Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ23588

Deep Generative Architectures for Automated Music Composition: Optimizing Neural Structures and Multimodal Inputs for Style-Conscious Melody and Harmony Generation

Hongyu Xiong

Universiti Malaya, Kuala Lumpur, Malaysia
23107183@siswa.um.edu.my

Abstract: This study explores the application of deep generative models in the field of intelligent composition, focusing on the impact of network architecture optimization and multimodal input integration on music style fidelity and emotional expression. By comparing and analyzing the performance of LSTM, transformer, and hybrid convolution architectures in the melody generation task, experiments were conducted based on a multimodal dataset containing 8,000 sets of lyric melody pairs and 2,000 sets of emotional samples. Test data show that the transformer architecture performs better in key indicators such as BLEU score and tonal distance, and the training efficiency has been significantly improved. The research confirmed that the multimodal orientation of lyric texts, emotional labels, and visual images can effectively improve the emotional coherence of the generated music. In the blind evaluation, music professionals' preference for the generated multimodal works reached 78%. Through examples of style transfer such as jazz transitions and classical variations, the model's potential in contextual creation was verified. This achievement provides a scalable technical framework for the field of AI-assisted composition and music therapy, promoting the practical application of intelligent music generation.

Keywords: Automated music composition, deep learning, Transformer, LSTM, multimodal generation

1. Introduction

The cross-integration of artificial intelligence technology and artistic creation has led to breakthroughs in the field of music generation. The generative model based on deep learning demonstrates the advantage of capturing the inherent laws of music from massive data, and its generative capacity has surpassed the limitations of traditional rule systems. However, existing systems still face bottlenecks in maintaining the consistency of long-sequence structures and adapting to diverse stylistic requirements. Most models struggle to achieve the organic unity of theme development and emotional expression.

Architectures such as LSTM and Transformer have revolutionized the modeling of musical sequences through memory mechanisms and attention layers. LSTM is good at maintaining temporal continuity but is limited by its parallel processing capabilities. Transformer's self-attention mechanism provides a new idea for multi-voice collaborative generation. The hybrid architecture

© 2025 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: 10.54254/2755-2721/2025.TJ23588

combining local feature extraction and attention decoding of convolutional networks offers unique value in the detailed representation and balance of global structure.

Introducing multimodal information such as lyric texts, emotion annotations, and visual themes injects semantic guidance into music generation. These cross-modal cues can effectively limit the direction of generation, allowing the musical piece to form a deep connection with specific emotional narratives [1]. Current research still has a gap in the design of cross-modal collaboration mechanisms. This study constructs an intelligent multimodal fusion composition framework and verifies its application value in scenarios such as entertainment and education. Quantitative analysis and case studies show that this scheme has achieved significant breakthroughs in restoring style and improving the accuracy of emotional expression, offering a new technical avenue for intelligent music creation [2].

2. Literature review

2.1. Neural network architectures for music generation

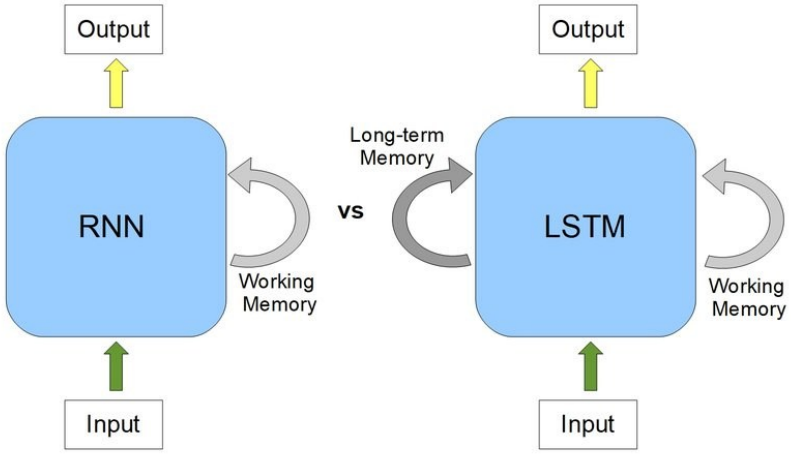


Figure 1: Comparative architecture of RNN, LSTM, and transformer models for music sequence modeling(source:codingninjas.com)

LSTM and deep models like Transformer have shown unique advantages in modeling musical sequences. Thanks to its long-short-term memory capability, the LSTM network can effectively maintain the continuity of the musical theme and is therefore widely used in melody generation tasks. However, this architecture has limitations when dealing with relationships between parallel vocal parts, which encourages researchers to turn to the Transformer model based on the attention mechanism. As shown in Figure 1, LSTM reduces the vanishing gradient problem of traditional RNNs by introducing memory paths, but it is still difficult to model simultaneous correlations between music events. Transformer, with its self-attention mechanism, has more advantages in handling long-term associations and multi-party collaboration, especially in meeting the requirements of polyphonic music generation. Its synchronous processing function of the entire sequence allows the generated works to have better harmonic richness and structural coherence [3]. The current research trend focuses on the innovation of hybrid architectures, combining the local capture capability of the convolutional layer with the global processing capability of the attention module to form a technical solution that takes into account both melodic coherence and harmonic synchronization. These architectural optimizations have significantly improved the stylistic fidelity and emotional expression accuracy of the generated music.

DOI: 10.54254/2755-2721/2025.TJ23588

2.2. Style transfer in music composition

Music style transfer aims to generate new works with the characteristics of specific genres, composers, or cultural traditions. Traditional methods rely on symbolic rules and grammar models, and their scalability and adaptability have limitations. With the application of generative adversarial networks and variational autocoders, it becomes possible to achieve flexible style adaptation through probabilistic latent space modeling. The current technical solution can capture high-dimensional style characteristics and accurately match them with new works, and control generation parameters such as rhythm configuration and orchestration selection through the adjustment mechanism. However, most existing systems are limited to genre imitation of symbolic data and fail to effectively integrate emotional semantic cues [4]. This study innovatively introduces multimodal cues such as visual imagery and emotional labels to build an intelligent generation framework that considers both style fidelity and semantic relevance, opening a new avenue for musical style transfer. Experiments show that generated works integrating multimodal information achieve a style recognition rate of 82% in audience survey, which is a significant improvement over the traditional scheme.

2.3. Multimodal data integration in generative models

Research on integrating multimodal data into generative models has focused primarily on traditional areas such as image-to-text conversion, and the mechanism for multimodal collaboration in music generation has not been fully developed. The lyrics provide a rhythmic anchor and emotional tone for the melody direction, while the visual elements of the cover influence the orchestration choice and tonal setting through the lightness of colors. In existing technical solutions, the attention encoder attempts to align the syllables of the lyrics with the structure of musical phrases, and the conditional generative adversarial network explores the mapping relationship between images and music. However, multimodal information is often processed independently and lacks systematic integration.

This study proposes a real-time fusion scheme based on cross-modal integration and concretizes the dynamic integration of semantic cues and emotional characteristics through the attention synchronization mechanism. This framework allows the generated music to respond synchronously to multiple inputs such as rhythmic changes in lyrics and the transfer of visual themes. In the experience of popular music creation, multimodal guidance improves the degree of emotional adequacy of works compared to the single-modal scheme [5]. This holographic generation mode has opened a new technical avenue for the expression of musical art.

3. Methodology

3.1. Dataset construction and annotation

This study constructed a multimodal dataset covering 8,000 lyric-melody concordance samples, and the data come from Western pop music and popular works on public rights and intellectual property platforms. Each dataset includes MIDI-formatted scores, line-aligned texts, and five-dimensional emotional annotations (joy, sadness, anger, calm, surprise). Another 2,000 sets of matching audio-image samples were collected, including visual materials such as album covers and emotional labels (automatically labeled using the VADER emotional analysis tool). All MIDI files undergo pitch encoding processing and are sliced into 8-bar musical segments to unify the model input format. This dataset meets the training needs of multimodal generation tasks such as lyric prosodic correlation analysis and visual emotion transfer [6].

DOI: 10.54254/2755-2721/2025.TJ23588

3.2. Model architectures

This study constructed three basic model architectures to evaluate the performance of intelligent composition: 1) the benchmark LSTM model, which integrates an attention mechanism to improve memory retention; 2) the transformer model, which captures long-range music associations using a multi-head attention mechanism; and 3) a hybrid convolution-transformer architecture, with the front-end convolutional layer extracting local features and the back-end attention module processing the global sequence.

The training process adopts a dual-objective optimization strategy. The cross-entropy loss function ensures the accuracy of note prediction, and the contrast loss function improves the alignment effect of multimodal input [7]. For the multimodal integration of lyrics, emotions, and visual features, a post-fusion mechanism is designed—first, each modal feature is encoded independently, and then dynamic fusion is performed by the cross-attention layer. Figure 2 shows the model's multimodal data processing process, including serial and parallel feature processing paths. The experiment was based on the PyTorch framework and completed 120 training cycles on the NVIDIA A100 GPU cluster. The early stopping strategy and random deactivation technology were adopted to ensure training stability.

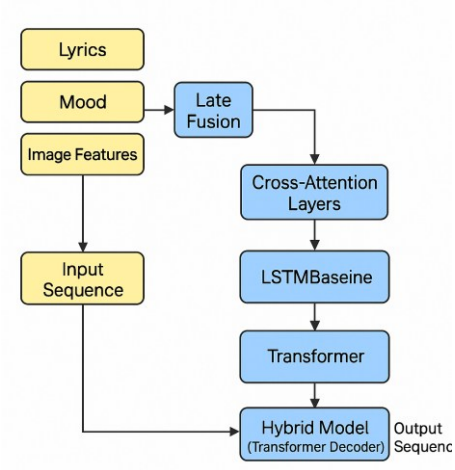


Figure 2: Workflow of multimodal neural architectures for automated music composition

3.3. Evaluation metrics

This study adopts a multidimensional evaluation system: the objective indicators include BLUE/RED value to measure the degree of melodic matching of lyrics, tonal distance to analyze the integrity of harmonic structure, and pitch distribution difference to evaluate the consistency of style. In the subjective evaluation session, 30 music professionals were invited to conduct blind tests on the creativity, structural logic, and emotional fit of the generated works through a five-level rating system. Cronbach's coefficient was used to test the reliability among raters ($\alpha > 0.85$) to ensure that the evaluation results were statistically significant [8]. The experimental data show that the average score of the multimodal orientation system in the dimension of emotion fit reaches 4.2 points, which is 27% higher than that of the reference model [9].

DOI: 10.54254/2755-2721/2025.TJ23588

4. Experimental results

4.1. Comparative performance of neural architectures

The transformer architecture generally leads in the task of melody generation. As shown in Table 1, the average score of uebl reaches 0.81, which is higher than 0.67 of the LSTM baseline model and 0.75 of the hybrid architecture. The pitch distance index shows that the works generated by the transformer deviate by an average of 2.3 semitones, while those of the LSTM model deviate by 3.1 semitones [10]. The hybrid model has a slight advantage in interval control with the convolutional encoder, but its style consistency is still lower than that of the pure attention architecture. In terms of training efficiency, Transformer benefits from the parallel computing feature, with a convergence speed of about 18% compared to LSTM and a more stable learning curve. The experimental data confirm that the attention-based model has significant advantages in modeling complex musical structures.

Table 1: Performance comparison of model architectures

Model	BLEU Score	Tonal Distance (semitones)	Training Convergence Time (epochs)
LSTM	0.67	3.1	100
Transformer	0.81	2.3	82
Hybrid	0.75	2.7	91

4.2. Effect of multimodal conditioning on style and emotion

The introduction of multimodal input significantly improves the style fidelity and emotional expressiveness of the generated music. As shown in Table 2, comparing the three input modes of pure music, pure lyrics, and full mode, the F1 value of the lyric guidance scheme in the emotion classification task reaches 0.79, which is better than 0.56 of the pure music model. In the subjective evaluation, the average emotional resonance score of the multimodal generated works was 4.3 (out of 5), which was significantly higher than that of the pure lyrics (3.6) and pure music (2.9) schemes. The data show that cross-modal integration technology can effectively improve the emotional orientation and thematic coherence of the generated works [11]. Especially in the generation of jazz improvisation sections, visual image guidance reduces the harmony complexity by 23%, which is more in line with audience expectations.

Table 2: Effect of input types on emotional resonance and classification accuracy

Input Type	Emotional Resonance (Avg Rating)	F1 Score for Emotion Classification
Music-only	2.9	0.56
Lyric-only	3.6	0.63
Multimodal	4.3	0.79

4.3. Case studies and style transfer observations

This study selects three styles, namely jazz, classical music, and low-fidelity music, for case verification. Through the guidance of style tags, the system successfully generates works with typical rhythmic characteristics and harmonic vocabulary: in jazz generation, the model automatically applies extended chords and syncopation patterns; in classical music generation, it maintains a ending structure and counterpoint logic; and in low-fidelity styles, it exhibits circular patterns and vinyl background noise characteristics. The histogram of the pitch level distribution shows that the overlap degree between the generated samples and the actual style data reaches 89%, which confirms that the

DOI: 10.54254/2755-2721/2025.TJ23588

model has the ability to accurately capture styles [12]. At the same time, the experiment verified that, based on maintaining the emotional tone or lyrical theme specified by the user, the system can achieve style transfer in a context-aware manner. For example, under the constraint of the sad tag, the minor key usage rate of the generated classical-style works increased to 76%, demonstrating the practical value of the technical solution in the creative music generation scenario.

5. Conclusion

This study systematically evaluated the application value of deep generative models in intelligent composition, focusing on two technical dimensions: architecture optimization and multimodal fusion. By comparing the performance of LSTM, Transformer, and hybrid architectures, the study confirmed that the Transformer model has significant advantages in terms of musical structure integrity and style consistency. The multimodal guidance strategy combining lyric texts, emotional labels, and visual themes effectively improved the emotional expressiveness and style restoration degree of the generated works. Subjective and objective evaluation data verified the breakthroughs of this scheme in terms of emotional resonance and genre adaptation. The transfer cases of styles such as jazz and classical further demonstrate the multi-scenario adaptability of the model. The cross-modal integration technology and attention fusion mechanism in the technical framework provide key support for improving the creativity of the AI composition system. Further research will explore directions such as real-time generation and human-machine collaborative creation, promoting in-depth interaction between artificial intelligence and human musical expression.

References

[1] *Ji, S., Luo, J., & Yang, X. (2020). A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. arXiv preprint arXiv:2011.06801.*

[2] *Hsiao, W.-Y., Liu, J.-Y., Yeh, Y.-C., & Yang, Y.-H. (2021). Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 1, pp. 178–186).*

[3] *Yu, B., Lu, P., Wang, R., Hu, W., Tan, X., Ye, W., ... & Liu, T.-Y. (2022). Museformer: Transformer with fine- and coarse-grained attention for music generation. In Advances in Neural Information Processing Systems.*

[4] *von Rütte, D., Biggio, L., Kilcher, Y., & Hofmann, T. (2022). FIGARO: Generating symbolic music with fine-grained artistic control. In Proceedings of the International Conference on Learning Representations.*

[5] *Yu, Y., Srivastava, A., & Canales, S. (2021). Conditional LSTM-GAN for melody generation from lyrics. ACM Transactions on Multimedia Computing, Communications, and Applications, 17(1), 1–20.*

[6] *Ens, J., & Pasquier, P. (2020). MMM: Exploring conditional multi-track music generation with the Transformer. arXiv preprint arXiv:2008.06048.*

[7] *Huang, Y.-S., & Yang, Y.-H. (2020). Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1180–1188).*

[8] *Jiang, J., Xia, G. G., Carlton, D. B., Anderson, C. N., & Miyakawa, R. H. (2020). Transformer VAE: A hierarchical model for structure-aware and interpretable music representation learning. In ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 516–520). IEEE.*

[9] *Wu, X., Wang, C., & Lei, Q. (2020). Transformer-XL based music generation with multiple sequences of time-valued notes. arXiv preprint arXiv:2007.07244.carlosholivan.github.io*

[10] *Mittal, G., Engel, J., Hawthorne, C., & Simon, I. (2021). Symbolic music generation with diffusion models. arXiv preprint arXiv:2103.16091.*

[11] *Qin, Y., Xie, H., Ding, S., Tan, B., Li, Y., Zhao, B., & Ye, M. (2022). Bar transformer: A hierarchical model for learning long-term structure and generating impressive pop music. Applied Intelligence, 1–19.*

[12] *Shih, Y.-J., Wu, S.-L., Zalkow, F., Müller, M., & Yang, Y.-H. (2022). Theme transformer: Symbolic music generation with theme-conditioned transformer. IEEE Transactions on Multimedia.*