses a lightweight LSTM decoder for forward chord generation. Transition probabilities supplied by the SNN guide the decoder so that inference can proceed by constrained forward sampling rather than repeated backtracking. At the same time, spike density associated with different durations regulates the rhythmic spacing of the generated progression. The LSTM therefore selects both the next chord and its corresponding rhythmic value under low computational cost, improving the real-time usability of the system.

In a conventional Transformer, the FFN or MLP layer mainly transforms the current hidden representation. After this transformation, explicit information about previous and future chord duration is not retained in a temporal form. As a result, the model may represent chord identity but cannot adequately describe how long a chord should continue or how its duration relates to surrounding events (Fig.2a, Fig.2b).

Musical perception is continuous rather than a simple succession of isolated notes. Gestalt theory suggests that artistic perception tends toward wholeness, balance, and relational organization [10]. For this reason, chord timing must be modeled together with harmonic content. HarmonyRFG uses SNNs inside the MLP position to capture duration, transition interval, and rhythmic activation. The SNN mechanism is inspired by neural dynamics: input accumulates as membrane potential, and a spike is generated only when the potential reaches a threshold. This means that a chord can remain active for a longer interval before the next event is triggered (Fig.2c). In the spike representation, sparse spikes indicate sustained chords, while denser spikes indicate faster harmonic movement (Fig.2d). Thus, the model can learn whether a chord should be prolonged, such as in the interval around coordinates 2-6, or quickly connected to the next chord, as around coordinates 7-8.

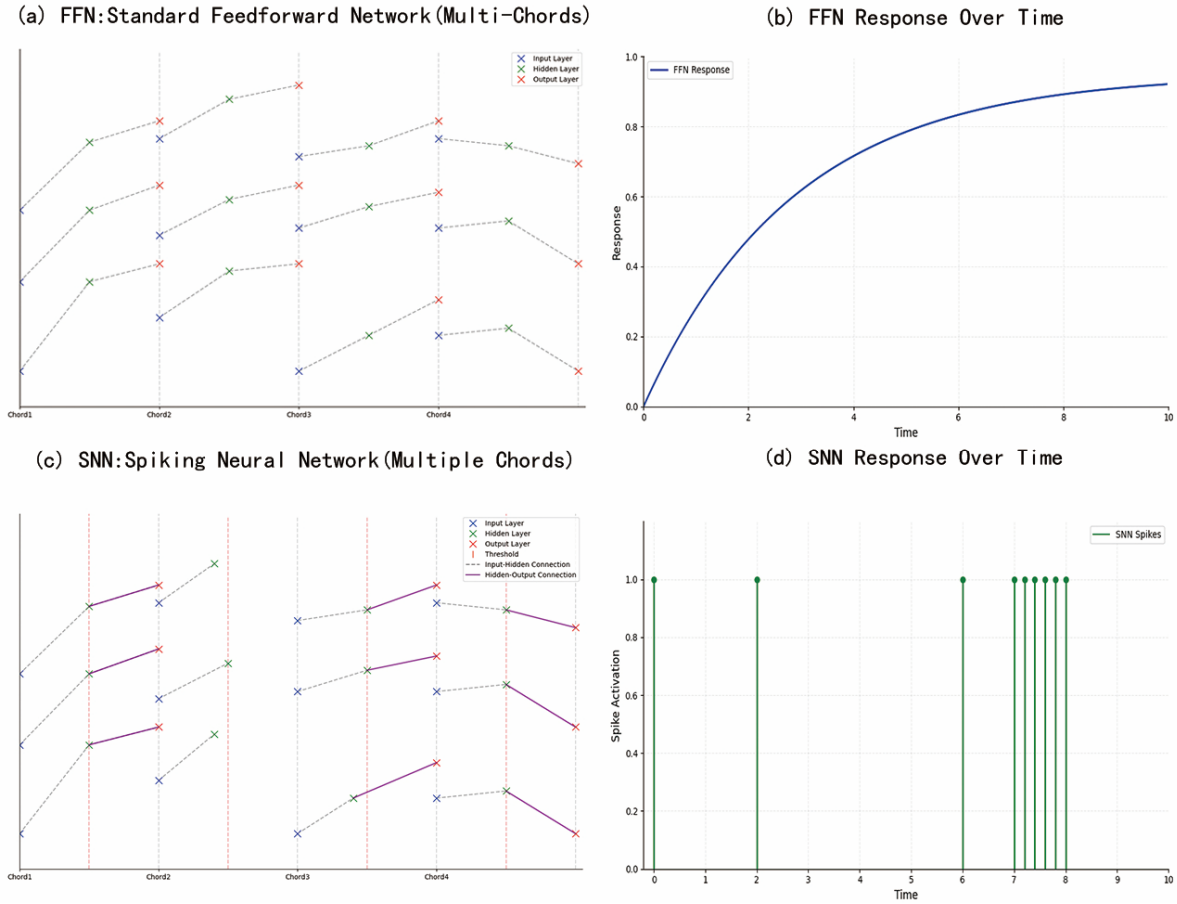

Figure 2. FFN and SNN Training Method Comparison Chart

The Transformer component complements this temporal mechanism through self-attention. With the Query-Key-Value structure, each time step can attend to chord information from the wider sequence, allowing the model to connect local events with larger harmonic patterns. This makes it possible to learn large-scale progressions, tonal return, thematic repetition, and variation. The hybrid structure therefore reduces the fragmentation and abrupt harmonic jumps that often appear in purely local sequence models.

Chord symbols are converted into fixed-dimensional Pitch Class Profile (PCP) vectors, which describe the distribution of pitch classes in each chord. These vectors are organized as

matrices so that similarities between chords can be measured and used to learn transition relations. Duration values are stored in a parallel matrix, ensuring that generated outputs include rhythmic change rather than chord labels alone. The combined harmonic and temporal information is then transformed into spike-based signals that can be processed by the SNN component.

To balance novelty with harmonic plausibility, Markov transition probabilities and rule-based chord scores are incorporated into the attention mechanism. The Markov component discourages unlikely jumps and stabilizes local continuity, whereas harmony rules introduce functional logic and coloristic variation. Relative frequencies of chord classes in the dataset are also considered, and different rule weights are assigned so that the model can distinguish functional progressions from color-oriented transitions. In this way, the system increases chord-color diversity while maintaining theoretical consistency.

4. Results and Analysis

The performance of HarmonyRFG is evaluated from two perspectives: computational behavior and artistic perception. The technical evaluation focuses on training loss, convergence behavior, and relative computational cost. The musical evaluation considers whether the generated progressions are coherent, expressive, and perceptually convincing.

For training behavior and energy use, we compare the loss curves of the SNN-based structure and the conventional FFN structure (Fig.3a, Fig.3b). The horizontal axis denotes epochs, and the vertical axis denotes loss. The FFN curve declines more slowly and stabilizes at a higher loss value, whereas the SNN curve reaches convergence earlier and with a lower final discrepancy. This suggests that the SNN block learns temporal chord dynamics more efficiently, saving approximately 50 epochs of training cost. Since SNN computation is event-driven and occurs primarily when spikes are produced, it avoids the full-step computation required by FFNs. In this experiment, the SNN-based design reduces model complexity by about 40%, supporting its use in low-power or embedded real-time generation contexts.

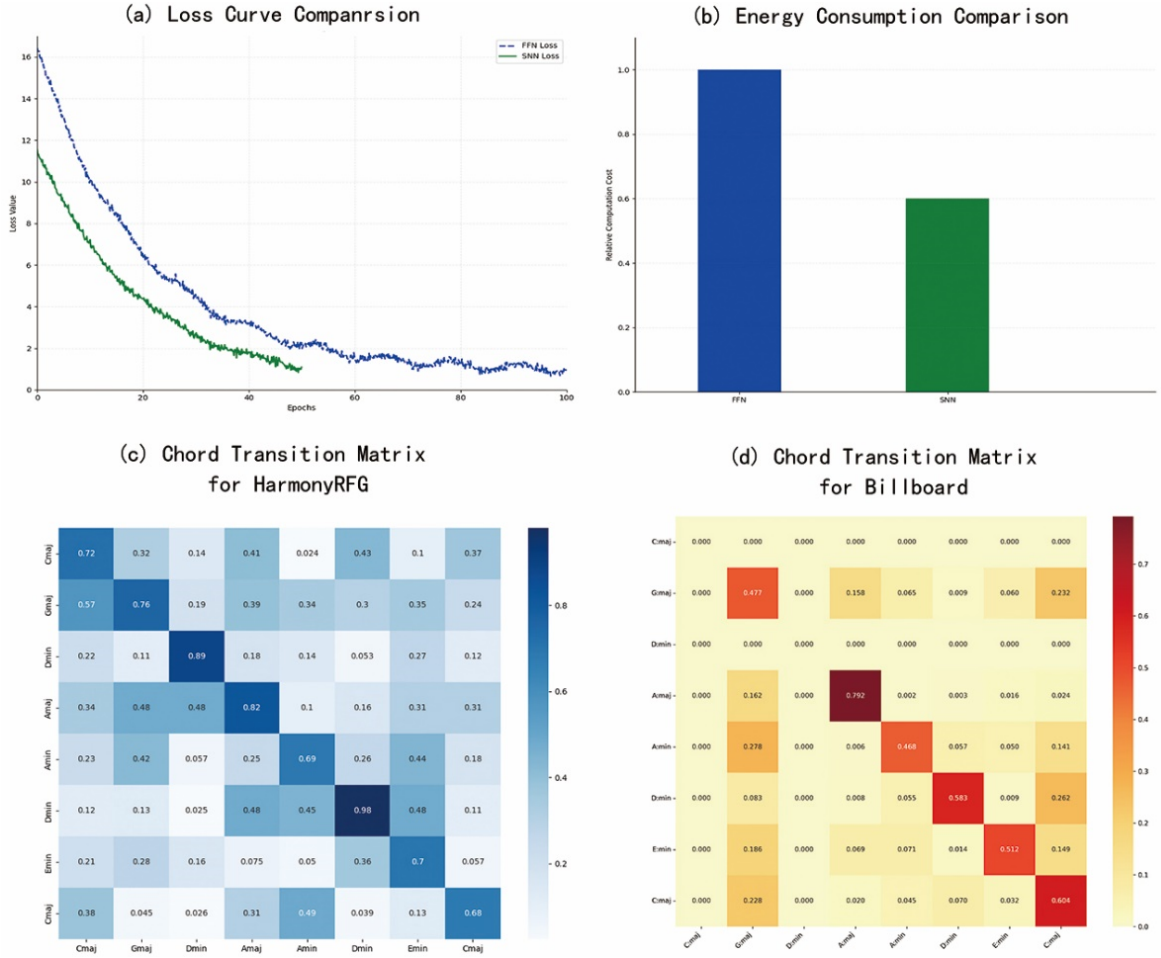

Figure 3. Model Training Results and Evaluation

After training, a randomly generated HarmonyRFG excerpt was visualized as a chord-transition probability heatmap (Fig.3c). The axes represent chord categories, and darker diagonal values indicate a strong tendency for certain chords to persist. For example, the value Dmin -> Dmin = 0.98 shows that repeated or sustained D minor events are highly probable. This behavior corresponds to the role of chord stability in music and indicates that the output maintains continuity. By contrast, uncommon leaps such as Cmaj -> Emin = 0.13 appear with lower probability, suggesting that the model favors smoother harmonic motion over abrupt change.

The generated transition matrix was then compared with the transition probabilities in the original dataset (Fig.3d). The blue matrix includes chord persistence more clearly, while the yellow matrix shows much weaker diagonal emphasis. This difference indicates that the trained model has learned the

possibility of chord sustain as a structural feature. In off-diagonal areas, some transitions remain similar to the dataset, such as Gmaj -> Cmaj, whereas others differ more substantially, including Cmaj -> Gmaj, Amaj -> Amin, Dmin -> Cmaj, and Emin -> Gmaj. These differences show that the model does not merely copy the empirical distribution; instead, it generalizes transition patterns and emphasizes functional relations that are consistent with traditional harmony.

Table 1 compares the chord-duration distributions of the generated data and the McGill-Billboard data. The two datasets show close average durations, which indicates that HarmonyRFG produces chord lengths broadly aligned with real musical material. The generated average duration is slightly longer, suggesting a more stable and slower-changing harmonic surface. The lower standard deviation and higher minimum value further indicate that the rule-constrained model suppresses extremely short chord events and tends to favor more stable duration patterns.

Table 1. The Chord Duration Distribution of HarmonyRFG Generated Data and Billboard Data

	HarmonyRFG data	McGill-Billboard data
Average duration	1.590 seconds	1.542 seconds
Median duration	1.718 seconds	1.531 seconds
Standard deviation	1.052 seconds	1.273 seconds
Minimum value	0.073 seconds	0.023 seconds

In creative practice, two parameters are especially important for controlling the auditory result: membrane-potential threshold and Markov-chain weight. The membrane-potential threshold determines how easily a neuron fires and initiates a new chord. A lower threshold produces faster response and is useful for interactive installations or live improvisation, while a higher threshold creates slower and steadier harmonic change suitable for background atmospheres. The Markov-chain weight controls the strictness of theoretical constraint. A higher value strengthens fluency and harmonic regularity, producing a stable and consonant sound world (Fig.4a). A lower value loosens the constraint, allowing more open, uncertain, and experimental chord movement (Fig.4b).

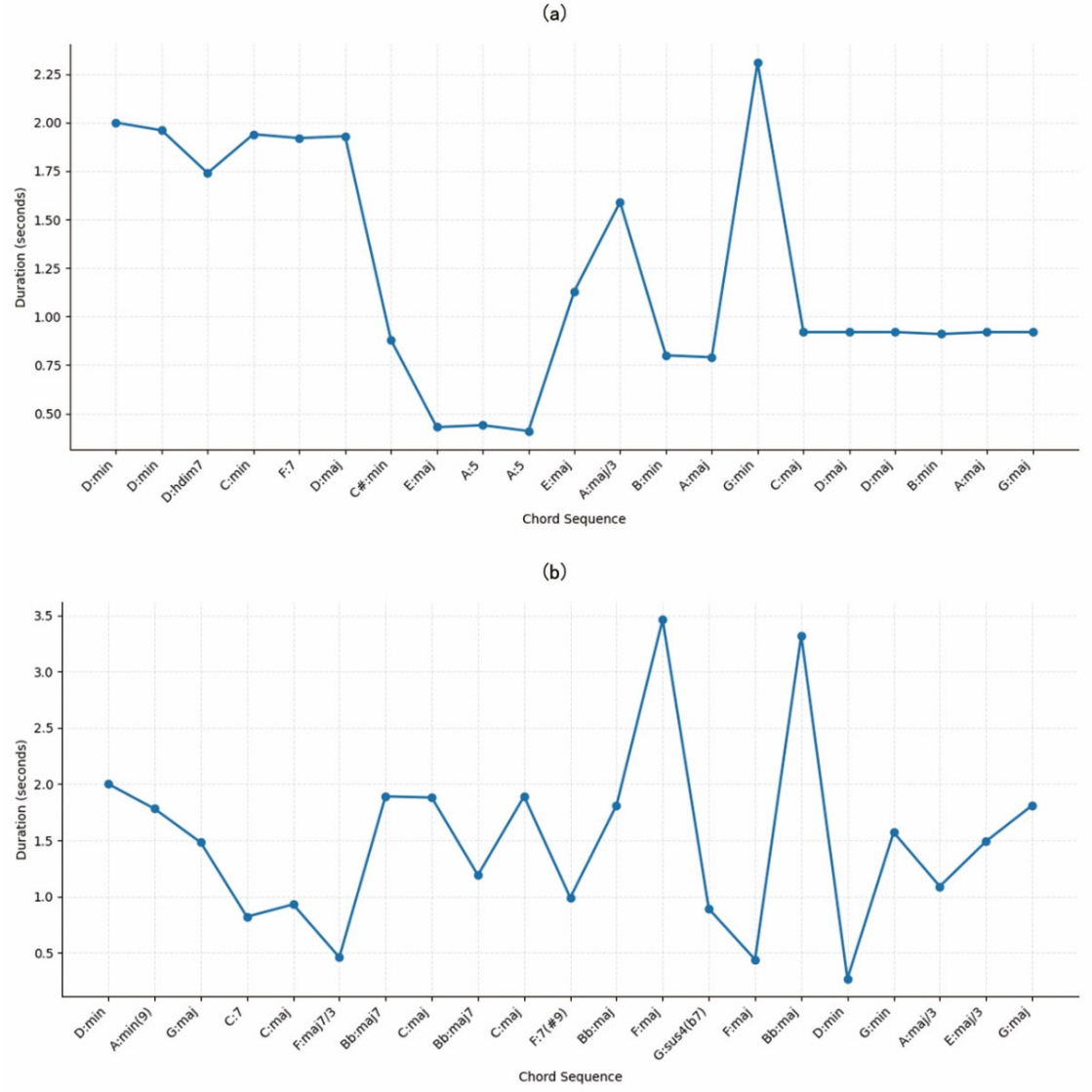

Figure 4. Comparison of Chord Durations Generated with Different Weights

A subjective listening test was also conducted to examine perceived musical quality. Twenty listeners aged 17-30 evaluated eight real-time generated clips under different parameter combinations. They scored each clip on a five-point scale for musical fluency, emotional expression, and innovativeness. Clips generated with higher weights were generally described as stable, harmonious, gradual, and relaxed. Clips generated with lower weights were perceived as freer, more rhythmically varied, and more experimental. Overall, the highest scores appeared under a medium-low threshold and medium Markov-chain weight, especially for emotional expression and artistic novelty.

Taken together, the technical tests and listening evaluations indicate that HarmonyRFG can generate chord progressions that are efficient, fast, and compatible with harmonic theory. Relative to FFN-based models, computational energy consumption is reduced by about 60%, and real-time generation speed increases by about 40%, making the framework practical for interactive artistic production.

In application, HarmonyRFG draws from the logic of parametric design in visual art and architecture. The system treats musical style and structure as adjustable processes rather than fixed outputs. McCormack and Dorin argue that parametric design and algorithmic composition both rely on controllable parameters, rules, and generative procedures [11], while Galanter emphasizes systematicity, autonomy, algorithmic operation, and process in generative art [12]. HarmonyRFG translates these ideas into musical generation by combining adjustable model parameters, SNN dynamics, and Markov-rule constraints. The result is a system with both parametric controllability and generative openness, allowing artists to explore style and affect through real-time interaction.

The practical use of HarmonyRFG is illustrated through an immersive art-healing case study. In this project, environmental variables drive changes in the generated music, demonstrating both parametric control and system interactivity. The cooperation of sound, space, and light further extends the work from a music-generation model into an immersive generative-art environment.

5. Case Study

The case study, Neither This Nor That, was presented at the Zhujiajiao Ultra-Immersive Art Season in Shanghai, China. The exhibition investigated the relationship between real-time sound and immersive spatial experience (Fig.5a). In this work, HarmonyRFG functions as the musical core of a healing-oriented environment. Instead of producing a fixed composition, it follows a process-based logic associated with generative music [13], using harmonic structure, rhythm, and sound flow to support relaxation and emotional regulation. The work aims to make the generated sound feel both familiar and unfamiliar: the model is trained on popular-music data, so it retains recognizable harmonic tendencies, while parameter adjustment reshapes balance, timbre flow, and continuity to create a smoother and more meditative sound environment [14].

HarmonyRFG operates as one part of a larger generative art system rather than as an isolated music engine. Light sensors inside the installation collect environmental information, including light intensity and temporal change. These signals are digitized and sent to the model as real-time control inputs. The resulting music interacts with the laser PVC structure, spatial layout, and eight positioning speakers (Fig.5b). According to the distribution of light in the space, HarmonyRFG assigns musical content to different speaker positions, forming a layered soundscape. When light is reflected and refracted by the laser PVC, sensor values change and the generated music shifts subtly. For instance, when a beam reaches a sensor area, the local speaker volume may increase and the rhythmic pattern may slightly adjust, as though light were conducting the motion of sound. The flicker of light and the rhythm of music are therefore linked into a combined visual-auditory experience. Stronger beams are associated with warmer, fuller harmonies, whereas weaker beams correspond to lighter and more restrained melodic textures [15]. In this way, the work presents perception as an embodied interaction among audience, space, sound, and environment rather than a passive reception of separate visual and auditory materials [16].

This environmental, real-time mode of generation allows Neither This Nor That to construct a multisensory field in which sound, light, space, and time continuously affect one another. Audiences are surrounded by spatialized sound while also responding to shifting reflections and shadows. As environmental conditions change, the music also changes, producing a calm and immersive state shaped by the interweaving of space, light, and sound [17].

During the exhibition, many visitors remained in the installation for extended periods and adjusted their bodily behavior, such as closing their eyes, listening quietly, or slowing their movement. Informal comments suggested that the fluidity and balance of the music helped some participants feel relaxed and focused, while the reverberant propagation of sound in the space intensified the affective experience [18] (Fig.6). The slow transformation of light and the softened musical texture together created a safe and contemplative atmosphere. More broadly, the parametric controllability of the system suggests a potential direction for art healing: because HarmonyRFG does not rely on a fixed score, the music can respond continuously to environmental parameters and may better resonate with different emotional states [19].

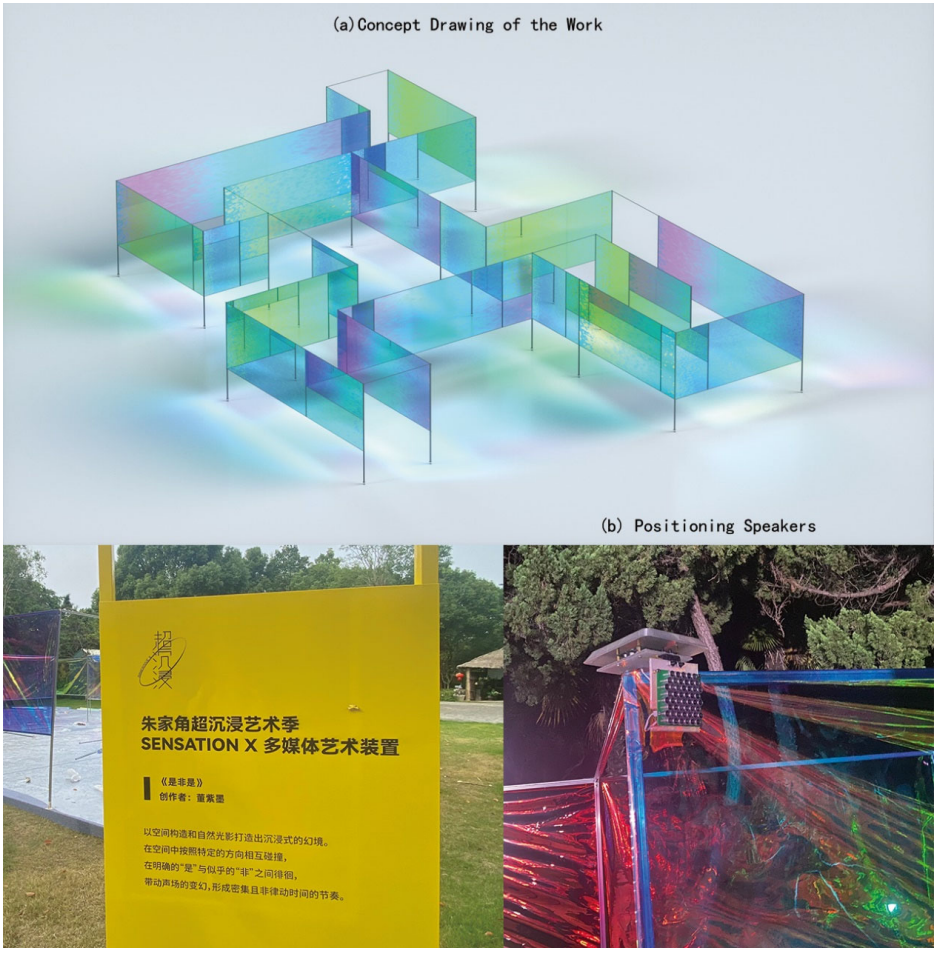

Figure 5. Zimo Dong, Neither This Nor That, Laser PVC, Aluminum and Positioning Speakers, 13mX9mX2m,2023

Figure 6. Zimo Dong, *Neither This Nor That*, Laser PVC, Aluminum and Positioning Speakers,13mX9mX2m,2023, On-site Audience and Artists Viewing the Works

6. Discussion and Summary

Technically, HarmonyRFG improves chord-sequence coherence, harmonic logic, and affective expression by integrating Transformer-based global modeling with SNN-based temporal perception. Compared with RNN-based systems or strictly rule-bound generators, it offers stronger real-time controllability, smoother harmonic movement, and more flexible mood shaping. The framework also demonstrates how data-driven learning can be combined with explicit harmony rules to support music healing, interactive art, and generative sound systems.

At the same time, HarmonyRFG should not be understood only as an efficient algorithm. It functions as a dynamic collaborator within the artistic process. Because the model allows real-time adjustment of harmonic structure, rhythmic density, and affective color, artists can respond to site conditions, audience behavior, or their own improvisational decisions during performance. This shifts evaluation away from technical indicators alone. The artistic value of the model also depends on whether it can move listeners, support emotional expression, and create new auditory experience. Future work will further develop the model's plasticity in artistic contexts and test its use across a wider range of generative art practices.

References

[1] Cideron,Geoffrey et al. "MusicRL: Aligning Music Generation to Human Preferences." ArXiv abs/2402.04229 (2024).

[2] Gaetan Hadjeres, Francois Pachet, Frank Nielsen, "DeepBach: a Steerable Model for Bach Chorales Generation," ICML'17: Proceedings of the 34th International Conference on Machine Learning, Volume 70, pp. 1362—1371 (2017).

[3] Mozer, M. C., "Neural network music composition by prediction: Exploring the benefits of psychophysical constraints and multiscale processing," Connection Sci., No. 6, pp. 247--280 (1994).

[4] Eck, D., J. Schmidhuber. "Learning the long-term structure of the blues," Proc.2002 Internat. Conf. Artificial Neural Networks ICANN, 284–289 (2002).

[5] N Boulanger-Lewandowski, Y Bengio, P Vincent,"Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription", ICML'12: Proceedings of the 29th International Coference on International Conference on Machine Learning, pp. 1881–1888(2012)

[6] YS Huang, YH Yang, "Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions," MM '20: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1180–1188(2020).

[7] Man Yao, Xuerui Qiu, Tianxiang Hu, Jiakui Hu, Yuhong Chou, Keyu Tian, "Scaling Spike-Driven Transformer With Efficient Spike Firing Approximation Training", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 47, Issue 4, pp.2973–2990(2025).

[8] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga, "An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis", Proceedings of the 12th International Society for Music Information Retrieval Conference, ed., pp. 633–38(2011).

[9] Christopher A. Harte, Mark B. Sandler, Samer A. Abdallah, and Emilia Gómez, "Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations", Proceedings of the 6th International Conference on Music Information Retrieval, ed., pp. 66–71(2005).

[10] Peter Simon Sapaty. "Towards Wholeness and Integrity of Distributed Dynamic Systems." Journal of Computer Science & Systems Biology, 9：3(2016).

[11] McCormack Jon, and Alan Dorin. "Artistic practice as research in generative art." Leonardo 38, No. 2. pp.101-109 (2005).

[12] Galanter Philip. "What is Generative Art? Complexity theory as a context for art theory." Generative art 1, No. 1 (2003).

[13] Miranda Eduardo Reck. "On computational models of musical creativity." Contemporary Music Review 22, No. 4. pp. 25-47 (2003).

[14] Stefan Koelsch, "Towards a neural basis of music-evoked emotions，" Trends in Cognitive Sciences Volume 14, Issue 3, pp. 131-137(2010).

[15] Bérigny C.D., et al. "EEG and Sonic Platforms to Enhance Mindfulness Meditation." Journal of Arts and Humanities， No.5, pp.1-12(2016).

[16] Barbara Maria Stafford, "From Visual Culture to Sensory Culture." Leonardo, Vol. 35, No. 4, pp. 401–04(2002).

[17] Bro et al.,Musical Breaks, "Live Music in a Hemodialysis Setting--A Qualitative Study on Patient." Nurse, and Musician Perspectives. Healthcare, 10.(2022).

[18] Barry Blesser,Linda-Ruth Salter, Spaces Speak, Are You Listening? (Cambridge: MIT Press, 2009). p 127-163

[19] Wang, et al., "Real-time Emotion-based Music Arrangement with Soft Transition." IEEE Transactions on Affective Computing. (2023)