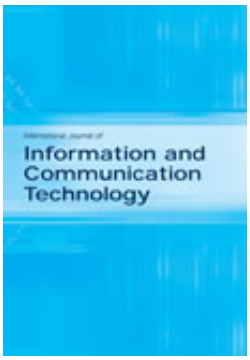
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Residual-enhanced transformer for affective multi-part music generation

Wenqi Li

DOI: 10.1504/IJICT.2025.10072786

Article History:

Received:	16 June 2025
Last revised:	02 July 2025
Accepted:	02 July 2025
Published online:	26 August 2025

Residual-enhanced transformer for affective multi-part music generation

Wenqi Li

Department of Arts,
SouthWest Petroleum University,
Sichuan 610500, China
Email: smpbjs123@163.com

Abstract: Music generation demands modelling intricate multi-dimensional sequences while preserving structural coherence and emotional expressiveness. To address transformer's limitations in detail retention, multi-track efficiency, and affective integration, we propose residual enhanced affective music transformer (REAM) with three key innovations: 1) residual dense blocks establishing inter-layer skip connections to enhance feature reuse and maintain fine-grained musical textures; 2) emotion-aware rotary positional encoding that dynamically modulates note relationships based on target sentiment vectors; 3) lightweight residual modules enabling efficient parallel generation of multi-track compositions. Through systematic ablation studies and perceptual evaluations, REAM demonstrates superior performance in both objective reconstruction metrics and subjective musicality assessments. This framework bridges symbolic precision with affective depth, enabling computationally efficient generation of structurally coherent, emotionally controllable multi-instrument music compositions.

Keywords: multi-part music generation; transformer; residual network; emotional modelling; lightweight architecture.

Reference to this paper should be made as follows: Li, W. (2025) 'Residual-enhanced transformer for affective multi-part music generation', *Int. J. Information and Communication Technology*, Vol. 26, No. 31, pp.88–104.

Biographical notes: Wenqi Li received his Bachelor's degree from Xi'an Conservatory of Music in 2014 and received his Master's degree from Conservatorio Niccolò Piccinni di BARI (Italy) in 2018. Currently, he works in the College of Arts in SouthWest Petroleum University. His research interest is music art and AI music generation.

1 Introduction

Music generation, situated at the intersection of artificial intelligence and art, has emerged as a highly promising yet challenging research frontier (Briot and Pachet, 2020; Kang et al., 2024). Compared to natural language or image generation tasks, music presents unique complexities: it is a structured, multi-dimensional, and emotionally rich temporal signal. In particular, multi-part music generation demands that models learn to simultaneously coordinate melody (Gao and Li, 2025), harmony, rhythm, and their

temporal evolution, all while maintaining overall coherence and conveying a targeted emotional atmosphere (Li et al., 2024a). These requirements place stringent demands on model design in terms of long-range dependency modelling, multi-voice interaction, and emotional expressiveness (Wang et al., 2024a).

Traditional approaches to music generation include rule-based systems (Wang et al., 2024b) and shallow statistical models (Wang et al., 2024c). Rule-based systems offer strong controllability but suffer from poor generalisation due to their reliance on hand-crafted music theory rules (Bhandari and Colton, 2024; Kwiecień et al., 2024; Xin, 2024). Shallow neural models such as RNNs or LSTMs have shown promise in modelling musical sequences, but they struggle with capturing long-term dependencies, inter-voice synchronisation, and dynamic emotional variation (Li, 2024; Li et al., 2024b; Wang et al., 2024d; Zhu et al., 2024). Furthermore, emotional modelling is often limited to using static emotion tags without dynamic integration into the music sequence, resulting in outputs that lack expressive consistency and emotional depth.

Recently, transformer-based models have demonstrated strong capabilities in sequence modelling and have been adopted in music generation due to their self-attention mechanisms, which allow for direct modelling of global dependencies across sequences (Ayres et al., 2024; Liu et al., 2024). While transformers offer improved performance over earlier architectures, they also face significant limitations in this domain. First, their purely stacked architecture lacks mechanisms for cross-layer feature reuse, which leads to the dilution of low-level musical details (Chen et al., 2024) – such as note-level rhythmic variations (Wang et al., 2024d) or ornamentation – at higher semantic levels (Huang, 2025). This weakens structural coherence and expressiveness. Second, their large parameter count and computational complexity make them inefficient for long multi-track sequences, limiting scalability. Third, standard positional encoding schemes fail to capture the nuanced evolution of emotion in music, leading to poor emotional controllability and ambiguous affective expression.

To address these challenges, we propose a novel framework, REAM: residual enhanced transformer for affective multi-part music generation, which incorporates three core innovations targeting architecture optimisation, cross-layer information flow, and emotion-aware sequence modelling:

1 We introduce residual dense blocks into both the encoder and decoder of the transformer. These blocks establish multi-layer skip connections that allow low-level musical features to flow into higher semantic representations, enabling deeper interaction across melody, harmony, and rhythm. This design ensures that fine-grained musical details are retained and integrated, enhancing the structural coherence and fidelity of generated music.

2 We extend rotary position embedding by explicitly incorporating emotion labels into the positional encoding process. This enables the model to dynamically capture the temporal and affective dependencies between notes and emotional states. By conditioning generation on labels such as ‘joyful’ or ‘melancholic’, the model can modulate the acoustic and structural features of multi-part sequences in alignment with the intended affective tone, allowing for fine-grained emotional control.

3 Inspired by advances in image compression, we employ three-layer lightweight residual blocks with nonlinear activations and normalisation layers. These modules reduce parameter overhead while preserving high expressiveness, achieving an effective trade-off between model depth and computational efficiency. This is particularly beneficial for multi-part music generation tasks, where both model scalability and training stability are critical.

2 Relevant technologies

2.1 Residual network

In the field of deep learning, there exists a complex relationship between network depth and model performance (Almukhalfi et al., 2024; Herrmann and Kollmannsberger, 2024). Theoretically, as the number of neural network layers increases, the model can learn more abstract and advanced feature representations, thereby enhancing its ability to model complex data patterns and continuously optimising performance (Zhao et al., 2025). However, in practical applications, when the network layers become excessively deep, thorny problems such as gradient vanishing, gradient explosion, and degradation arise. Gradient vanishing refers to the exponential decay of gradients during backpropagation as the number of network layers increases, making it difficult to update the parameters of layers close to the input layer (Wu et al., 2024). Gradient explosion, on the contrary, means that gradients continuously increase during backpropagation, causing the parameter update magnitude to be too large and preventing the model from converging (Yang et al., 2024). Degradation manifests as a decline in model performance on both the training and test sets when the network depth reaches a certain level (Zohra et al., 2024). Even with the application of optimisation techniques such as batch normalisation (Li et al., 2024a), it is challenging to effectively address this issue. These problems severely restrict the development and application of deep neural networks.

The proposal of the residual network (ResNet) provides an innovative solution to the above-mentioned challenges (Khan et al., 2025; Wang et al., 2024d). Its core structure is the residual block, which breaks the linear connection pattern between layers in traditional neural networks by introducing skip connections. Suppose the input of a residual block is x, the output is y, and the mapping after a series of nonlinear transformations (Wei et al., 2024) (such as convolution operations, batch normalisation operations, activation functions, etc.) is $F(x)$. The output of the residual block can be expressed as:

$$y = F(x) + x \tag{1}$$

where x represents the input feature map of the residual block. In the network architecture, it is typically the output result of the previous network layer after operations such as convolution and pooling, carrying the feature information extracted by the preceding network. $F(x)$ is the residual mapping with respect to the input x. This mapping is composed of multiple convolutional layers, batch normalisation (BN) layers, and activation functions (such as the ReLU function) connected in series. Its core objective is to learn the difference, i.e., the residual part, between the input x and the target output. y is the output feature map of the residual block, which integrates the original input information and the learned residual information. This unique structural design enables

the network to focus on learning the residual part instead of directly learning complex identity mapping relationships, greatly reducing the learning difficulty of the network. Meanwhile, the existence of skip connections allows input information to be directly transmitted to subsequent layers, effectively preventing information loss caused by excessive network depth and providing strong support for alleviating gradient vanishing and explosion phenomena.

During the backpropagation process, the gradient calculation mechanism of the residual network undergoes a fundamental change due to skip connections. According to the chain rule, the gradient of the loss function L with respect to the input x can be derived as follows:

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial x} = \frac{\partial L}{\partial y}\left(1 + \frac{\partial F(x)}{\partial x}\right) \tag{2}$$

In the above formula, L represents the loss function of the entire network. It measures the degree of difference between the model's predicted results and the true labels through specific metrics (Connor et al., 2024; Huang and Ren, 2024) (such as cross-entropy loss function, mean squared error loss function, etc.), and it is a key indicator guiding the network's parameter optimisation. $\frac{\partial L}{\partial x}$ is the gradient of the loss function L with respect to the input x of the residual block. This gradient information determines the direction and magnitude by which the network parameters corresponding to the input x should be adjusted during the parameter update process to reduce the loss. $\frac{\partial L}{\partial y}$ is the gradient of the loss function L with respect to the output y of the residual block, reflecting how changes in the output y affect the loss function L. $\frac{\partial F(x)}{\partial x}$ is the gradient of the residual mapping $F(x)$ with respect to the input x, representing the sensitivity of changes in the residual part to the input x. It can be clearly seen from this formula that the gradient of the residual network consists of two parts. The constant term 1 is crucial as it ensures that even if the value of $\frac{\partial F(x)}{\partial x}$ approaches 0 during backpropagation, the gradient $\frac{\partial L}{\partial x}$ will not vanish. This makes the training process of deep networks more stable, effectively avoiding the problem of network non-convergence caused by gradient vanishing and greatly improving the efficiency of network training and the performance of the model.

Based on the excellent characteristics of the residual network, numerous researchers have conducted in-depth improvement and expansion work around it and widely applied it to multiple fields such as computer vision and natural language processing. As the backbone network, it can effectively extract the semantic information of images, providing rich features for subsequent pixel-level classification. This enables the model to accurately classify each pixel in the image, achieving high-quality semantic segmentation results. These research achievements and application practices are based on the residual network.

2.2 Transformer

In the development history of deep learning, RNNs and their variants, such as LSTMs and gated recurrent units (GRUs), were once the mainstream models for processing sequential data (Wang et al., 2024c). However, these models have significant drawbacks. For example, due to the vanishing gradient problem, RNNs struggle to handle long sequential data, and their recurrent computation mechanism prevents parallel processing, resulting in low training efficiency. Although LSTMs and GRUs alleviate the gradient problem by introducing gating mechanisms, they still cannot fundamentally solve the bottleneck of sequential computation. The introduction of the transformer architecture completely breaks this dilemma. Replacing the traditional recurrent structure with the novel self-attention mechanism, it has achieved revolutionary breakthroughs in numerous fields, including natural language processing and computer vision, providing crucial theoretical and technical support for the research of this paper (Madarapu et al., 2024).

The core of the transformer lies in the self-attention mechanism, which enables the model to dynamically calculate the degree of correlation between each position and other positions when processing sequential data, thus capturing long-range dependencies within the sequence (Pu et al., 2024). Suppose the input sequence is $x = [x_1, x_2, \ldots, x_n]$. For each position i, the input is first mapped to three different vector spaces through linear transformations to obtain the query vector q_i, the key vector k_i, and the value vector v_i:

$$q_i = W_q x_i k_i = W_k x_i v_i = W_v x_i \tag{3}$$

where x_i represents the vector at the i^{th} position in the input sequence, carrying the information corresponding to that position. W_q, W_k and W_v are all learnable weight matrices. Through training, the parameters are adjusted so that the model can learn appropriate mapping relationships, transforming the input into vectors suitable for calculating attention weights. q_i is used to calculate the degree of correlation with other positions, k_i is used to be queried by other positions for calculating the correlation, and v_i contains the actual information of this position.

Next, the similarity between the query vector q_i and all key vectors k_j ($j = 1, 2, \ldots, n$) is calculated using dot product and then normalised to obtain the attention weight α_{ij}:

$$\alpha_{ij} = \frac{\exp\left(q_i \cdot k_j / \sqrt{d_k}\right)}{\sum_{j=1}^{n} \exp\left(q_i \cdot k_j / \sqrt{d_k}\right)} \tag{4}$$

In this formula, d_k is the dimension of the key vector k. The introduction of $\sqrt{d_k}$ for scaling is to prevent the dot product result from being too large, which could lead to the vanishing gradient problem of the softmax function. α_{ij} represents the attention weight of the i^{th} position to the j^{th} position, reflecting the correlation strength between the two positions. The larger the weight, the more relevant the information of the two positions is in the current calculation.

Finally, the attention weights are weighted and summed with the value vectors v_j to obtain the output z_i of the self-attention mechanism:

$$z_i = \sum_{j=1}^{n} \alpha_{ij} v_j \tag{5}$$

where Z_i integrates the information related to position i throughout the entire input sequence. In this way, the model can effectively capture long-range dependencies within the sequence.

In practical applications, the transformer usually employs the multi-head attention mechanism. It consists of multiple independent self-attention heads that calculate in parallel, and then the results are concatenated and linearly transformed:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(head_1, \ldots, head_h\right) W^O \tag{6}$$

where $head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, Q, K, and V are the query matrix, key matrix, and value matrix generated from the input sequence, respectively. h represents the number of attention heads. W_i^Q, W_i^K, W_i^V and W^O are all learnable weight matrices. The multi-head attention mechanism can capture rich information from different representation subspaces, further enhancing the model's ability to extract sequential features.

Based on the transformer architecture, numerous pre-trained models, such as BERT and the GPT series, have emerged and achieved remarkable success in natural language processing tasks. In text classification tasks, BERT learns general language representations through pre-training on large-scale corpora, and after fine-tuning, it can reach leading performance in various classification tasks. In the field of machine translation, the transformer architecture significantly improves translation efficiency and quality due to its parallel computing advantage, becoming the fundamental architecture of current mainstream translation models.

3 Residual enhanced transformer for affective multi-part music generation

3.1 Residual dense connection encoder architecture

The encoder of the proposed music emotion analysis model employs a residual dense connection (RDC) mechanism to achieve hierarchical modelling of polyphonic music through cross-layer feature fusion. Traditional transformer architectures face challenges in handling complex musical structures due to information propagation bottlenecks. In contrast, our approach introduces cross-layer connections after each transformer module, integrating low-level features with high-level abstract representations. Specifically, the output H^l of the l^{th} encoder layer is computed as follows:

$$H^l = \text{LayerNorm}\left(\text{TransformerLayer}\left(H^{l-1}\right) + \text{Concat}\left(H^{1\ldots l-1}\right) + X\right) \tag{7}$$

where $H^l \in \mathbb{R}^{n \times d}$ denotes the hidden state matrix of the l^{th} layer, with n being the sequence length and d the feature dimension, and TransformerLayer(·) represents a standard.

Transformer layer operation, including multi-head self-attention and feed-forward networks, Concat($H^{1\ldots l-1}$) signifies the concatenation of all hidden states from the first to the $(l-1)^{\text{th}}$ layer along the feature dimension, $X \in \mathbb{R}^{n\times d}$ is the input embedding matrix, preserving the original input information.

To enhance the flexibility of feature fusion, a gating mechanism is introduced to selectively integrate cross-layer features:

$$G^l = \sigma\left(W_g \cdot \text{Concat}\left(H^{l-1}, H^{l-2}\right)\right) \tag{8}$$

$$H^l = \text{LayerNorm}\left(\text{TransformerLayer}\left(H^{l-1}\right) + G^l \odot \text{Concat}\left(H^{1\ldots l-1}\right) + X\right) \tag{9}$$

where σ is the sigmoid activation function, W_g is a learnable weight matrix, and $\odot$ denotes element-wise multiplication. This gated residual dense connection (GRDC) mechanism allows the model to adaptively select relevant historical features, enabling simultaneous capture of temporal dependencies across bass, middle, and treble voices while mitigating the vanishing gradient problem in deep networks.

3.2 Emotion-guided lightweight decoder

In the decoder design, we propose an emotion-guided sublayer residual connection (EGSRC) mechanism to integrate emotion label information E into the music feature decoding process. The following connection is introduced between the self-attention and feed-forward network sublayers:

$$S_2 = \text{LayerNorm}\left(\text{MultiHead}\left(S_1, E\right) + S_1 + X\right) \tag{10}$$

where S_1 and S_2 represent the input and output of the sublayer, respective, and MultiHead(S_1, E) denotes the multi-head attention computation with S_1 as queries and E as keys/values, X is the cross-layer input to enhance feature propagation.

To reduce model complexity, lightweight residual blocks (LRBs) are employed to replace traditional fully connected layers. Each LRB consists of a 1 × 1 dimensionality reduction convolution, a 3 × 3 feature extraction convolution, and a 1 × 1 dimensionality expansion convolution. The parameter count is calculated as:

$$\text{Params}_{light} = C_{in} \cdot k + k \cdot 3^2 \cdot k + k \cdot C_{out} \tag{11}$$

where $\text{Params}_{conv} = C_{in} \cdot 3^2 \cdot C_{out}$, C_{in} and C_{out} are the number of input and output channels, respectively, k is the intermediate layer compression factor.

Additionally, depthwise separable convolutions (DSConv) are introduced to further optimise the model structure:

$$\text{DSConv}(x) = \text{Pointwise}(\text{Depthwise}(x)) \tag{12}$$

where Depthwise convolution applies spatial convolution to each input channel individually, while Pointwise convolution combines the output channels using 1 × 1 convolutions. Through these optimisations, the decoder achieves a significant reduction in parameter count while maintaining feature representation capabilities and improving inference efficiency.

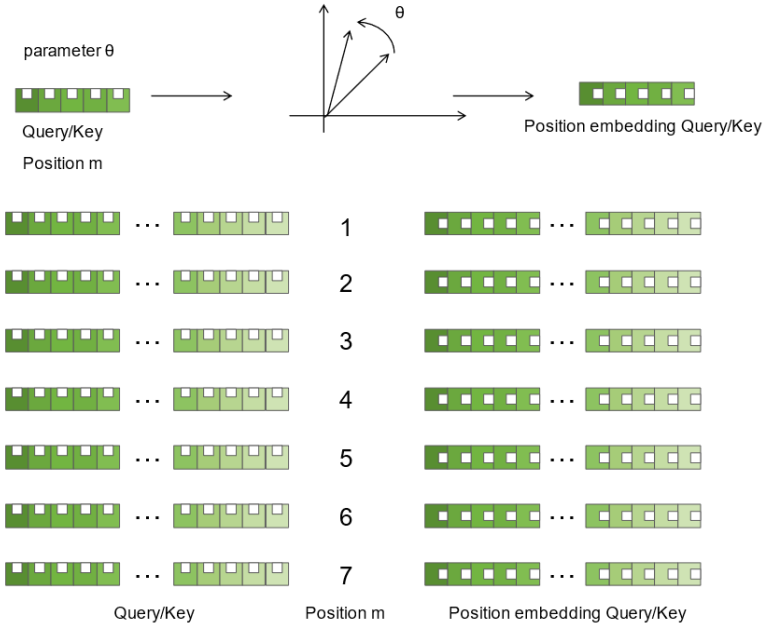
3.3 Rotary position embedding mechanism

To address the challenges of modelling temporal dependencies in long music sequences, rotary position embedding (RoPE) is adopted to replace traditional sinusoidal position encodings, as shown in Figure 1. Given query vector q_m and key vector k_n, the attention score is computed using RoPE as follows:

$$\text{Attention}\left(q_m, k_n\right) = \frac{\left(W_q x_m \cdot R_{\theta(m-n)}\right)\left(W_k x_n\right)}{\sqrt{d}} \tag{13}$$

where x_m and x_n are the input embeddings at positions m and n, respectively, W_q and W_k are linear transformation matrices, $R_{\theta(m-n)}$ is a rotation matrix whose parameters θ depend on the position difference ($m - n$).

Figure 1 Rotary position embedding mechanism (see online version for colours)



The rotation matrix R_θ is structured as:

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{14}$$

To adapt to the periodic nature of music data, we further propose periodic rotary position embedding (PRoPE), which modifies the parameter calculation method of the rotation matrix. This mechanism integrates absolute position information into vector representations, enabling the model to better capture long-range temporal dependencies and significantly enhancing the modelling of extended music sequences.

3.4 Emotion and music feature fusion framework

The emotion label E is represented by a one-hot encoding $e \in \mathbb{R}^K$ (where K is the number of emotion categories) and fused with the music symbol embedding $x_t \in \mathbb{R}^d$ as follows:

$$x_{t'} = x_t + W_e \cdot e \tag{15}$$

where $W_e \in \mathbb{R}^{d \times K}$ is a learnable projection matrix that maps the emotion label to the same dimensional space as the music features. Building on this, we introduce emotion-conditional layer normalisation (ECLN):

$$\mathrm{ECLN}(x_t) = \gamma_e \odot \frac{x_t - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta_e \tag{16}$$

where μ and σ are the mean and standard deviation of the input, respectively, γ_e and β_e are emotion-conditioned scaling and shifting parameters derived from the emotion label e.

The entire model is trained end-to-end by minimising a multi-task loss function:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{seq} + \lambda_2 \mathcal{L}_{con} + \lambda_3 \mathcal{L}_{adv} \tag{17}$$

where $\mathcal{L}_{cls}$ is the emotion classification loss, using the cross-entropy loss function, $\mathcal{L}_{seq}$ is the sequence reconstruction loss, preserving musical structural information, $\mathcal{L}_{con}$ is the contrastive learning loss, enhancing the discriminability of emotion features, $\mathcal{L}_{adv}$ is the adversarial training loss, strengthening the model's ability to extract emotion-related features, λ_1, λ_2, λ_3 are hyperparameters balancing the contributions of each loss component.

4 Experimental results and analyses

To comprehensively evaluate the performance of the proposed REAM model, experiments were designed from four dimensions: multi-model performance comparison, ablation analysis of core modules, cross-emotion generation capability verification, and subjective human preference assessment.

4.1 Comparative experiments: multi-model performance evaluation

The REAM model was compared with five representative baseline models:

- *Transformer-XL:* a representative architecture in long-sequence generation, it effectively mitigates the performance degradation of traditional positional encoding when handling long-range dependencies through relative positional encoding, demonstrating excellent performance in long-sequence modelling tasks.
- *CP transformer:* this model focuses on optimising sequence feature representation. It enhances the model's ability to understand and represent text semantic structures by designing a composite word encoding strategy.

- *CEG-Transformer:* an architecture that uses emotion labels as guidance. It achieves the deep integration of emotional information and the transformer architecture through the construction of emotion-driven attention mechanisms and network structures.
- *LSTM-Attn:* a typical example of the combination of recurrent neural networks and attention mechanisms. This model combines the advantages of long short-term memory networks (LSTM) in temporal data processing with the focusing ability of attention mechanisms, showing good adaptability and generalisation ability in temporal data modelling tasks.
- *Rule-based:* an expert system built based on harmony rules. It relies on established music theory rules to generate music.

The experiment employed three evaluation metrics: emotional consistency (EC), measured by the classification accuracy of the generated music using the DUPSO-DSKSVM algorithm to assess its alignment with the target emotion; spatial interaction complexity (SIC), evaluated via the note synchronisation index (SI) to gauge the rhythmic and pitch coordination among voices; and structural rationality (SR), scored on a 10-point scale based on the correctness of chord progressions according to music theory rules.

In terms of the emotional consistency (EC) metric, the rule-based model scored the lowest at 62.3%. This is because it completely relies on preset rules and lacks the ability to flexibly learn emotional features. LSTM-Attn and Transformer-XL scored in the middle range, indicating that traditional temporal modelling methods and relative positional encoding can capture certain emotional information, but they still have limitations in the accuracy of emotional expression in polyphonic music. CP transformer and CEG-Transformer significantly improved the EC index by optimising feature representation and introducing emotion guidance mechanisms. However, the REAM model further increased the EC to 87.5% with innovative designs such as the residual dense connection encoder and emotion-conditional layer normalisation, demonstrating its superiority in capturing and expressing emotional information.

Table 1 Comparative experiments result

Model	*EC (%)*	*SIC (SI)*	*SR*
Rule-based	62.3	0.21	7.8
LSTM-Attn	75.1	0.34	8.2
Transformer-XL	72.5	0.39	8.5
CP transformer	82.5	0.45	8.8
CEG-Transformer	85.0	0.48	8.9
REAM	87.5	0.56	9.2

Regarding the spatial interaction complexity (SIC), REAM's value of 0.56 was significantly higher than that of other baseline models. Especially compared with the best-performing baseline model, CEG-Transformer (0.48), the improvement exceeded 20%. This benefit comes from its cross-layer feature fusion mechanism in the encoder and the lightweight residual block design in the decoder, which can better model the complex rhythm and pitch coordination among multiple voices. In contrast, the

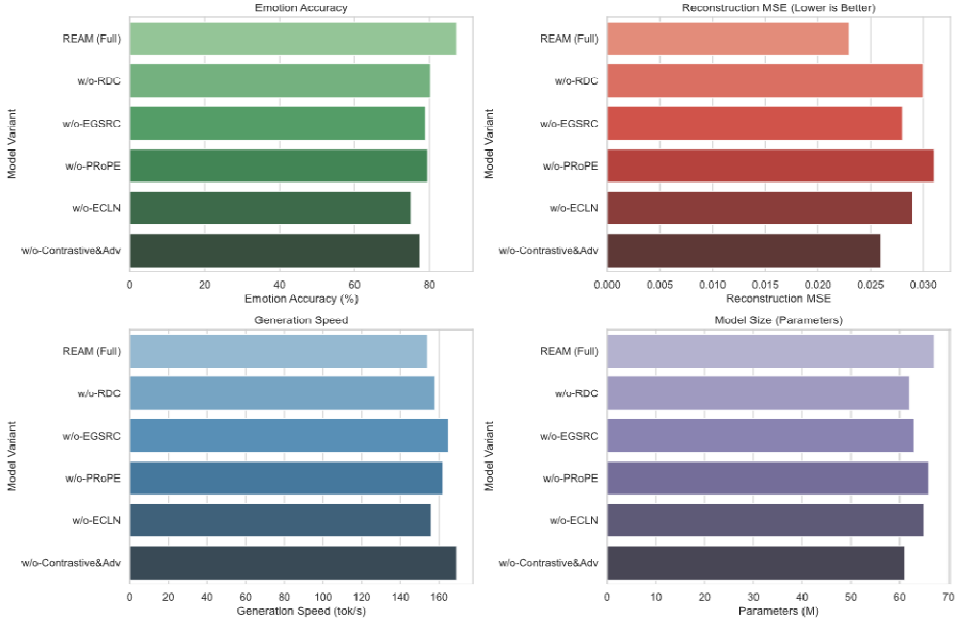
rule-based model only had an SIC of 0.21, reflecting that the generation method based on fixed rules has difficulty handling complex voice interactions.

In the structural rationality (SR) metric, the REAM model led with a score of 9.2, indicating that the music it generated was more in line with professional standards in terms of chord progressions and other music theory structures. Among the other baseline models, CP transformer and CEG-Transformer scored 8.8 and 8.9 respectively, indicating that the improved models based on transformer have certain advantages in structure generation. However, the REAM model further improved the rationality of music structure while ensuring emotional expression through the optimisation of the multi-task loss function.

4.2 Ablation experiments: verification of the effectiveness of key modules

To evaluate the effectiveness of each key component in the proposed REAM model, we conduct a comprehensive ablation study. This analysis investigates the impact of removing or replacing individual modules on emotion classification accuracy, multi-part music reconstruction quality, long-range dependency modelling, and model efficiency. By comparing simplified variants with the full REAM model, we assess the actual contribution of each module to overall performance.

Figure 2 REAM model variants performance comparison (see online version for colours)



The complete REAM architecture includes the RDC encoder, the EGSRC decoder, the PRoPE positional embedding, the emotional fusion mechanism (e.g., ECLN), and a multi-task loss function composed of classification loss, reconstruction loss, contrastive loss, and adversarial loss. Based on this full model, we design five ablated variants as follows:

- *w/o-RDC:* replaces the RDC encoder with a standard transformer encoder, removing cross-layer residual dense connections
- *w/o-EGSRC:* removes the emotion-guided stacked residual connections in the decoder, directly concatenating emotion labels to musical features
- *w/o-PRoPE:* substitutes the emotion-aware rotary positional encoding with conventional sinusoidal positional encoding
- *w/o-ECLN:* removes the emotional fusion module and uses standard LayerNorm without conditional normalisation
- *w/o-Contrastive&Adv:* removes contrastive and adversarial loss terms, keeping only classification and reconstruction loss.

Evaluation metrics include emotion classification accuracy (higher is better), music reconstruction error measured by mean squared error (MSE, lower is better), generation speed (tokens per second), and model size in parameter count.

Results show that w/o-RDC exhibits a significant drop in emotion classification accuracy and a noticeable increase in reconstruction error, indicating that the RDC encoder improves representational capacity through cross-layer feature reuse and mitigates gradient vanishing. As shown in Figure 2, the w/o-EGSRC variant demonstrates decreased emotion consistency in generated music, suggesting that the decoder's hierarchical emotion integration is critical for expressing intended emotional tones. w/o-PRoPE performs worst in long-sequence modelling tasks, validating that PRoPE effectively captures the periodic and hierarchical structures inherent in music. The w/o-ECLN variant yields poor classification accuracy and reduced generation quality, confirming that emotional fusion via conditional normalisation enhances the interaction between musical and emotional features. Finally, w/o-Contrastive&Adv retains reasonable reconstruction accuracy but produces emotionally monotonous outputs, highlighting the importance of contrastive and adversarial learning in constructing a more discriminative and expressive emotion space.

In conclusion, all core components contribute substantially to REAM's ability to generate emotionally consistent, structurally coherent, and computationally efficient multi-part music. The full REAM model consistently outperforms its ablated variants across all key evaluation metrics.

4.3 Emotion transfer experiments: cross-emotion generation ability

To evaluate the REAM model's ability to generate emotionally controllable multi-part music, we conducted an experiment by fixing the model parameters and varying only the input emotion labels – namely, 'excited', 'calm', 'sad' and 'tense'. For each emotion label, the model generated corresponding multi-voice music fragments. We then performed a comparative analysis of melodic and harmonic feature variations across the different emotional contexts.

From the melodic perspective, the REAM model exhibits the capacity to adjust note pitch ranges and rhythmic density based on emotional input:

- for 'excited' music, the increase in high-pitch melodies (+32%), the predominance of sixteenth notes (65%), and the overwhelming use of major chords (89%) together

create a bright, energetic musical atmosphere, aligning with listeners' perception of excitement

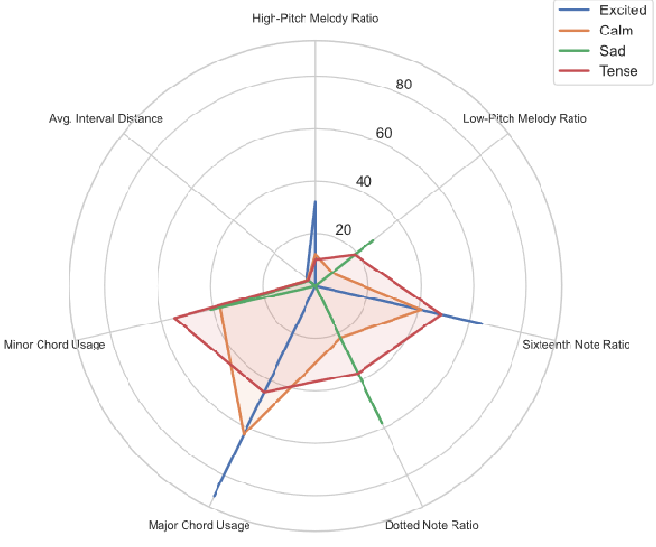
- for 'sad' music, the shift to lower-pitch melodies (+28%), higher ratio of dotted rhythms (58%), and increased use of minor chords (+41%) effectively convey a melancholic and subdued emotional tone
- the 'calm' and 'tense' types exhibit intermediate characteristics, with 'calm' music emphasising moderate pitch and rhythmic balance, while 'tense' music leans towards irregular rhythmic structures and minor harmonies, reflecting emotional tension.

From the perspective of multi-voice interaction, we observe notable differences in the average interval distance (i.e., the average pitch distance between simultaneous notes in different voices) under different emotional contexts:

- the 'excited' type exhibits the largest average interval (4.2 degrees), creating a more spatially open and harmonically dynamic sound texture, which enhances musical tension and emotional intensity
- conversely, the 'sad' type maintains a more compressed voice spacing (3.1 degrees), resulting in a tighter and more introspective auditory experience
- 'calm' and 'tense' show moderate interval distributions, reflecting their intermediate affective positioning.

These results demonstrate that REAM can translate abstract emotional semantics into concrete and diverse musical features, both in terms of pitch-rhythm construction and polyphonic coordination, as shown in Figure 3. The model's ability to dynamically adapt melodic and harmonic elements, as well as voice interactions based on emotion, confirms its fine-grained control over emotional expression and its musical interpretability.

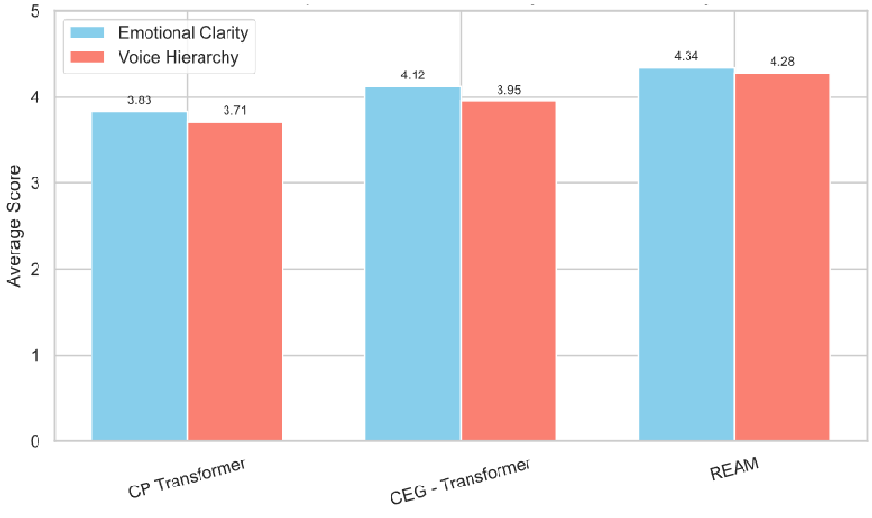
Figure 3 Radar analysis diagrams of music characteristics based on different emotions (see online version for colours)



4.4 Subjective listening experiments: human preference assessment

The experiment was set up as follows: 30 listeners, including 10 professional musicians and 20 ordinary listeners, were invited to rate the emotional expression clarity and voice hierarchy of the generated music on a 5-point scale.

Figure 4 Model comparison on emotional clarity and voice hierarchy (see online version for colours)



To highlight the REAM model's effectiveness from a human-perception angle, typical evaluations from different listener groups are provided, as shown in Figure 4. Professional musicians noted, "in the polyphonic fragments generated by REAM, the emotional cues of each voice are clear, and the emotional echo between the violin melody and the piano accompaniment is natural", emphasising its superior ability in creating harmonious, emotionally-consistent polyphony. These comments from both professional and non-professional perspectives jointly validate the REAM model's excellence in emotional music generation.

In terms of emotional expression clarity, the REAM model led with an average score of 4.34, surpassing the CP transformer (3.83) and CEG-Transformer (4.12). This is due to its unique emotion-music feature fusion framework. Through emotion-conditional layer normalisation and the optimisation of the multi-task loss function, the emotional cues of the generated music are more prominent and explicit. Both professional musicians and ordinary listeners can more clearly perceive the emotions conveyed by the music.

In the evaluation of voice hierarchy, the REAM model's average score of 4.28 was also higher than that of the other two baseline models. The evaluation from professional listeners, "the emotional cues of each voice in the polyphonic fragments generated by REAM are clear, such as the natural emotional echo between the violin melody and the piano accompaniment", indicates that the residual dense connection and cross-layer feature fusion mechanism of its encoder can effectively construct a multi-voice structure with distinct layers and harmonious coordination. The feedback from ordinary listeners, "the music fragments of different emotions can be easily distinguished", further

demonstrates that the REAM model not only performs well in emotional expression and voice construction but also has significant advantages in the overall perceptibility and distinguishability of music fragments. From the perspective of human subjective feelings, it fully verifies the practicality and effectiveness of the model.

5 Conclusions

In this paper, a residual enhanced transformer (REAM) for affective multi-part music generation is proposed, addressing challenges in long-range dependency modelling, multi-voice interaction, and emotional expression. Three core innovations are introduced: hierarchical residual dense connections enabling cross-layer feature fusion, emotion-aware rotary position encoding (ERoPE) for dynamic emotional modelling, and lightweight residual modules to balance efficiency and expressiveness. Experimental results on multiple metrics show significant improvements over baselines. The following conclusions can be drawn:

- The hierarchical residual dense connections in the encoder retain low-level musical details and enhance structural coherence by facilitating cross-layer feature interaction.
- ERoPE dynamically integrates emotion labels into positional encoding, improving the model's ability to capture temporal-affective dependencies and achieve fine-grained emotional control.
- Lightweight residual modules with depthwise separable convolutions reduce parameter overhead while maintaining expressive power, optimising computational efficiency for multi-track sequences.
- The multi-task loss function, including classification, reconstruction, contrastive, and adversarial losses, enhances emotional discriminability and structural rationality.
- The proposed REAM demonstrates superiority in generating emotionally coherent and structurally sophisticated multi-part music. However, the current focus on specific emotion categories may limit generalisation to highly nuanced emotional expressions. Future work will explore integrating more diverse emotional datasets and real-time interaction mechanisms to expand the model's applicability in practical music creation scenarios.

Declarations

The author declares that he has no conflicts of interest.

References

Almukhalfi, H., Noor, A. and Noor, T.H. (2024) 'Traffic management approaches using machine learning and deep learning techniques: a survey', *Engineering Applications of Artificial Intelligence*, Vol. 133, p.108147.

Ayres, L.B., Gomez, F.J., Silva, M.F., Linton, J.R. and Garcia, C.D. (2024) 'Predicting the formation of NADES using a transformer-based model', *Scientific Reports*, Vol. 14, No. 1, p.2715.

Bhandari, K. and Colton, S. (2024) 'Motifs, phrases, and beyond: the modelling of structure in symbolic music generation', *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, Springer, pp.33–51.

Briot, J-P. and Pachet, F. (2020) 'Deep learning for music generation: challenges and directions', *Neural Computing and Applications*, Vol. 32, No. 4, pp.981–993.

Chen, J., Ma, X., Li, S., Ma, S., Zhang, Z. and Ma, X. (2024) 'A hybrid parallel computing architecture based on CNN and transformer for music genre classification', *Electronics*, Vol. 13, No. 16, p.3313.

Connor, R., Dearle, A., Claydon, B. and Vadicamo, L. (2024) 'Correlations of cross-entropy loss in machine learning', *Entropy*, Vol. 26, No. 6, p.491.

Gao, B. and Li, Q. (2025) 'Social entertainment robot based on neural network algorithm in personalized music course simulation', *Entertainment Computing*, Vol. 52, p.100771.

Herrmann, L. and Kollmannsberger, S. (2024) 'Deep learning in computational mechanics: a review', *Computational Mechanics*, Vol. 74, No. 2, pp.281–331.

Huang, J. (2025) 'DSSViT: multi-scale adaptive fusion vision transformer with dense feature reuse for robust pneumonia detection in chest radiography', *International Journal of Imaging Systems and Technology*, Vol. 35, No. 3, p.e70127.

Huang, Y. and Ren, R. (2024) 'A GARCH model selection and estimation method based on neural network with the loss function of mean square error and model confidence set', *Journal of Forecasting*, Vol. 43, No. 8, pp.3177–3193.

Kang, J., Poria, S. and Herremans, D. (2024) 'Video2music: suitable music generation from videos using an affective multimodal transformer model', *Expert Systems with Applications*, Vol. 249, p.123640.

Khan, S.U.R., Zhao, M. and Li, Y. (2025) 'Detection of MRI brain tumor using residual skip block based modified MobileNet model', *Cluster Computing*, Vol. 28, No. 4, p.248.

Kwiecień, J., Skrzyński, P., Chmiel, W., Dąbrowski, A., Szadkowski, B. and Pluta, M. (2024) 'Technical, musical, and legal aspects of an ai-aided algorithmic music production system', *Applied Sciences*, Vol. 14, No. 9, p.3541.

Li, F. (2024) 'Chord-based music generation using long short-term memory neural networks in the context of artificial intelligence', *The Journal of Supercomputing*, Vol. 80, No. 5, pp.6068–6092.

Li, P., Liang, T-m., Cao, Y-m., Wang, X-m., Wu, X-j. and Lei, L-y. (2024a) 'A novel Xi'an drum music generation method based on Bi-LSTM deep reinforcement learning', *Applied Intelligence*, Vol. 54, No. 1, pp.80–94.

Li, R., Hu, M., Gao, R., Wang, L., Suganthan, P.N. and Sourina, O. (2024b) 'TFormer: a time-frequency Transformer with batch normalization for driver fatigue recognition', *Advanced Engineering Informatics*, Vol. 62, p.102575.

Liu, Z., Qian, S., Xia, C. and Wang, C. (2024) 'Are transformer-based models more robust than CNN-based models?', *Neural Networks*, Vol. 172, p.106091.

Madarapu, S., Ari, S. and Mahapatra, K. (2024) 'A deep integrative approach for diabetic retinopathy classification with synergistic channel-spatial and self-attention mechanism', *Expert Systems with Applications*, Vol. 249, p.123523.

Pu, Q., Xi, Z., Yin, S., Zhao, Z. and Zhao, L. (2024) 'Advantages of transformer and its application for medical image segmentation: a survey', *BioMedical Engineering Online*, Vol. 23, No. 1, p.14.

Wang, H., Zou, Y., Cheng, H. and Ye, L. (2024a) 'Diffuseroll: multi-track multi-attribute music generation based on diffusion model', *Multimedia Systems*, Vol. 30, No. 1, p.19.

Wang, J., Hong, S., Dong, Y., Li, Z. and Hu, J. (2024b) 'Predicting stock market trends using LSTM networks: overcoming RNN limitations for improved financial forecasting', *Journal of Computer Science and Software Applications*, Vol. 4, No. 3, pp.1–7.

Wang, L., Zhao, Z., Liu, H., Pang, J., Qin, Y. and Wu, Q. (2024c) 'A review of intelligent music generation systems', *Neural Computing and Applications*, Vol. 36, No. 12, pp.6381–6401.

Wang, S., Tian, J., Liang, P., Xu, X., Yu, Z., Liu, S. and Zhang, D. (2024d) 'Single and simultaneous fault diagnosis of gearbox via wavelet transform and improved deep residual network under imbalanced data', *Engineering Applications of Artificial Intelligence*, Vol. 133, p.108146.

Wei, Y., Wu, D. and Terpenny, J. (2024) 'Remaining useful life prediction using graph convolutional attention networks with temporal convolution-aware nested residual connections', *Reliability Engineering & System Safety*, Vol. 242, p.109776.

Wu, X., Xiang, B., Lu, H., Li, C., Huang, X. and Huang, W. (2024) 'Optimizing recurrent neural networks: a study on gradient normalization of weights for enhanced training efficiency', *Applied Sciences*, Vol. 14, No. 15.

Xin, Y. (2024) 'MusicEmo: transformer-based intelligent approach towards music emotion generation and recognition', *Journal of Ambient Intelligence and Humanized Computing*, Vol. 15, No. 8, pp.3107–3117.

Yang, Z., Wang, Z., Cao, X., Chen, B., Fan, R. and Lu, Y. (2024) 'Influences of concentration gradients and ignition positions on unconfined inhomogeneous hydrogen explosion', *International Journal of Hydrogen Energy*, Vol. 50, pp.857–869.

Zhao, Z., Li, Y., Peng, Y., Camilleri, K. and Kong, W. (2025) 'Multi-view graph fusion of self-weighted EEG feature representations for speech imagery decoding', *Journal of Neuroscience Methods*, Vol. 418, p.110413.

Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X. and Wang, R. (2024) 'A review of key technologies for emotion analysis using multimodal information', *Cognitive Computation*, Vol. 16, No. 4, pp.1504–1530.

Zohra, F.T., Webb, C.J., Lamb, K.E. and Gray, E.M. (2024) 'Degradation of metal hydrides in hydrogen-based thermodynamic machines: a review', *International Journal of Hydrogen Energy*, Vol. 64, pp.417–438.