ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Design and implementation of Intelligent Composition generation Model driven by Multi-modal music features

Lianghua Li[1,*]

[1] Hunan First Normal University the College of Music and Dance. Chang sha 410205, P. R. China

SUMMARY: *Generative artificial intelligence (AI) has accelerated its entry into digital content production scenarios and promoted the evolution of intelligent composition from single sequence prediction to multi-modal collaborative modeling. Aiming at the composition task driven by multi-modal music features, this paper constructs a generative model that integrates audio, MIDI, lyric text, style labels and emotion labels. Through unified feature representation, cross-modal attention fusion, hierarchical sequence generation, melody rhythm and harmony synergy constraints, as well as bi-conditional modulation of style and emotion and structure verification of music score. The closed-loop design from feature input to symbolic output is realized. Experimental results show that the melody coherence, style matching and emotional accuracy of the model reach 91, 90 and 87 points respectively, and the comprehensive quality score is 88.4, which is still 84.7 under 20% disturbance. In the lyric-assisted composition scene, the comprehensive score is 89.1, and the style preservation rate is 91.4%. The research shows that this method can improve the structure stability and expression consistency of the generated music, which has reference significance for the engineering implementation of intelligent composition system.*

KEYWORDS: *Multi-modal music feature; Intelligent composition; Music generation model; Cross-modal fusion*

1 Introduction

With the continuous evolution of digital audio processing, deep learning and generative artificial intelligence technologies, music composition is moving from the experience-led manual writing mode to a new stage of data-driven and intelligence-assisted parallelism. Traditional composition activities emphasize the composer's overall grasp of melody direction, rhythm organization, harmony configuration and emotional expression, and the formation process of works has strong personal style and aesthetic dependence. This method can reflect distinct artistic personality, but there are also practical problems such as long creation cycle, more repetitive labor, and limited style transfer efficiency. Especially in the context of the rapid expansion of application scenarios such as short video soundtracks, game interactive music, digital content production and personalized music services, we see that how to improve the efficiency of music generation and enhance the ability of work style adaptation with computer technology has become the focus of continuous attention in intelligent music research.

From the existing research, intelligent composition has developed from rule-based and probability statistics methods to relying on deep generative frameworks such as recurrent neural

*lilianghua84816@163.com
https://doi.org/10.65102/is2026215
Published: 30 April, 2026

networks, Transformers, and diffusion models. Related methods have made significant progress in melody continuation, segment completion, harmony generation and style imitation, but there are still some shortcomings. One kind of method mainly relies on a single modal input and only learns music rules from MIDI note sequences or audio signals, which is difficult to make full use of multi-source information such as lyrics semantics, emotional labels, rhythm structure, and performance style, resulting in deviations in the generated results in terms of emotional consistency, structural integrity and style control. The other kind of methods introduce multi-modal information, but often simply concatenate different modalities, failing to effectively solve the problems of inconsistent time scale, inconsistent semantic level and insufficient feature coupling, which affects the ability of the generation model to describe the internal logic of music. For intelligent composition, we believe that what is truly valuable is not the generation of a number of notes in isolation, but the joint action of multimodal cues to form a complete musical fragment that takes into account melody audibility, rhythm coordination, harmonic rationality and emotional expression.

Based on this, this paper focuses on the intelligent composition task driven by multi-modal music features, and tries to build a composition generation model that takes into account feature expression, generative modeling and system implementation. Our research focuses on the unified representation of multi-modal music information, the construction of cross-modal fusion mechanism, the introduction of generation constraints, and the optimized output of generated results. We strive to make the model not only learn the statistical law of music sequences, but also perceive the guiding role of lyrics semantics, emotional tendencies and style labels on composition results. On this basis, we further combine experimental evaluation, comparative analysis and system implementation to verify the generation quality and application feasibility of the model. Through this research, we hope to promote the music generation technology from single note prediction to multi-dimensional information collaborative modeling, improve the artistic expression and engineering implementation ability of intelligent composition system, and provide certain reference for related research and application practice.

2 Literature Review

In recent years, the research of intelligent composition has gradually shifted from single note prediction to multi-modal collaborative generation. Liu et al.(2025) proposed the MusDiff framework, which introduces text and images into the music generation process, and combines IP-Adapter and KAN optimization feature fusion and modal alignment based on the diffusion model, indicating that multi-modal conditional constraints are helpful to improve the semantic consistency and detail performance of generated music [1]. Zhang and Yu(2025) further combined GAN and Transformer into multimodal music synthesis, emphasizing the synergistic relationship between generation authenticity, temporal modeling ability and joint expression of multi-source information [2]. Combining these studies, we can see that generative models are no longer limited to melody continuation, but begin to focus on the co-shaping of music content and expressivity by cross-modal conditions.

Around the research of music understanding, multimodal emotional modeling has become an important support. Wang(2025) starts from the joint modeling of audio time-frequency features and text semantics, and uses improved CNN, BiLSTM and cross-modal Transformer to build an emotion recognition model, which shows that emotion labels can not only be used for classification, but also serve the reverse direction of emotion control in the generation process [3]. Although Zhu et al.(2025) discussed language-guided cross-modal semantic fusion retrieval, the idea of realizing semantic alignment through unified embedding space provided a

method reference for modeling the association between music, text and visual information [4]. Chen and Wu(2022) earlier verified the advantages of multimodal signals in emotion discrimination from the analysis of music types and children's emotions [5]. Shi et al.(2024) combined music theory inspired features with self-supervised representation to further reduce the distribution differences between different modalities, reflecting the research trend of the fusion of "knowledge prior + deep representation" [6]. From this research vein, we can find that multimodal emotion modeling has gradually become an important support link between music understanding and music generation.

At the level of system implementation and engineering application, Zhang(2025) proposed TransVAE-Music system, which combines audio and video learning, Transformer and SketchVAE, and introduces real-time perception and Bayesian optimization mechanism in Internet of Things scenarios, indicating that intelligent composition is moving from offline generation to deployable system [7]. Hao et al.(2025) focuses on real-time music emotion recognition, providing a foundation for real-time interactive composition by leveraging Bi-LSTM, feature fusion, and adaptive sampling to balance accuracy and latency [8]. Zhang(2025) proposed the TS-Resformer model, which used residual network and Transformer in parallel for music signal classification, indicating that the joint extraction of time series information and spectral features is still an important direction of music modeling [9]. Pan et al.(2025) realizes hierarchical cross-modal generation under the condition of missing modalities, which provides inspiration for modal defect completion in multimodal music tasks [10]. Liu(2025) applied large language models to piano music style classification and trend analysis, showing the potential of the combination of symbolic features, text metadata and LLM in music understanding and style prediction [11]. In general, the existing research has covered multi-modal generation, emotion recognition, style classification and cross-modal alignment, but there are still some shortcomings in the unified modeling for composition tasks: first, the generation model is not closely connected with the emotion control module, second, the multi-modal features mostly stay at the parallel fusion level, and third, the structural constraints of the generated results and the output of the score are not enough attention. Combined with the above research progress, we believe that these shortcomings also constitute the entry point for further research in this paper.

Table 1: Main directions and implications of related research

Research Direction	Representative References	Main Content	Implications for This Study
Multimodal conditional music generation	[1][2]	Introduces conditions such as text and images into music generation to enhance modal alignment and generative performance	Indicates the need to build a cross-modal generative backbone for composition tasks
Music emotion recognition and modeling	[3][5][6][8]	Uses audio, text, physiological, or other multimodal signals to identify emotional states	Suggests that emotional information can serve as an important control variable in music composition generation
Cross-modal semantic alignment	[4][10]	Achieves cross-modal associations through unified embedding spaces or missing-modality reconstruction	Provides methodological support for aligning lyrics, music, and visual information
Automatic composition system implementation	[7]	Integrates generative models with real-time perception and parameter optimization into a complete system	Indicates that this study should balance model design with system implementation
Music style classification and trend analysis	[9][11]	Combines deep networks and large language models for music genre classification and trend prediction	Provides references for style control and style-label modeling

3 Proposed method

3.1 Overall system architecture and multimodal composition generation process

The intelligent composition system constructed in this paper consists of a multi-modal input layer, a feature encoding layer, a cross-modal fusion layer, a music generation layer, and a result optimization output layer. The input receives audio clips, MIDI event sequences, lyrics text, emotion labels and style labels at the same time. There are obvious differences in time granularity, semantic level and expression form of information from different sources. Therefore, the system first extracts rhythm, pitch, timbre, semantic and emotion features respectively, and then realizes cross-modal alignment through a unified representation space. In order to ensure that the generated results have both melodic coherence and style controllability, the conditional guidance vector is introduced in the fusion stage, and the emotional state and style constraints are embedded into the composition process. Let the audio, symbol sequence and text modal inputs be xa,xm and xt respectively, and the corresponding encoding representations are h_a,h_m and h_t, then the multimodal fusion representation can be written as follows:

$$z = \phi(W_a h_a + W_m h_m + W_t h_t + b) \tag{1}$$

where W_a,W_m,W_t are the mapping matrices of each modality, b is the bias term, and $\phi(\cdot\ )$ is the nonlinear activation function. Based on the fusion representation z and the condition vector c, the generator outputs a sequence of note events in time steps with conditional probabilities written as follows:

$$\mathrm{P}(\mathrm{Y}|\mathrm{z},\mathrm{c}) = \prod_{\mathrm{i}=1}^{\mathrm{T}} \mathrm{P}\,(\mathrm{y_i}|\mathrm{y_i} < \mathrm{z},\mathrm{c}) \tag{2}$$

where $\mathrm{Y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{T}\}$ denotes the generation of the phrase. Then, cross-modal fusion and conditional constraint injection are performed. Then, the generated network outputs the melody, rhythm and harmony information. Finally, the mode consistency check, beat legitimacy check and music symbolic output are completed in the post-processing module. Such a process not only retains the complementary value of multi-source music information, but also provides a unified entry for subsequent model training, style control and system implementation. Figure 1 shows the overall architecture of the intelligent composition generation model driven by multimodal music features.

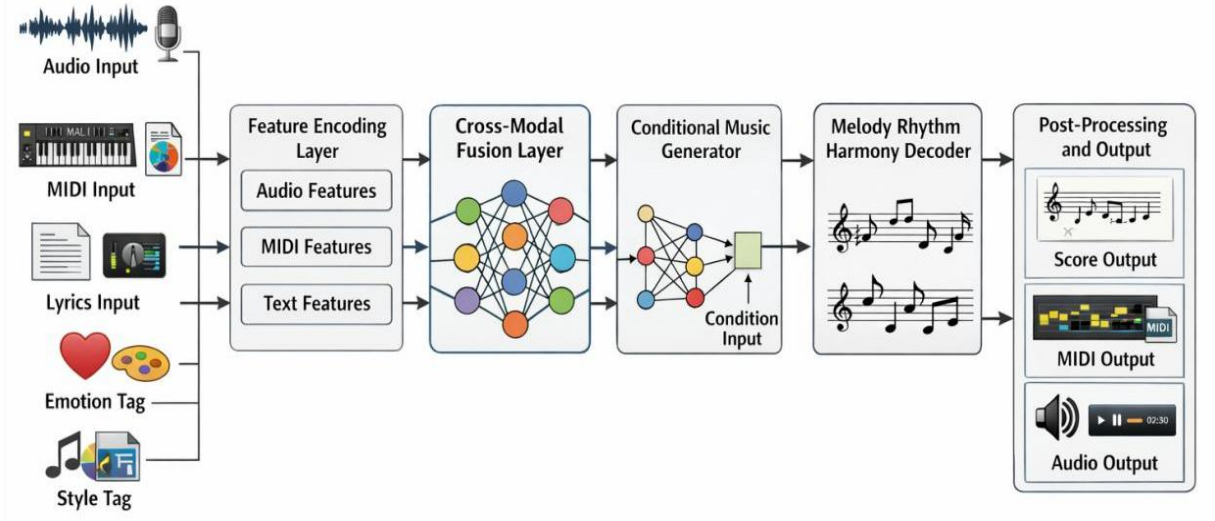

Figure 1: Overall architecture diagram of intelligent composition generation model driven by multimodal music features

3.2 Multimodal music feature representation and cross-modal fusion mechanism

In this paper, the audio modality is used as the source of acoustic performance information, and the continuous features such as MEL spectrum, rhythm intensity, timbre envelope and dynamic change are extracted. The MIDI modality is used as the source of symbol structure information to encode discrete events such as pitch, duration, strength, beat position and chord direction. The lyrics text was used as the source of semantic and emotional clues, and word vectors, syntactic dependencies and emotional polarity were extracted. Style and emotion labels are embedded into the generator as conditional control signals. The original distributions of different modalities are quite different, and if they are directly input into the generation network, it is easy to cause the problems of semantic dislocation and time step inconsistency. Therefore, this paper first performs linear mapping and time sequence alignment of each modality, and then enters the fusion module to complete the joint modeling.

Let the audio, MIDI and text inputs be denoted as x_a,x_m and x_t respectively, and the corresponding encoder outputs be h_a,h_m and h_t, then the unified projection representation of each modality can be written as follows.

$$\bar{h}_i = W_i h_i + b_i, \quad i \in \{a, m, t\} \tag{3}$$

where Wi is the modal mapping matrix and bi is the bias term. Considering that the relationship between lyrics semantics and music structure is not linear, this paper introduces a cross-modal attention mechanism in the fusion stage, takes MIDI event sequence as the dominant timeline, and uses audio and text features to enhance the context. The calculation form is as follows.

$$\alpha_{ij} = \text{softmax}\left(\frac{(Q\tilde{h}_m^i)(K\tilde{t}_j)^T}{\sqrt{d}}\right), z_i = \sum_j \alpha_{ij} V\tilde{h}_j + W_c c \tag{4}$$

where Q,K and V represent the query, key and value transformation matrices respectively, d is the feature dimension, and c is the sentiment and style condition vector. After this process, the fused representation zi simultaneously preserves the correlation information between melody timing, semantic sentiment, and acoustic performance. In order to reduce the interference of

single modal noise on the fusion results, the entry control weight is also added in the training to suppress the modes with low contribution, so that the generated network can maintain a more stable input quality in different scenarios. The cross-modal fusion mechanism constructed in this way provides a more detailed feature basis for subsequent melody generation, harmony organization and style control. Table 2 shows the representation and fusion of music features of each modality.

Table 2: Representation and fusion of musical features of each modality

Modality Type	Main Input Content	Core Features	Representation Form	Role in Fusion
Audio modality	Raw audio clips, performance segments	Mel spectrograms, rhythmic intensity, timbre envelope, dynamic variations	Continuous vector sequences	Provides acoustic expression and emotional details
MIDI modality	Note events, time signatures, chord information	Pitch, duration, velocity, beat position, harmonic progression	Discrete event embeddings	Provides melodic skeleton and structural constraints
Text modality	Lyrics, theme descriptions, semantic prompts	Word embeddings, syntactic relations, sentiment polarity, keywords	Semantic embedding sequences	Provides semantic guidance and emotional cues
Emotion labels	Joyful, lyrical, tense, soothing, etc.	Emotion categories, intensity levels	Conditional vectors	Controls the emotional tone of generated music
Style labels	Classical, pop, jazz, electronic, etc.	Style categories, style weights	Conditional embeddings	Controls the stylistic tendency of the composition
Cross-modal fusion layer	Unified multimodal input	Attention weights, gating coefficients, shared semantic representations	Joint feature vectors	Establishes inter-modal associations and outputs unified representations

3.3 Intelligent composition generation model construction and training target design

After the unified representation of multi-modal features is completed, the intelligent composition task is modeled as a sequence generation process with "structure leading, event progressive, and controllable conditions". Different from the way that only relies on a single note history to predict, the generative model in this paper simultaneously considers the continuous constraints on the output by the phrasp-level structural intention, the timestep level note events, and the style emotional conditions. The structure planner is responsible for forming the measure level latent representation, the condition generator gradually generates event information such as pitch, duration and intensity according to the latent structure, and the output mapper converts the hidden state into a decoded symbol sequence. In order to avoid the problem that the generated results are locally reasonable but overall loose, this paper adopts the hierarchical generation idea. Firstly, the structure summary of the music segment is established at the paragraph level, and then the fine-grained expansion is performed at the event level.

Let the sequence after cross-modal fusion be represented as $\mathcal{Z} = \{z_1, z_2, \ldots, z_T\}$,The structure vector of the KTH phrase is denoted as pk, and its construction mode is defined as follows.

$$p_k = \tanh\left(W_p \cdot \frac{1}{|S_k|} \sum_{t \in S_k} z_t + b_p \right) \tag{5}$$

Here, S_k represents the set of time steps corresponding to the KTH phrase, and W_p and b_p are trainable parameters, respectively. This formula is used to extract the phrase level summary from the local temporal features, so that the model not only depends on the preceding events, but also retains the overall structural intention of the current phrase when generating the note at a certain time. For any time step t, the state update of the conditional generator is expressed as follows.

$$\mathrm{s_t} = \sigma\big(\mathrm{W_s}[\mathrm{s_{t-1}} \| \mathrm{e_{t-1}} \| \mathrm{p_{\gamma(t)}} \| \mathrm{c}] + \mathrm{b_s}\big) \tag{6}$$

where s_{t-1} is the hidden state at the last moment,e_{t-1} is the event embedding of the previous note, $p_{\gamma(t)}$is the structure vector of the phrase at the current moment, c is the encoding of style and emotion conditions, and // represents vector concatenation. The design enables melodic development, local context, and global conditions to work synergistically within the same state space. Furthermore, the model outputs a distribution of note events for each time instant:

$$\hat{\mathrm{y}}_\mathrm{t} = \mathrm{Softmax}(\mathrm{W_o}(\mathrm{s_t} \odot \mathrm{g_t}) + \mathrm{b_o}) \tag{7}$$

Here, g_t is the gated modulation vector and $\odot$ represents element-wise multiplication. The gating mechanism can automatically adjust the participation strength of structural information and conditional information according to the current context, so that the model can generate different generation bias in the chorus, transition segment or bandar segment.

In terms of training objectives, we construct a joint optimization objective consisting of event accuracy term, mode consistency term and rhythm density balance term.

$$\mathcal{L} = \mathcal{L}_{evt} + \lambda_1 \mathcal{L}_{key} + \lambda_2 \mathcal{L}_{rh} \tag{8}$$

The event prediction term is used to ensure the basic generation correctness. The mode consistency term is used to constrain the deviation degree between the generated note and the target mode center. The rhythm density term is used to suppress extremely sparse or overly dense note distributions. Their specific forms are respectively defined as follows.

$$\mathcal{L}_{key} = \frac{1}{T} \sum_{t=1}^{T} (1 - \rho(\hat{n}_t, \kappa)) \tag{9}$$

$$\mathcal{L}_{rh} = \frac{1}{B} \sum_{b=1}^{B} \left| \frac{m_b}{l_b} - r^* \right| \tag{10}$$

where, $\rho(\hat{n}_t, \kappa)$ represents the matching degree between the generated note $\hat{n}_t$ and the target mode set κ, m_b represents the number of notes in the BTH bar, lb represents the duration length of the bar, and $r*$ is the expected rhythm density. Through this joint target design, the model can not only improve the quality of note level prediction, but also take into account the mode stability, rhythm balance and style controllability as a whole. The main modules of the intelligent composition generation model and the design of training objectives are shown in

Table 3.

Table 3: Main modules and training objective design of intelligent composition generation model

Module Name	Main Input	Main Output	Core Function	Corresponding Training Objective
Structure planner	Cross-modal fused feature sequence	Phrase-level structural vectors	Extracts segment summaries and forms global layout priors	Enhances structural continuity
Conditional generator	Historical events, structural vectors, conditional encodings	Temporal hidden states	Progressively generates events such as pitch, duration, and velocity	Improves event prediction capability
Gated modulation unit	Hidden states, emotion and style conditions	Modulation vectors	Dynamically allocates the influence weights of different conditions on generation	Enhances style and emotion control
Output mapper	Hidden states and modulation results	Probability distribution of note events	Completes the mapping to compositional symbolic outputs	Ensures decodability of generated results
Tonality constraint branch	Generated note sequences, target tonality	Tonality consistency error	Controls tonal stability of generated melodies	Reduces the risk of out-of-key notes
Rhythm constraint branch	Measure-level event statistics	Rhythm density error	Balances note distributions across measures	Avoids rhythmic imbalance

3.4 Generative constraint mechanism for melodic rhythm and harmony synergy

In the process of intelligent composition, if we only rely on the generative model to predict the probability of the next note event, although we can obtain a formal continuous note sequence, it is often prone to problems such as excessive melody jump, rhythm center drift and loose harmony support. In order to make the output more in line with the basic law of music composition, we introduce three types of collaborative constraints in the generation stage, which are used as the basis for structure modification in the decoding process. Instead of simply filtering the generated results, the proposed mechanism continuously measures the matching degree between the current note and the local phrase, beat frame and harmonic background during the time step, so as to improve the overall musical sense and structural stability.

Melodic constraints mainly control the smoothness of the pitch direction and the singability within the phrase. Let the pitch of the t-th note be qt, then the melodic continuity cost is defined as follows.

$$C_{mel} = \frac{1}{T-1}\sum_{t=2}^{T}[\max(0, |q_t - q_{t-1}| - \delta) + \mu 1(q_t = q_{t-1} = q_{t-2})] \tag{11}$$

Here, δ represents the allowable natural jump threshold and μ is the polyphonic penalty coefficient. This formula not only suppresses excessive interval mutation, but also avoids the rigidity of lines caused by the melody staying at the same pitch for a long time.

Rhythm constraints focus on the downbeat correspondence within the bar and the balance of time value distribution. The timing value of the JTH note in bar b is denoted as $u_{b,j}$, and its beat weight is denoted as $\omega_{b,j}$. Then the rhythm coordination degree can be expressed as follows.

$$C_{rhy} = \frac{1}{B}\sum_{b=1}^{B}\left|\frac{\sum_{j=1}^{N_b}\omega_{b,j}\,u_{b,j}}{\sum_{j=1}^{N_b}u_{b,j}} - \tau_b\right| \tag{12}$$

Here, τ_b is the target rhythm center of gravity in this subsection. Through this constraint, the model can reduce the phenomenon of downbeat misplacement and weak beat stacking, and make the rhythm advance more in line with the beat sign logic.

The harmony constraint is then used to measure the compatibility between the melodic tone and the accompanying chord. Let rt be the set of chords at time t, and ψ(qt,rt) denote the attribution score of the melody tone in the current chord. Then the harmony consistency cost is written as follows.

$$C_{har} = \frac{1}{T}\sum_{t=1}^{T}(1 - \psi(q_t, r_t)) \tag{13}$$

When the melody tone belongs to the core tone of the chord, the value of this term is small. The cost rises when persistent non-sum sounds are present. After synthesizing the three types of constraints, the collaborative control item is written as follows.

$$\Omega = \alpha C_{mel} + \beta C_{rhy} + \gamma C_{har} \tag{14}$$

Here, α,β,γ represent the melody, rhythm and harmony constraint weights, respectively. With the help of this collaborative mechanism, the model generation results form a unified and mutually supporting relationship among melody fluency, rhythm organization and harmony fit, which lays a more stable structural foundation for subsequent style control and score output.

3.5 Music style control and emotional expression guidance strategy

Intelligent composition is not only to complete the continuous generation of note sequences, but also to make the work present recognizable style characteristics and stable emotional direction in the overall listening sense. In this paper, a dual-conditional modulation strategy of style control and emotion guidance is introduced into the generative network, which considers the style label as a long-term constraint signal and the emotional state as a dynamic guidance signal, and acts on the composition process together through hierarchical injection. Among them, style control is responsible for limiting the overall performance framework of the work, such as classical, popular, jazz and other creative orientations. Emotional guidance focuses on the strength change, tone distribution, rhythm tension and melody direction in local paragraphs, so that the generated results retain enough emotional fluctuations and expression levels under

the premise of overall unity.

Let style label embedding be v_s and sentiment label embedding be v_e, then the bi-conditional joint control vector is defined as follows.

$$g_t = \eta_t \odot \tanh(W_s v_s) + (1 - \eta_t) \odot \tanh(W_e v_e) \tag{15}$$

Here, W_s and W_e denote the style mapping matrix and sentiment mapping matrix, respectively, and ηt is the modulation coefficient related to the time step. This formula is not a static average fusion of style and emotion, but dynamically adjusts the participation ratio of the two types of conditions according to the current generation context. When the work is in the paragraph expansion stage, the model emphasizes more on style consistency. When the work enters the climax, transition or bandha position, the weight of the emotional signal is correspondingly enhanced, so that the generated content is more in line with the actual change law of musical expression.

In order to ensure that the dual-condition control can deeply affect the hidden state update process, we further adopt the conditional affine modulation mechanism to adjust the scaling and offset of the decoder hidden representation. Let the original hidden state of the generator at time t be ht, then the modulated state is expressed as follows.

$$\hat{h}_t = \Gamma(g_t) \odot h_t + \mathcal{B}(g_t) \tag{16}$$

Here, Γ(·) and B(·) represent the scaling and offset terms generated by the control vector, respectively. After this process, the model no longer just appends style and emotional conditions from the outside, but directly changes the state distribution in the hidden space, so that the melody fluctuation, rhythm density and tone layout can reflect the corresponding expression tendency. In order to further constrain the output effects of style and sentiment, this paper sets the style consistency loss and sentiment matching loss:

$$\mathcal{L}_{style} = -\sum_{k=1}^{K} y_k^{(s)} \log \hat{y}_k^{(s)}, \mathcal{L}_{\text{emo}} = 1 - \frac{u(\hat{Y}) \cdot u(Y^e)}{\|u(\hat{Y})\| \|u(Y^e)\|} \tag{17}$$

where, $\hat{y}_k^{(s)}$ represents the prediction probability of the generated result on the KTH style, and $y_k^{(s)}$ is the target style label. $u(\hat{Y})$ represents the sentiment representation vector extracted from the generated music, and $u(Y^e)$ represents the target sentiment vector. The former is used to strengthen the stability of style attribution, and the latter is used to narrow the deviation between the target sentiment and the generated sentiment. Therefore, the model not only learns the generation law at the note level, but also learns the style boundary and emotional direction simultaneously during training, so that the output works form a closer correspondence between the macro style and the micro expression. Figure 2 shows the dual-condition control mechanism of style and emotion.

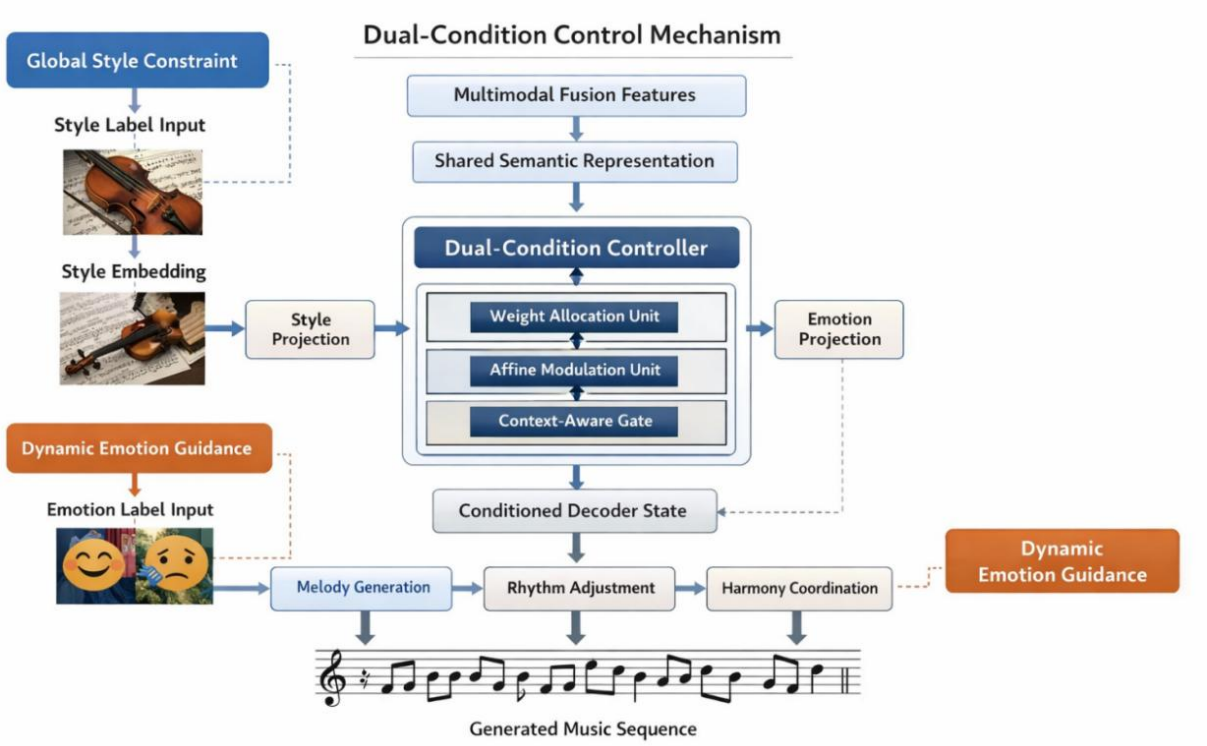

Figure 2: Control mechanism diagram of style and emotion dual condition

3.6 Optimization of generated results and verification method of musical score structure

After decoding by the generative model, although the results have the basic melody shape, there may still be problems such as uneven bar values, local range out of bounds, unstable end of sentences, and too many repeated segments in the continuous output condition. If the result is directly written into MIDI or music score, it will not only weaken the audibility of the work, but also affect the standardization of the subsequent symbolic output. Based on this, this paper sets up a result optimization and structure verification module after the generation end, which performs beat alignment, pitch range screening, mode checking, termination judgment and symbol correction on the candidate note sequence in turn, so that the model output is further changed from "generable" to "usable".

Let the sequence of candidate works be $M = \{(p_i, d_i, v_i)\}_{i=1}^{N}$, where p_i,d_i and v_i denote the pitch, duration and intensity of the ith note, respectively. For the structural integrity of subsection, this paper defines the time value deviation term as follows.

$$D_{bar} = \frac{1}{B}\sum_{b=1}^{B}\frac{\left|\sum_{j=1}^{n_b} d_{b,j} - T_b\right|}{T_b} \tag{18}$$

where B is the total number of bars, Tb is the theoretical timing capacity of the b bar, db,j is the timing value of the JTH note within the bar. This formula is used to detect the case of over-full or over-empty sections, and the correction is completed by splitting, merging or extending the rest in post-processing. Aiming at the possible range imbalance problem in the melody output, the range penalty term is further constructed as follows.

$$D_{rng} = \frac{1}{N}\sum_{i=1}^{N}[\max(0, p_i - p_{max}) + \max(0, p_{min} - p_i)] \tag{19}$$

where, p_{max} and p_{min} represent the highest and lowest tone allowed in the target style, respectively. This item can avoid the long-term deviation of the generated results from the reasonable singing or playing interval.

In order to ensure the natural wrapping of phrases, this paper also introduces an index of termination stability. Let the ending sound of the u-th phrase be e_u and the corresponding set of terminating harmonics be Ω_u, then:

$$Q_{cad} = \frac{1}{U}\sum_{u=1}^{U} 1(e_u \in \Omega_u) \tag{20}$$

where U is the total number of phrases and $1(\cdot\ )$ is the indicator function. The higher this index is, the easier it is to form a stable drop at the end of a sentence. After integrating all constraints, the post-processing score of the candidate result is written as follows.

$$J(M) = \lambda_1 e^{-D_{bar}} + \lambda_2 e^{-D_{rng}} + \lambda_3 Q_{cad} - \lambda_4 R_{rep} \tag{21}$$

Here, R_{rep} represents the proportion of repetitive segments, and λ_1,λ_2,λ_3, and λ_4 are the weight coefficients. The system selects the sequence with the largest J(M) from the candidate set as the final output, and completes the symbolic arrangement of articulation, rest, strength mark and clause boundary on this basis. After this process, the generated results have been significantly improved in the aspects of beat closure, pitch range distribution, sentence end stability and notation specification, and are more in line with the needs of subsequent experimental evaluation and system deployment. Figure 3 shows the process of optimizing the generated results and verifying the music score structure.

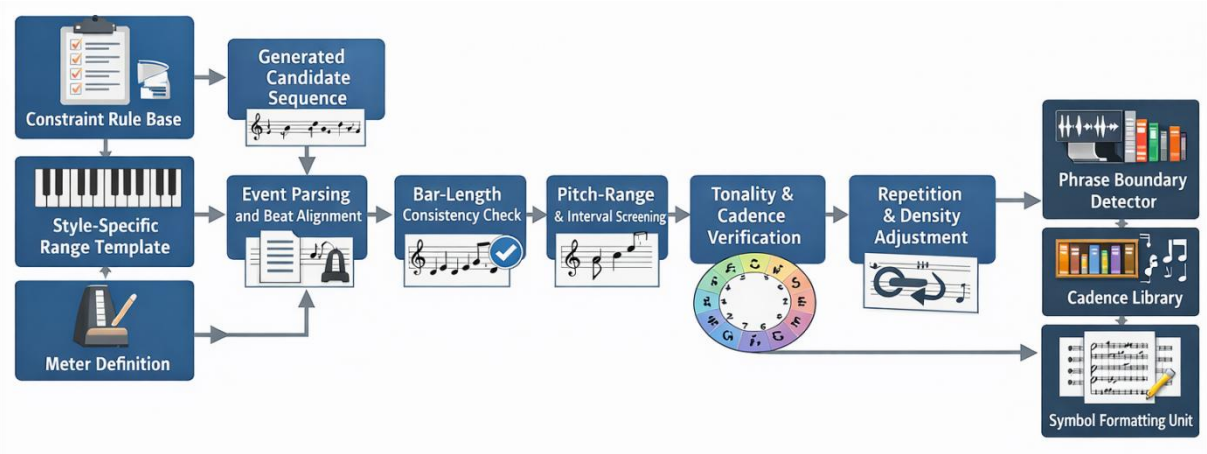

Figure 3: Flow chart of optimization of generated results and verification of musical score structure

4 Experimental results and discussion

4.1 Data set construction Data division and evaluation system design

In order to verify the composition generation ability of the proposed model under multi-modal conditions, this paper constructs an experimental dataset consisting of audio clips, MIDI symbol sequences, lyrics text, style labels and emotion labels. The data sources include public music clips, manually curated lyrics corpus, and style and emotion labels processed by unified annotation rules. Considering the differences in length, sampling rate and semantic granularity between different modalities, audio frame truncation, MIDI event normalization, lyric clause alignment and label cleaning are completed before the experiment, and then unified samples are formed with phrases as the basic units. In order to avoid information leakage in the training process, this paper divides the training set, validation set and test set according to the ratio of 7∶1.5∶1.5, and ensures that the same piece of work does not reappear across sets. Both objective and subjective indicators are introduced into the evaluation system: the objective level mainly examines the melody coherence, rhythm rationality, harmony consistency, style matching and emotional accuracy, and the subjective level invites reviewers with music learning background to rate the audibility, completeness and expressiveness of the generated results. Through this design of "data division-index evaluation-manual review", the actual performance of the model in the multi-modal intelligent composition scene can be more comprehensively reflected. Table 4 shows the composition and division of the data set.

Table 4: Composition and division of multimodal music dataset

Data Modality	Original Sample Size	Valid Sample Size	Main Content	Training Set	Validation Set	Test Set
Audio clips	4820	4516	Melody, rhythm, timbre, dynamic variations	3161	677	678
MIDI sequences	4820	4468	Pitch, duration, beat positions, chord events	3128	670	670
Lyric text	4360	4092	Segmented lyric text, semantic themes, emotion words	2864	614	614
Style labels	4820	4516	Classical, pop, jazz, electronic, etc.	3161	677	678
Emotion labels	4820	4516	Joyful, lyrical, tense, soothing, etc.	3161	677	678
Aligned multimodal samples	4510	4200	Joint samples of audio, MIDI, lyrics, and labels	2940	630	630

4.2 Comparative experiment and composition generation quality assessment

In order to verify the comprehensive performance of the proposed model in intelligent composition tasks, LSTM-Composer, MusicGAN and Transformer-Composer are selected as comparison methods, and experiments are carried out under the same data division, unified training rounds and consistent evaluation standards. The evaluation results are normalized by the percentage system, and the five indexes are mainly investigated: melody coherence, rhythm rationality, harmony consistency, style matching and emotional accuracy. The comparison results show that the traditional cycle model has a certain stability in melody continuity, but has a general performance in harmony organization and style maintenance. The generated results of GAN-like models are more variable, but prone to loose local structure and emotional expression fluctuation. The Transformer model is superior to the previous two models in long-range dependency modeling, but it still has the problem that the style boundary is not clear enough when the multi-modal condition constraints are insufficient. In contrast, with the help of cross-modal fusion, dual-condition control and post-processing verification mechanism, the proposed model shows better balance in the five indicators, which not only improves the overall coordination degree of melody and rhythm, but also enhances the consistency of generated results in style recognition and emotional expression. The comparison results of composition generation quality of different models are shown in Figure 4.

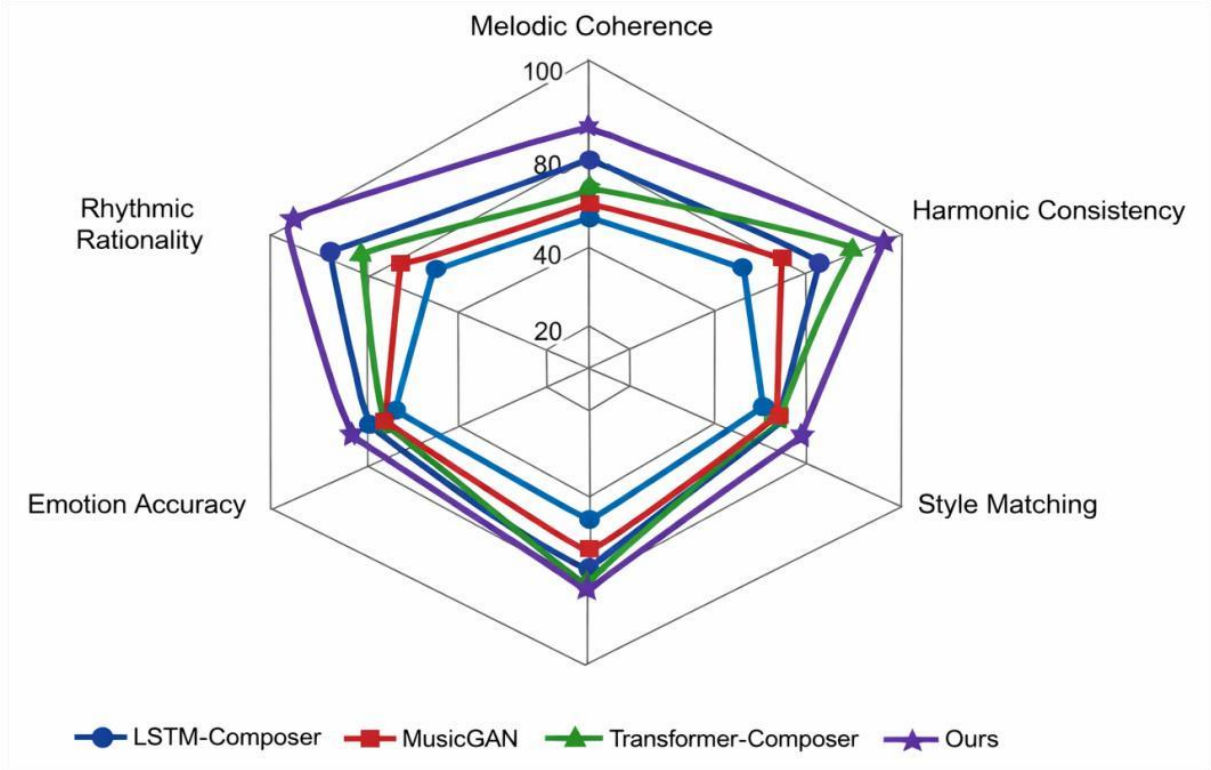

Figure 4: Radar plots of composition generation quality for different models

It can be seen from Figure 4 that the proposed model is at the highest level in five indicators, and the melody coherence reaches 91 points, which is 7 points higher than that of Transformer-Composer. The style matching degree reached 90 points, which was 19 points higher than that of MusicGAN. The emotional accuracy reaches 87 points, which is 21 points higher than that of LSTM-Composer, indicating that the proposed method can more stably balance the quality of music structure and emotional expression effect under the guidance of multi-modal

conditions.

4.3 Confidence interval of comprehensive quality score and stability analysis of results

In order to evaluate the interval distribution characteristics and result stability of different models in composition generation quality, this paper conducts five independent repeated experiments on LSTM-Composer, MusicGAN, Transformer-Composer and the proposed model. The 95% confidence intervals are also estimated based on 2000 Bootstrap resampling. The statistical analysis takes the comprehensive quality score as the core index, and at the same time combines the melody coherence, harmony consistency, style matching and emotional accuracy for auxiliary judgment. The results show that the mean performance and interval stability of the proposed model are better than those of the comparison methods, and the mean comprehensive quality score of the proposed model is 88.4, and the 95% confidence interval is [86.9, 89.9]. Transformer-Composer with mean 79.6, range [77.8, 81.4]; The mean values of MusicGAN and LSTM-Composer are 72.8 and 71.2, respectively. Compared with the strongest baseline Transformer-Composer, the comprehensive score of the proposed model is improved by 8.8 points, and the confidence intervals are basically non-overlapping, indicating that the advantages of the model have strong stability. In this section, error bar plots with 95% confidence intervals are used to show the interval distribution and stability differences of the comprehensive quality scores of different models. Figure 5 shows the error bar graph of key indicators.

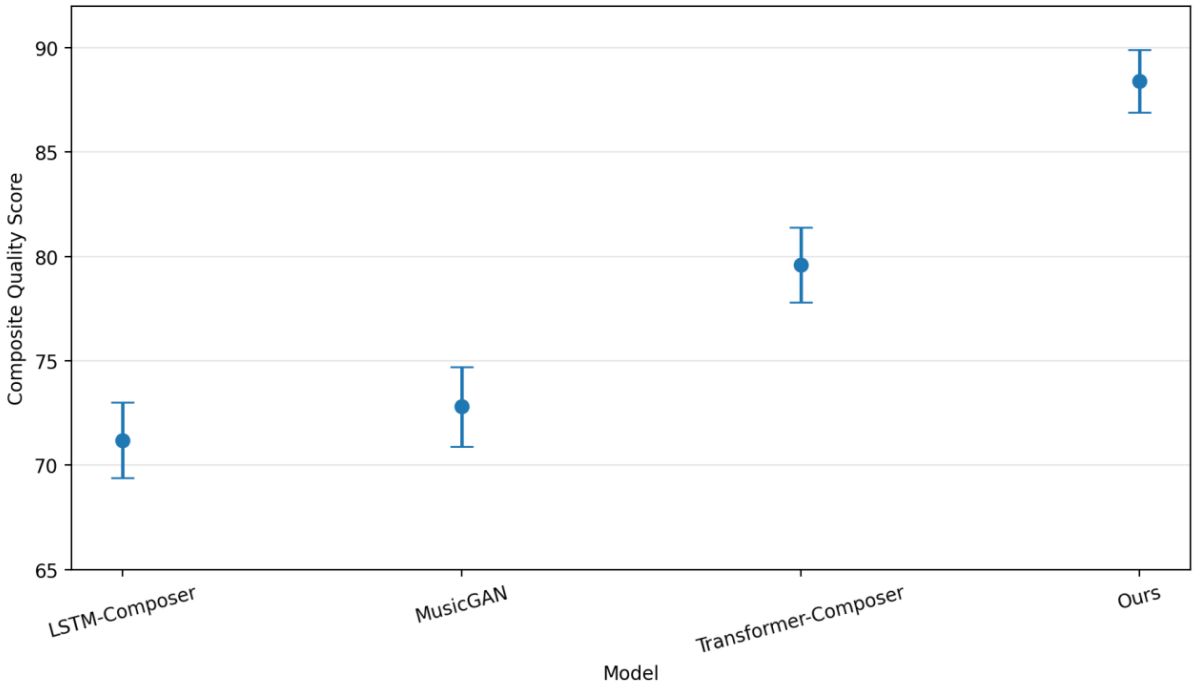

Figure 5: Error bar plots of key indicators

Figure 5 shows that the proposed model not only has the highest mean value, but also has the most concentrated error interval, indicating that multimodal fusion, dual-condition control and post-processing verification jointly improve the stability of model output. Among them, the comprehensive quality score of the proposed model is 8.8 points higher than that of Transformer-Composer, 15.6 points higher than that of MusicGAN, and 17.2 points higher than

that of LSTM-Composer, showing a stable interval advantage.

4.4 Analysis of model robustness

In order to investigate the stability of the model in a complex generation environment, this paper sets up a step-by-step perturbation experiment to simulate common uncertainties in real applications, such as input noise, missing lyrics, style label offset and local MIDI event loss, and uses the perturbation strength of 0%, 10%, 20%, 30% and 40% as the test gradient. The comprehensive generation quality of different models is compared. The results show that with the enhancement of disturbance, the scores of each model show a downward trend, but the decline of the proposed model is significantly slower, indicating that it has stronger stability in multi-modal collaborative modeling and conditional constraint maintenance. Specifically, under the condition of no disturbance, the comprehensive score of the proposed model is 88.4. When the perturbation intensity is increased to 20%, the score remains at 84.7 with a decrease of only 3.7 points, while Transformer-Composer decreases from 79.6 to 74.1 with a decrease of 5.5 points. MusicGAN and LSTM-Composer drop 6.0 and 6.1 points, respectively. When the perturbation is further increased to 40%, the proposed model still reaches 78.9, which is higher than 66.7 of Transformer-Composer, 59.6 of MusicGAN, and 57.8 of LSTM-Composer. This shows that the proposed model effectively alleviates the chain effect of single modal distortion on the overall generation results through cross-modal fusion, dual-condition control and post-processing verification mechanism, so that the melody, rhythm and style expression can still maintain high consistency under disturbance conditions. Figure 6 shows the change curve of model robustness under different disturbance intensities.

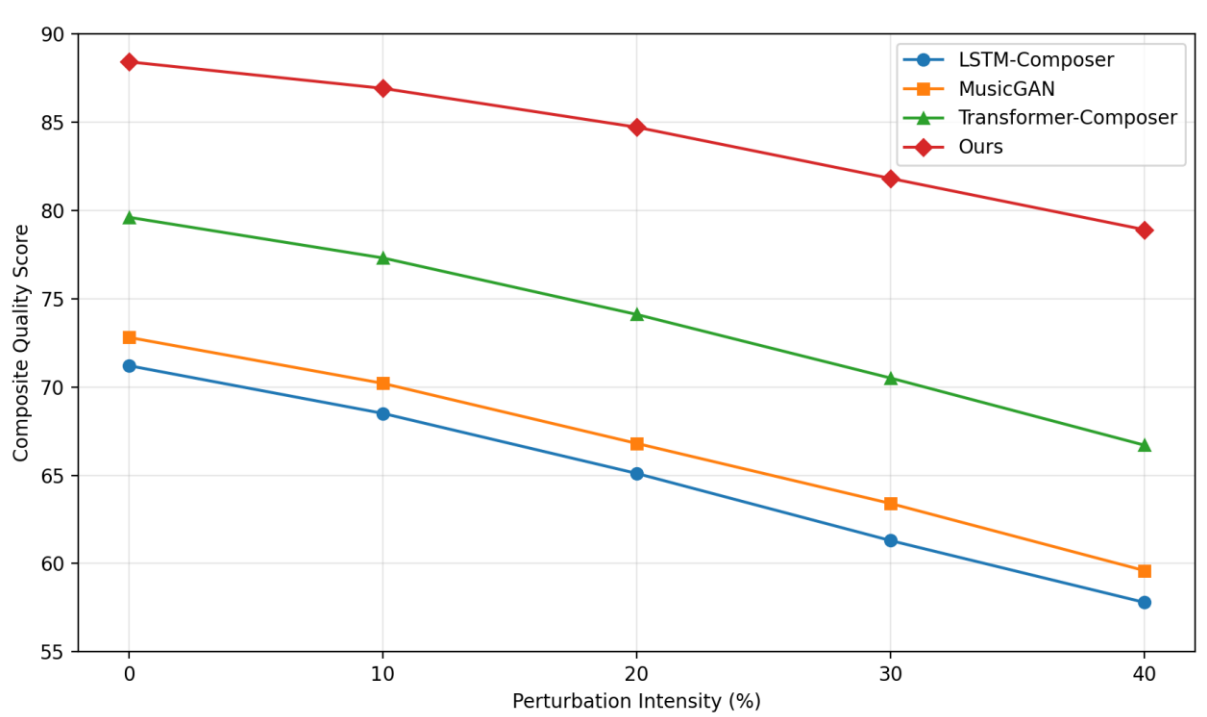

Figure 6: Graph of model robustness variation

It can be seen from Figure 6 that the proposed model maintains the highest score in all perturbation intervals. Under 40% perturbation, it is still 12.2 points higher than that of Transformer Composer, 19.3 points higher than that of MusicGAN, and 21.1 points higher than that of LSTM-Composer. It shows that the proposed method still has good robustness and

availability in scenarios with incomplete or disturbed multimodal information.

4.5 Scalability and cross-scenario generalization analysis

In order to further verify the adaptability of the proposed model in different application environments, this paper selects four scenes including short video soundtrack, game background music, lyric assisted composition and teaching melody generation to carry out generalization tests, and only adjusts the input prompt length and generation time Settings while keeping the main model parameters unchanged. The results show that the proposed model maintains high output quality in different tasks, indicating that the multimodal feature representation and the dual-condition control mechanism have good transfer ability. From the perspective of generation efficiency, with the increase of scene complexity, the average generation time is still in an acceptable range, although it increases to a certain extent. From the perspective of content quality, the model has small fluctuations in style retention rate and emotional consistency, showing a stable cross-scene adaptation ability. In the lyric-assisted composition scenario, the comprehensive score of the model reaches 89.1, and the style preservation rate is 91.4%. In the game background music scene, the average generation time is controlled at 1.42 s, and the comprehensive quality of more than 85 points is still maintained. Table 5 shows the comparison results of model generalization performance in different scenarios.

Table 5: Comparison results of model generalization performance in different scenarios

Application Scenario	Data Characteristics	Overall Quality Score	Style Retention Rate / %	Emotion Consistency / %	Average Generation Time / s
Short-video soundtrack generation	Distinct rhythm and relatively short duration	87.6	89.8	86.9	1.08
Game background music generation	Strong cyclicity and continuous atmosphere	85.9	88.3	84.7	1.42
Lyric-assisted composition	Clear semantic guidance and strong emotional constraints	89.1	91.4	88.6	1.27
Educational melody generation	Regular structure and high melodic readability requirements	86.8	87.9	85.5	0.96

It can be seen from Table 5 that the comprehensive quality scores of the proposed model in the four types of scenes remain above 85.9, with the highest reaching 89.1, and the average generation time is controlled within 1.5 s, indicating that it has good application potential in terms of scalability and cross-scene generalization.

4.6 System implementation and configuration of software and hardware environment

In order to ensure that the intelligent composition model driven by multi-modal music features

can stably complete training, reasoning and result output, this paper completes the system deployment on a unified experimental platform, and constructs a complete implementation process around data preprocessing, model training, condition control, result decoding and music export. The system uses Python 3.10 as the development language and runs under Ubuntu 22.04 LTS environment. The deep learning part is implemented based on PyTorch 2.3, and CUDA 12.1 is used to accelerate model training and parallel computing. In terms of hardware configuration, the experimental platform is equipped with Intel Xeon Silver 4314 processor, NVIDIA RTX 4090 24 GB graphics card, 128 GB memory and 2 TB SSD, which can meet the computational requirements of multi-modal feature extraction, long sequence modeling and batch generation tasks. PrettyMIDI 0.2.10 is used for MIDI event parsing and reconstruction. Transformers 4.41 is used for semantic encoding of lyrics. In the training process, the batch size was set to 32, the initial learning rate was set to 0.0005, the optimizer was selected AdamW, and the iteration rounds were uniformly set to 120. At the same time, the random seed was fixed and the log information was recorded to reduce the fluctuation of the experiment and ensure the reproducibility of the results. Finally, the system can output MIDI, MusicXML, score image and other results, which can better support various needs such as model training, generation and verification, and application display.

4.7 Ablation experiment and contribution analysis of key modules

In order to further verify the actual contribution of each key module to the generated results, this paper sets up three groups of ablation schemes based on the complete model, removing the cross-modal fusion module, the dual-condition control module, and the post-processing and music verification module respectively, and carries out comparative tests under the same data set, the same training rounds and consistent evaluation standards. The ablation experiment focuses on four indicators: comprehensive quality score, harmony consistency, style matching degree and emotional accuracy. The results show that the cross-modal fusion module has the most direct impact on the overall composition quality. After removal, the model is difficult to fully integrate the complementary information between audio, MIDI and text, resulting in a significant decline in melody organization and harmonic cohesion. The dual-condition control module mainly affects the clarity of style boundaries and the stability of emotional expression. Although the generated results still maintain the basic structure after removal, the personality and emotional orientation of the works are significantly weakened. The post-processing and score verification module plays an important role in output normalization and local detail correction, and the beat closure and sentence end stability decrease after the absence of the module. The comparison results of model performance under different ablation configurations are shown in Table 6.

Table 6: Comparison results of model performance under different ablation configurations

Model Configuration	Overall Quality Score	Harmonic Consistency	Style Matching Degree	Emotion Accuracy
Without the cross-modal fusion module	80.1	77.4	79.2	78.5
Without the dual-condition control module	82.7	80.3	81.6	79.8
Without the post-processing and score validation module	84.3	82.1	86.4	83.2
Full model	88.4	86.0	90.0	87.0

It can be seen from Table 6 that the full model achieves optimal results on all four indicators. Among them, compared with the removal of the cross-modal fusion module, the comprehensive quality score of the complete model is improved by 8.3 points, and the harmony consistency is improved by 8.6 points. Compared with removing the dual-condition control module, the style matching degree is increased by 8.4 points, and the emotional accuracy is increased by 7.2 points. Compared with the post-removal processing and score verification module, the comprehensive quality score is still increased by 4.1 points, indicating that the above three modules play an irreplaceable role in the composition generation process.

4.8 Discussion

Combined with the above experimental results, it can be seen that the modeling method driven by multi-modal music features can better balance the structural integrity and expression refinement compared with the composition method relying on a single sequence input. On the one hand, audio, MIDI and lyric text have obvious complementarity at the information level: audio features are closer to the real listening sense, MIDI sequences are more conducive to describing pitch, duration and harmony, and text and label information provide external semantic support for style control and emotional guidance. Because these types of information form joint constraints in the generation process, the model shows more stable advantages in melody coherence, style matching and emotional accuracy. On the other hand, the dual-condition control mechanism does not stay at the simple label superposition level, but continuously instills style information and emotional information into the decoding process through hidden state modulation, which makes the generated results not only "can hear", but more similar to the works with "explicit expression intention".

From the experimental phenomenon, the cross-modal fusion module is the most key to the overall quality improvement, which solves the problems of inconsistent timing and semantic granularity between different modalities. Although the post-processing and score verification modules are not directly involved in the main body generation, they significantly improve the beat closure, sentence end stability and output normalization, which indicates that the structure correction after generation should not be ignored if the intelligent composition system is to have real application value. The robustness experiments and cross-scenario tests also show that the proposed model can still maintain high quality when the input is disturbed or the application scenario changes, which means that the proposed method is not only suitable for standard experimental environments, but also has certain potential for practical deployment.

Of course, there is still room for further improvement of the research in this paper. Although the current experimental sample size can support model training and comparative analysis, it is still not sufficient in the subdivision level of music style, the coverage of complex harmonic texture and the generation of long-term structure. Although the human evaluation part introduces the evaluators with music learning background, the subjective feeling itself still has some fluctuations. Subsequent research can continue to expand on the basis of larger scale, multi-style and multi-language lyrics data, and introduce more fine-grained paragraph control, orchestration generation and human-computer collaborative creation mechanisms, so that the model can further enhance the editability and artistic expression of works while maintaining the quality of generation.

5 Conclusion

Focusing on the problem of intelligent composition driven by multi-modal music features, this paper completes the model design, system implementation and experimental verification. By

constructing a unified representation and cross-modal fusion mechanism, this study integrates heterogeneous information such as audio, MIDI and lyrics into the same generation framework, and combines hierarchical generation, melody rhythm and harmony synergy constraints, dual-condition control of style and emotion, and post-processing verification to improve the overall quality and output standardization of the generated results. Experiments show that the comprehensive quality score of the complete model reaches 88.4, which is 8.8 points higher than that of Transformer-Composer. The score of 78.9 is still maintained under 40% disturbance, which reflects good robustness. In scenes such as short video soundtrack, game background music and lyric assisted composition, the comprehensive quality score remains above 85.9, showing strong generalization ability. In general, the proposed method has achieved relatively stable results in music structure organization, style preservation and emotional expression, but there is still room for improvement in complex harmony texture, long-term structure generation and large-scale multi-style data adaptation. In the future, it can continue to be expanded in the direction of orchestration generation, human-computer collaborative creation and finer-grained control.

References

[1] Liu L, Gong R, Yang Y.MusDiff: A multimodal-guided framework for music generation [J]. Alexandria Engineering Journal, 2025, 129(000):128-136. DOI:10.1016/j.aej.2025.05.053.

[2] Zhang Y, Yu S.Harmonizing AI: A GAN–Transformer fusion for expressive multimodal music synthesis in IoT systems[J]. Alexandria Engineering Journal, 2025, 131(c): 368-382. DOI:10.1016/j.aej.2025.07.043.

[3] Wang S.Music Emotion Recognition and Modeling Based on Multimodal Signal Fusion[J].Traitement du Signal, 2025, 42(4).DOI:10.18280/ts.420446.

[4] Zhu L, Zhou F, Wang S,et al.A language-guided cross-modal semantic fusion retrieval method[J].Signal Processing, 2025:234.DOI:10.1016/j.sigpro.2025.109993.

[5] Chen W, Wu G.A Multimodal Convolutional Neural Network Model for the Analysis of Music Genre on Children's Emotions Influence Intelligence[J]. Computational intelligence and neuroscience, 2022, 2022:5611456.DOI:10.1155/2022/5611456.

[6] Shi X, Li X, Toda T.Multimodal Fusion of Music Theory-Inspired and Self-Supervised Representations for Improved Emotion Recognition[J].Interspeech 2024, 2024: 3724-3728. DOI:10.21437/interspeech.2024-2350.

[7] Zhang Y.An IoT-enhanced automatic music composition system integrating audio-visual learning with transformer and SketchVAE[J].Alexandria Engineering Journal, 2025, 113(000):378-390.DOI:10.1016/j.aej.2024.10.115.

[8] Hao X, Li H, Wen Y.Real-time music emotion recognition based on multimodal fusion[J].Alexandria Engineering Journal, 2025, 116(000):586-600.DOI:10.1016/j.aej.2024.12.060.

[9] Zhang Y.TS-Resformer: a model based on multimodal fusion for the classification of music signals[J]. FRONTIERS IN NEUROROBOTICS, 2025, 19(000).DOI:10.3389/

fnbot.2025.1568811.

[10] Pan Z, Jiang S, Yang X,et al.Hierarchical Cross-Modal Image Generation for Multimodal Biometric Recognition With Missing Modality[J].Information Forensics and Security, IEEE Transactions on, 2025, 20(000):4308-4321. DOI:10.1109/TIFS. 2025.3559802.

[11] Liu Z.AI-driven classification and trend analysis of piano music genres using large language models[J]. International Journal of Information and Communication Technology, 2025, 26(12):32-48.DOI:10.1504/IJICT.2025.146164.

[12] Song J.Constructing a Multimodal Music Teaching Model in College by Integrating Emotions[J].Applied Mathematics and Nonlinear Sciences, 2024, 9(1). DOI:10.2478/amns-2024-1202.

[13] Wu M.Music Emotion Classification Model Based on Multi Feature Image Fusion[J]. 2024 First International Conference on Software, Systems and Information Technology (SSITCON), 2024:1-6.DOI:10.1109/ssitcon62437.2024.10796069.

[14] Li S.Multimodal Data Fusion Method for Music Education Evaluation: Comprehensive Application of Image Processing and Audio Analysis[J]. 2023 International Conference on Intelligent Computing, Communication & Convergence (ICI3C), 2023:167-172. DOI:10.1109/ici3c60830.2023.00042.

[15] Liu Z, Liu X, Chen S,et al.Multimodal Fusion for Talking Face Generation Utilizing Speech-Related Facial Action Units[J].ACM Transactions on Multimedia Computing, Communications, and Applications, 2024, 20(9):1-24.DOI:10.1145/3672565.

[16] Pandeya Y R, Lee J.Deep learning-based late fusion of multimodal information for emotion classification of music video[J].Multimedia Tools and Applications, 2021, 80(38):1-19.DOI:10.1007/s11042-020-08836-3.

[17] Hu Y.Music Emotion Research Based on Reinforcement Learning and Multimodal Information[J]. JOURNAL OF MECHANICS, 2022, 2022(000):9. DOI:10.1155/2022/2446399.

[18] Yu Q, Lin Y.Research on Emotion Recognition Algorithm of Piano Music Based on Multimodal Learning[J].Advances in Computer and Engineering Technology Research, 2024, 1(4):250.DOI:10.61935/acetr.4.1.2024.p250.

[19] Sable R Y, Sayyed A, Kalyane B,et al.Enhancing Music Mood Recognition with LLMs and Audio Signal Processing: A Multimodal Approach[J].International Journal for Research in Applied Science and Engineering Technology, 2024, 12(7):628-642. DOI:10.22214/ijraset.2024.63590.

[20] Gu X, Ou L, Zeng W,et al.Automatic Lyric Transcription and Automatic Music Transcription from Multimodal Singing[J].ACM Transactions on Multimedia Computing, Communications, and Applications, 2024, 20(7).DOI:10.1145/3651310.