Traitement du Signal
Vol. 42, No. 4, August, 2025, pp. 2389-2399
Journal homepage: http://iieta.org/journals/ts

Music Emotion Recognition and Modeling Based on Multimodal Signal Fusion

Shan Wang

School of General Education, Shanghai Technology and Innovation Vocational College, Shanghai 201620, China

Corresponding Author Email: ws03130717@163.com

https://doi.org/10.18280/ts.420446

Received: 18 December 2024
Revised: 27 May 2025
Accepted: 2 July 2025
Available online: 14 August 2025

Keywords:
multimodal signal fusion, music emotion recognition, feature extraction, cross-modal interaction, emotion modeling

ABSTRACT

With the rapid development of the digital music industry, a vast amount of music resources has emerged, and the demand for accurate music emotion matching is becoming increasingly urgent. The transmission of music emotion involves multimodal information, such as audio and text. Single-modal recognition, due to its inability to fully capture the emotional nuances, has limitations, and multimodal signal fusion has become the key to achieving a more comprehensive recognition of music emotion. In current research on music emotion recognition, some methods rely on single modalities, such as audio feature-based recognition, which cannot interpret the deeper emotions in lyrics, resulting in low recognition accuracy for lyrical music. Some multimodal fusion studies use early feature concatenation or simple weighted strategies, failing to establish dynamic relationships between modalities. As a result, recognition errors are significant in cross-modal conflict scenarios, and robustness to cross-modal noise is insufficient. Against this backdrop, researching music emotion recognition and modeling based on multimodal signal fusion is of great significance. This study proposes a multimodal signal fusion-based music emotion recognition model, which makes breakthroughs through four core modules: in the feature extraction phase, improved Convolutional Neural Network (CNN) is used to extract emotional features from the audio time-frequency domain, and Bidirectional Long Short-Term Memory (BiLSTM) combined with the attention mechanism captures the semantic emotional tendencies of the text; the cross-modal interaction learning module designs a dynamic attention weight matrix, quantifying the contribution of different modalities in different emotional dimensions based on mutual information entropy; the feature fusion module introduces a cross-modal Transformer, which maps audio temporal features and text semantic features to a unified emotional vector space to address modality heterogeneity; the emotion classification layer uses a multi-output loss function to optimize both discrete emotional categories and continuous emotional dimension predictions. This research aims to improve the accuracy and robustness of music emotion recognition, providing a scalable model architecture and technical standards for multimodal emotion computation.

1. INTRODUCTION

With the widespread popularity of digital music platforms and the lowering of barriers to music creation, music works around the world are growing at an exponential rate, and the daily music information that users are exposed to has surpassed the massive level [1-4]. In this context, relying solely on manual selection or simple label classification can no longer meet users' needs for precise emotional matching. The transmission of music emotion is a complex multi-dimensional process: elements such as pitch fluctuations, rhythm intensity, and timbre in the audio signal form the "auditory skeleton" of emotional expression [5-9], while imagery choices, semantic tendencies, and rhetorical usage in lyrics form the "semantic flesh and blood" of emotional transmission [10-13]. Relying solely on audio signals makes it difficult to capture the deep emotions contained in metaphorical lyrics, and relying only on text information loses the emotional tension carried by melody and rhythm. Therefore, achieving three-dimensional recognition of music emotion through multimodal signal fusion has become the key path to breaking the current technical bottleneck.

Conducting related research has multiple levels of significance. Theoretically, it can reveal the mapping rules and collaborative mechanisms of music emotion across different modalities, providing a new paradigm for cross-modal information processing in the field of emotion computation, and promoting a deeper understanding of human emotional perception and expression. At the application level, in addition to music recommendation and psychological therapy, it can also empower intelligent music education by optimizing teaching strategies based on recognizing learners' emotional feedback on music works; assist in film and television advertisement music composition, automatically matching the most suitable music clips according to the emotional requirements of the scene; and even assist in music creation, generating works with both auditory and semantic consistency based on users' preset emotional goals.

Existing research in the field of music emotion recognition has significant limitations. References [14-16] use single-modal modeling, such as classifying emotions based only on audio features. These methods cannot interpret emotional expressions such as "sad autumn" or "happy rain" in lyrics, leading to recognition accuracy for lyrical music generally being lower than 60%. Another attempt, such as the research in references [17-20] on multimodal fusion, mostly uses early feature concatenation or simple weighted fusion strategies, failing to build dynamic relationships between modalities. When the audio is happy but the lyrics are sad, the fusion result often leads to emotional deviation, with the average recognition error exceeding 25% compared to the ideal state. Moreover, references [21-25] show insufficient robustness to cross-modal noise, such as background noise in audio or ambiguous words in text, which distorts the information in the feature extraction stage, further reducing recognition stability.

This paper addresses the above issues by constructing a music emotion recognition model based on multimodal signal fusion, with core innovations in the collaborative design of four modules: in the feature extraction phase, an improved CNN is used to extract time-frequency domain emotional features for audio, while a BiLSTM combined with the attention mechanism captures semantic emotional tendencies in text, achieving fine-grained feature representation; the cross-modal interaction learning module introduces a modality attention weight matrix, dynamically calculating the mutual information entropy between audio and text features, quantifying their contribution in different emotional dimensions; the feature fusion module adopts a cross-modal Transformer structure, transforming the temporal features of audio and the semantic features of text into a unified emotional vector space, addressing modality heterogeneity; the emotion classification layer designs a multi-output loss function to simultaneously optimize the prediction accuracy of both discrete emotional categories and continuous emotional dimensions. This study aims to significantly improve the accuracy of emotion recognition by constructing a deep cross-modal fusion framework, substantially reducing recognition errors in cross-modal conflict scenarios, and providing a reusable model architecture and technical standards for multimodal emotion computation.

2. MUSIC EMOTION RECOGNITION AND MODELING BASED ON MULTIMODAL SIGNAL FUSION

The use of DeBERTa for extracting text features and ResNet50 for extracting audio features is a targeted choice made in this paper based on the inherent characteristics of text and audio information in music emotion recognition and the practical scene requirements. In terms of text, music-related texts such as lyrics and reviews often contain rich emotional metaphors and semantic layers. For example, the word “rain” in the lyrics may symbolize loneliness or represent cleansing, with emotional ambiguity due to polysemy, and there are significant differences in textual expression across different music styles. As a pre-trained language model, DeBERTa can deeply analyze contextual semantic associations in the text through dynamic masking and enhanced positional encoding, precisely capturing emotional tendencies in complex expressions like "using scenery to metaphor emotions," addressing the shortcomings of traditional text feature extraction methods in semantic understanding depth, and meeting the actual demand for delicate and variable emotional expression in music texts. In terms of audio, emotional information in music is hidden in the time-frequency changes of the spectrogram, such as the brightness of major scale audio versus the melancholy of minor scale, the excitement of fast tempos versus the relaxation of slow tempos. These features require the model to have the ability to extract deep-level hierarchical features. ResNet50, with its residual connections, effectively alleviates the gradient vanishing problem in deep networks, and can extract features from basic frequencies to higher emotional features from audio spectrograms, especially resistant to interference from environmental noise in live versions of music, adapting to the characteristic of audio signals being easily contaminated by noise in practical scenes.

This paper further introduces the self-attention mechanism and multi-head cross-attention mechanism to address the core issues of feature fragmentation and weak modality correlations in music emotion recognition, aligning with the high demands for accuracy and robustness in emotional recognition for practical applications. In real scenarios, the emotional expression of music often presents a “fragmented feature” state: the audio features of a song may be calm in the verse and intense in the chorus, and the text features may mix unrelated narratives with core emotional sentences, diluting the effective emotional features. The self-attention mechanism, by calculating the association weights between features, can automatically focus on the most emotionally distinguishing spectrogram segments in the audio and the key emotional words in the text, enhancing the representation strength of critical emotional features, thus solving the problem of emotional expression ambiguity and feature distribution dispersion. The multi-head cross-attention mechanism is designed to meet the collaborative needs of multimodal information. In practical applications, music emotion is often a joint expression of audio and text. For example, sad lyrics combined with a somber melody will reinforce the sadness, but there are also conflicting scenarios such as “happy melody + sad lyrics.” This mechanism, through multiple attention heads, learns different correlation patterns between modalities in parallel. It allows the rhythm intensity of audio and the emotional vocabulary in text to complement and verify each other, and in the case of modality conflicts, adjusts the weights to calibrate the information, ensuring that even in complex emotional scenes, the model can efficiently mine cross-modal emotional associations, ultimately improving the accuracy of emotional classification and meeting the strict requirements for emotional recognition in music recommendation, psychological therapy, and other scenarios.

Specifically, the proposed multimodal emotion analysis model consists of four parts: audio and text feature extraction, cross-modal interaction learning module, audio and text emotion feature fusion module, and emotion classification. The audio and text feature extraction module and the cross-modal interaction learning module form the foundational layer of the model for understanding multimodal music data, with their operation principles closely aligned with the expressive characteristics of music emotion in different modalities. The multi-head cross-attention in the cross-modal interaction learning module, designed for the actual situation where music emotion is often expressed jointly or in conflict by audio and text, uses multiple attention heads to focus on different dimensions of the correlation between hidden states of different modalities. For instance, one attention head focuses on the emotional word “crying” in the lyrics and its correspondence with the low tone in the audio, while another focuses on the matching of words like “laughing” with the bright rhythm. It even captures the potential emotional logic in conflict scenes like “happy melody + sad lyrics,” finely integrating the interaction information and enabling the model to initially understand the emotional associations between multimodal data. The audio and text emotion feature fusion module and emotion classification module form the core layer of the model for refining and outputting emotional information, directly serving the accurate capture of the overall emotion of music. In practice, music emotion is often scattered across different sections, such as the calm narrative in the verse and the emotional outburst in the chorus, and there is a deep semantic connection between the emotional associations of audio and text. The self-attention mechanism calculates the association weights of features across the entire audio and text, automatically focusing on the intense audio segments in the chorus and the key emotional sentences in the lyrics, solving the issue of dispersed emotional features. The Transformer encoder processes the features in parallel at multiple levels, learning feature representations at different levels of abstraction. For example, at a lower level, it associates the word “love” in the lyrics with a gentle melody in the audio, and at a higher level, it distills the overall emotion of “sweetness,” deeply capturing the cross-modal semantic connections. The fused features are then input into the emotion classification module, which outputs precise emotional classification results, combining the need for music emotion recognition to clearly distinguish categories like “joy” and “sadness” in recommendation systems, or quantifying pleasantness and arousal in psychological therapy, realizing the transformation from multimodal data to clear emotional labels.

2.1 Text and audio feature extraction

Figure 1 shows the schematic of text and audio feature extraction. The text feature extraction module uses the *DeBERTa* model, whose core principle is to accurately capture the complex emotional connotations in music texts through global attention mechanisms and deep contextual semantic analysis, in order to adapt to the emotional expression characteristics of text information such as lyrics in the music scene. In real music scenarios, lyrics often contain a large number of rhetorical devices such as metaphors and puns. For example, the word "falling leaves" might symbolize the melancholy of passing time in folk music, or it could represent the resoluteness of breaking free from constraints in rock music. The emotional tendency of the same word varies significantly depending on the context. As an improved version of *BERT*, *DeBERTa* uses dynamic masking technology to randomly mask words in the input text and predict them, forcing the model to deeply learn contextual associations. Meanwhile, enhanced positional encoding can accurately distinguish the grammatical position and semantic weight of words in the sentence. For example, the word "love" in "I love you" and "You love me" is the same word but conveys different emotional meanings due to its position in the sentence. This design allows it to effectively address the ambiguity and polysemy of emotional expression in lyrics, providing high-purity text emotional features for subsequent cross-modal fusion and meeting the need for recognizing complex emotions such as "a theme of heartbreak but with cheerful lyrics" in music recommendations.

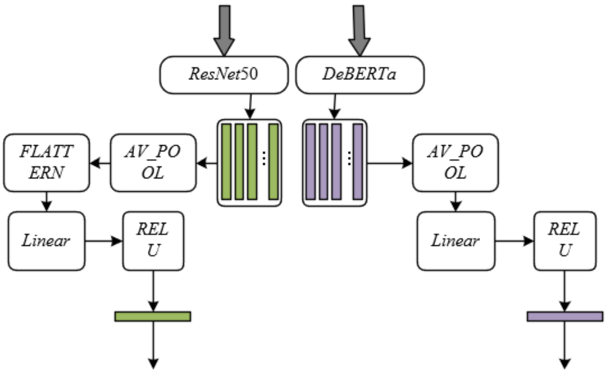

Figure 1. Schematic of text and audio feature extraction

The specific computational process of text feature extraction focuses closely on generating high-dimensional emotional features to adapt to the sequential characteristics and emotional complexity of music texts. For the input text information S= {s_1, s_2,..., s_v}, each word s_u is first converted into a fixed-dimensional word vector through the embedding layer, while enhanced positional encoding is incorporated to mark the relative position relationships of words in the sequence. Then, the text sequence enters an encoder composed of multiple layers of Transformer, where the global attention mechanism in each layer calculates the association weight of each word with all other words. For example, in the lyrics "The rain falls all night, my love overflows like the rain," the attention weight between "rain" and "love" will be significantly higher than that of other words, highlighting the core emotional connection. The dynamic masking mechanism continuously randomly masks part of the words, and through the model's prediction error for the masked words, the parameters are optimized in reverse, strengthening the understanding of contextual semantics. Finally, after passing through multiple layers of encoding, the text sequence is pooled to generate a fixed-dimensional high-dimensional text feature vector. This vector not only contains explicit emotional tendencies such as "positive," "neutral," and "negative," but also contains implicit features of more subtle emotions like "missing" or "letting go," providing a rich semantic foundation for subsequent cross-modal interactions. Specifically, define each input sample as (U, S), where the text information S contains v words, and the audio information is $U∈R^{G×Q×Z}$. Assume that the hidden state feature information of each token in the text sequence is represented by $G_{TE}∈R^{V×F}$, the entire text sequence feature information is represented by $D_S∈R^{V×C}$, the length of the text sequence is V, and the hidden state information of the text sequence and the dimensionality of the entire text sequence are represented by F and C, respectively. The trainable initialization weights and biases in the DeBERTa model are represented by q_0 and y_0, and the average pooling function is represented by *AV_POOL*, with the activation function represented by *RELU*. The computational process expression for text feature extraction is:

$$G_{TE} = DeBERTa(S) \tag{1}$$

$$D_S = RELU\left(q_0\left(AV\_POOL\left(G_{TE}\right)\right)+y_0\right) \tag{2}$$

The audio feature extraction module uses the *ResNet50* model, which is based on the synergy of deep CNN and residual connections, to extract emotion-related features layer by layer from the audio spectrogram in order to deal with the diversity and complexity of music audio in practical scenes. Emotional information in music audio is hidden in the dynamic changes of the time-frequency domain, such as the high-frequency harmonics of the violin conveying a melodious feeling in classical music, or the low-frequency pulses of electronic music conveying an exciting feeling. Additionally, live-recorded audio often contains applause, noise, and other interferences. *ResNet50* converts the audio signal into a Mel spectrogram $R^{G\times Q\times Z}$, and extracts features through multiple convolutional layers: shallow convolutions capture low-level features such as edges and textures in the spectrogram, while deep convolutions integrate these low-level features to form high-level emotional features such as "joyful melody" or "somber tone." Crucially, residual connections bypass certain convolution layers and directly pass features, effectively addressing the gradient vanishing problem in deep networks, enabling stable processing of song audio as long as 5 minutes or chorus segments lasting 30 seconds, adapting to music scenes with different lengths and styles, and providing robust and distinguishable audio emotional features for cross-modal fusion.

The specific computational process of audio feature extraction aims to accurately capture emotion-related audio features, achieving effective encoding of complex audio information through convolution operations and feature aggregation. The input audio spectrogram $U \in R^{G\times Q\times Z}$ first passes through an initial convolutional layer, where a 3×3 convolution kernel is used to perform sliding window calculations, extracting the frequency changes and temporal continuity features of the local regions in the spectrogram, such as sudden increases or continuous decreases in certain frequency bands. Then, the feature map enters a stacked structure consisting of multiple residual blocks. Each residual block contains two convolution layers and one shortcut connection: convolution layers further extract multi-scale features using convolution kernels of different sizes, such as 1×1 convolution to compress the channel dimension and reduce computational load, and 3×3 convolution to capture broader frequency associations. The shortcut connection adds the input features to the convolution output, retaining the original information while adding new features, thus avoiding feature degradation in deep networks. After processing by multiple residual blocks, the feature map is converted into a fixed-dimensional vector via a global average pooling layer. This vector integrates multi-dimensional features such as rhythm intensity, timbre brightness, and emotional tension in the audio, and can complement text features. For example, when lyrics ambiguously express "excitement," the audio's rapid rhythm and high-pitched frequency can reinforce this emotional determination, improving the model's accuracy in music emotion classification. Specifically, assume that the hidden state feature information of the audio sequence is represented by $G_{IM} \in R^{M\times F}$, the feature information of the entire audio sequence is represented by $D_U \in R^{M\times C}$, the length of the audio sequence is M, and the hidden state information and the dimensionality of the entire audio sequence are represented by F and C, respectively. The trainable initialization weights and biases in the *ResNet50* model are represented by q_1 and y_1, and the flatten operation is represented by the *FLATTEN* function. The computational process of audio feature extraction is as follows:

$$G_{IM} = RESNET50(U) \tag{3}$$

$$\begin{aligned} D_U &= G_{TE} = DeBERTa(S) \\ RELU&\begin{pmatrix} q_1\left(FLATTEN\left(AV\_POOL\left(G_{IM}\right)\right)\right) \\ +y_1 \end{pmatrix} \end{aligned} \tag{4}$$

2.2 Cross-modal interaction learning

Figure 2 shows the schematic of the cross-modal interaction learning principle. The cross-modal interaction learning module adopts the core principle of multi-head cross attention, aiming to break through the limitations of traditional multimodal fusion methods that overlook local associations and fine-grained interactions, accurately capturing emotional associations between audio and text at different levels and dimensions in music, thus adapting to the complexity and dynamics of music emotion expression. In practical music scenarios, the emotional interaction between audio and text is not a simple one-to-one correspondence at the overall level, but rather hides a large number of local, fine-grained associations: for example, in the verse of a song, the lyrics may tell a mundane story, but the low chords of the piano accompaniment may hint at underlying sadness; in the chorus, the word "shouting" in the lyrics may resonate with the high-pitched distorted sound of the electric guitar. Traditional methods directly fuse the entire text sequence and audio features, which may mask such local interactions and lead to misjudgments of the overall emotion. The multi-head cross attention mechanism, by using multiple attention heads in parallel to focus on different local regions of interaction, can separately capture strong associations between the chorus text and audio, weak associations between the verse text and audio, and even dynamic emotional mappings of the same text segment with different audio sections, providing richer associative information for subsequent fusion, thus meeting the need for adapting to complex emotional scenes in music emotion recognition.

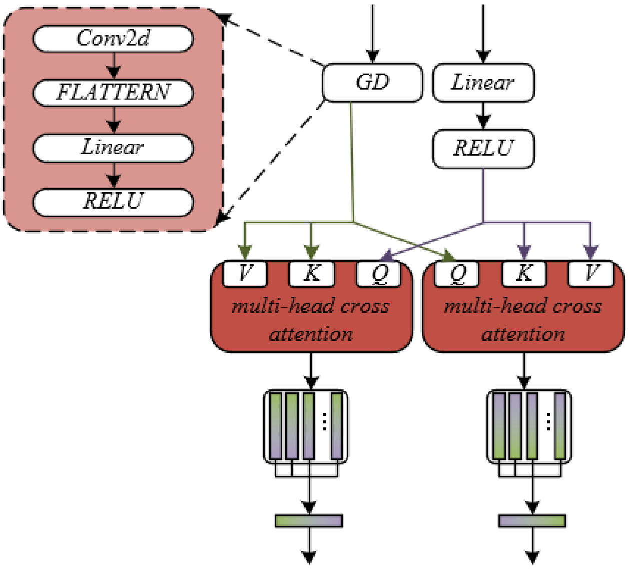

Figure 2. Schematic of cross-modal interaction learning

The design of the multi-head cross attention mechanism allows it to analyze cross-modal emotional associations from multiple dimensions, with each attention head focusing on a specific aspect of the interaction between audio and text in music, enabling a deep exploration of potential semantic structures. In music emotion expression, the interaction between text and audio involves multiple dimensions: firstly, the correlation between emotional words and timbre, such as "warm" lyrics often matching the soft timbre of strings; secondly, the synchronization of narrative rhythm and music speed, such as fast-paced rap lyrics requiring a corresponding dense drumbeat; thirdly, the correspondence between semantic transitions and melodic fluctuations, such as "sudden departure" in lyrics often accompanying a sharp drop in melody. The multi-head cross attention mechanism, by assigning different weight parameters to each attention head, enables it to focus on interaction learning in a specific dimension: for example, one head focuses on the correlation between emotional adjectives and the centroid of the audio spectrum, another head focuses on the match between the length of text phrases and the spacing of audio beats, and yet another captures the correspondence between the semantic transitions in text and changes in the audio melody contour. This multi-dimensional focus mechanism enables the model to understand the emotional dialogue between "words and sounds" in music from different perspectives, similar to a human listener, thus avoiding information omissions caused by a single perspective.

The module processes feature by reasonably setting query and key-value sets, enabling deep reciprocal feedback between audio and text emotional information, to adapt to the dynamic influence relationships between modalities in music. In actual music creation, the emotional influence between audio and text is bidirectional: the semantic tendency of text guides the listener's interpretation of the emotional tone of the audio, and the emotional tone of the audio also influences the understanding of the text. The module first processes the text and audio features through activation functions and linear layers to extract more representative local features. This step highlights the semantic weight of emotional words in the text and the significance of emotional features in the audio. Specifically, after processing G_{IM} and G_{TE} through activation functions, linear layers, and other operations, the local text and audio feature information is obtained, assuming the trainable initialization weights and biases are represented by q_2 and y_2. A multi-layer perceptron consisting of *ReLU*, *Linear*, *Flatten*, and *Conv2d* is represented by G_D. The query set of the cross-modal multi-head cross attention mechanism is represented by G_S, and its key-value set is represented by G_U. The specific computational formula is as follows:

$$G_S = RELU\left(q_2 G_{TE} + y_2\right), G_U = G_D\left(G_{IM}\right) \tag{5}$$

Subsequently, the module performs bidirectional cross-attention calculations: first, using text features as queries and audio features as keys, to capture the guiding role of text in interpreting the emotional tone of audio; then using audio features as queries and text features as keys, to capture the emotional attribution of audio to text semantics. Specifically, in the cross-attention mechanism with G heads, each head learns different attention weights and focuses on emotional feature information in the audio and text data from different angles. Let the operations of the cross-attention mechanism and the cross-modal multi-head attention mechanism be represented by *CROSS_ATT* and *MHCA*, respectively, assuming the query, key, and value vectors of dimension f are represented by $W_{Sg}∈R^{V×f}$, $J_{Ug}∈R^{M×f}$, and $N_{Ug}∈R^{M×f}$, respectively. The number of heads in the cross-attention mechanism is represented by g, and the g-th attention head is represented by $HEAD_g$. The dimensions of W and J are represented by f_j, and the weight parameters for linear projection are represented by Q_x. The learned projection matrices are Q_{WSg}, Q_{JUg}, and $Q_{NUg}∈R^{F×f}$, and the specific computational formulas are as follows:

$$W_{Sg} = Q_{WSg} G_S, J_{Ug} = Q_{JUg} G_U, N_{Ug} = Q_{NUg} G_U \tag{6}$$

$$\begin{aligned} & CROSS\_ATT\left(G_S, G_U\right) \\ & = SOFTMAX\left(\frac{W_{Sg} J_{Ug}^S}{\sqrt{f_j}}\right) N_{Ug} \end{aligned} \tag{7}$$

$$HEAD_g = CROSS\_ATT\left(G_S, G_U\right) \tag{8}$$

$$\begin{aligned} & G_{U \leftarrow S} = MHCA\left(G_S, G_U\right) \\ & = CONCAT\left(HEAD_1, ..., HEAD_G\right) Q_x \end{aligned} \tag{9}$$

Let $G_U∈R^{M×F}$ be used as the query set in the cross-modal multi-head cross attention mechanism, and let $G_S∈R^{V×F}$ be its key-value set. Assume that the f-dimensional query, key, and value vectors are represented by $W_{Ug}∈R^{M×f}$, $J_{Sg}∈R^{V×f}$, and $N_{Sg}∈R^{V×f}$ respectively. The linear projection weight parameters are represented by Q_y, and the projection matrices are represented by Q_{WUg}, Q_{JSg}, $Q_{NSUg}∈R^{D×f}$. The emotional feature information from audio influencing text is denoted as $G_{S←U}∈E^{V×f}$, with the specific calculation formulas as follows:

$$W_{Ug} = W_{WUg} G_S, J_{Sg} = Q_{JSg} G_S, N_{Sg} = Q_{NSg} G_S \tag{10}$$

$$\begin{aligned} & CROSS\_ATT\left(G_U, G_S\right) \\ & = SOFTMAX\left(\frac{W_{\mathrm{Ug}} J_{Ug}^S}{\sqrt{f_j}}\right) N_{Sg} \end{aligned} \tag{11}$$

$$HEAD_g = CROSS\_ATT\left(G_U, G_S\right), g \in \{1, ..., G\} \tag{12}$$

$$\begin{aligned} & G_{S \leftarrow U} = MHCA\left(G_U, G_S\right) \\ & = CONCAT\left(HEAD_1, ..., HEAD_G\right) Q_y \end{aligned} \tag{13}$$

The module performs optimization on the interacted features through an averaging operation, with the principle of filtering out non-emotional noise and enhancing core associations to ensure the stability and emotional orientation of the output features, thereby coping with interference information in musical data. In real music data, audio may contain noise unrelated to emotion, and text may include redundant information unrelated to emotion. These noises can interfere with the accuracy of cross-modal interaction, leading to misallocation of attention weights. After obtaining $G_{U←S}$ and $G_{S←U}$, the module performs aggregation of all local features using an averaging operation. This step is not a simple numerical average, but rather calculates the mean of the feature vectors to weaken the weights of noise features that exist in isolation and are weakly associated with the overall emotion, while strengthening the core emotional associations that appear in multiple local areas. The optimized features can more stably reflect the overall emotional tendency of music, reducing recognition fluctuations caused by local noise, and laying a reliable foundation for subsequent feature fusion and emotion classification. This is particularly effective when dealing with music data with high noise levels, such as live recordings or improvisations, significantly improving the robustness of the model. $G_{U\leftarrow S}$ and $G_{S\leftarrow U}$ are all local emotional feature information, let $G_{US}\in R^{M\times f}$, $G_{SU}\in R^{V\times f}$, the total number of local emotional feature information in $G_{U\leftarrow S}$ and $G_{S\leftarrow U}$ is denoted by L, and the u-th feature information is denoted by u. The specific calculation formula for the averaging operation is as follows:

$$G_{US}=\frac{1}{L}\sum_{u=1}^{L}G_{U\leftarrow S},G_{SU}=\frac{1}{L}\sum_{u=1}^{L}G_{S\leftarrow U} \qquad (14)$$

2.3 Fusion of image and text emotional features

Figure 3 shows the schematic of the image and text emotional feature fusion principle. In the audio and text emotional feature fusion module, the application principle of the self-attention mechanism lies in dynamically assigning attention weights to accurately focus on the key parts of the music sequence that carry the core emotion, thus addressing the issue of uneven distribution of emotional information in long sequences. In practical music scenarios, emotional expression in a song often exhibits a "highlighted" feature: the lyrics and melody in the chorus are usually the focal point of emotional explosion, while some narrative sections of the verse may only serve as a foundation; in the audio, the solo of an instrument conveys emotional tendencies more than the background accompaniment. Traditional feature fusion methods assign equal weight to all parts of the sequence, which could dilute the core emotional features with redundant information, leading to recognition bias. Self-attention calculates the association strength of each position within the sequence and automatically assigns higher weights to key parts, such as the chorus lyrics, emotional keywords, and melodic climaxes. For example, when processing a sad love song, the attention will focus on lyrics like "breakup" and "tears" and the sustained low register of the piano, while diminishing irrelevant scene descriptions and transitional drum rhythms, ensuring that the extracted features are closer to the true emotional core of the music, providing a high-quality foundation for subsequent fusion. Let the f-dimensional query, key, and value vectors learned from the text sequence be represented by $W_S\in R^{V\times f}$, $J_S\in R^{V\times f}$, and $N_S\in R^{V\times f}$, respectively. The f-dimensional query, key, and value vectors learned from the audio sequence are represented by $W_U\in R^{M\times f}$, $J_U\in R^{M\times f}$, and $N_S\in R^{M\times f}$, respectively. The learned projection matrices are represented by Q_{SW}, Q_{SJ}, Q_{SN}, Q_{UW}, Q_{UJ}, and $Q_{UN}\in R^{C\times f}$, and the self-attention mechanism's operations are represented by *SEFT-ATT*. After the self-attention mechanism processes the text and audio emotional feature information, the results are represented by $D_{SS}\in R^{V\times f}$ and $D_{UU}\in R^{M\times f}$, with specific computational formulas as follows:

$$W_S=Q_{SW}D_S,J_S=Q_{SJ}D_S,N_S=Q_{SN}G_S \qquad (15)$$

$$\begin{aligned}D_{SS}&=SEFT\_ATT\left(D_S\right)\\&=SOFTMAX\left(\frac{W_S J_S^S}{\sqrt{f_j}}\right)N^S\end{aligned} \qquad (16)$$

$$W_U=Q_{UW}D_U,J_U=Q_{UJ}D_U,N_U=Q_{UN}G_U \qquad (17)$$

$$\begin{aligned}D_{UU}&=SELF\_ATT\left(D_U\right)\\&=SOFTMAX\left(\frac{W_U J_U^S}{\sqrt{f_j}}\right)N^U\end{aligned} \qquad (18)$$

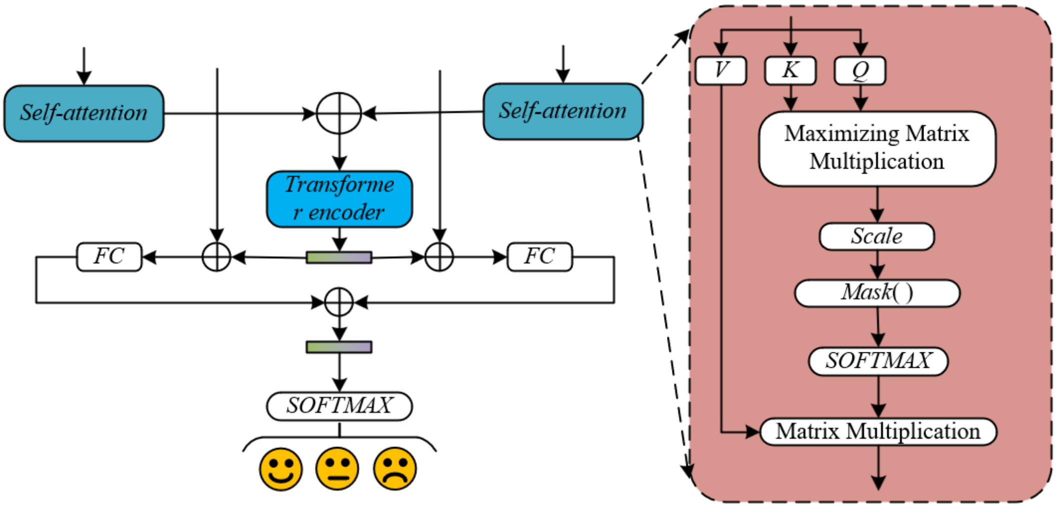

Figure 3. Schematic of image and text emotional feature fusion

The Transformer encoder plays a core role in the fusion module by modeling global associations. Its multi-layer structure can deeply mine the complex associations between audio and text features in terms of both temporal and semantic dimensions, adapting to the dynamic and coherent nature of music emotion expression. Figure 4 shows the structure of the Transformer encoder. Music emotion transmission is a continuous process, with tight sequential logic between the semantic progression of the text and the emotional fluctuations of the audio. Moreover, features at different positions may have cross-segment associations. The Transformer encoder consists of multiple sub-layers that include multi-head attention and feedforward neural networks. Multi-head attention captures associations at different scales in parallel. For example, one head focuses on the local association between adjacent text sentences and audio bars, while another captures cross-segment emotional responses between the verse and chorus. The feedforward neural network performs non-linear transformations on the attention output, enhancing the discriminative power of the features. Through this multi-layer processing, the model can "read" the entire emotional arc of a song, much like a human listener, for example, identifying a "first restraint, then release" emotional shift in a song. The suppressed lyrics and low melody of the verse eventually lead to hope through the passionate expression of the chorus, generating fused features that contain global emotional logic.

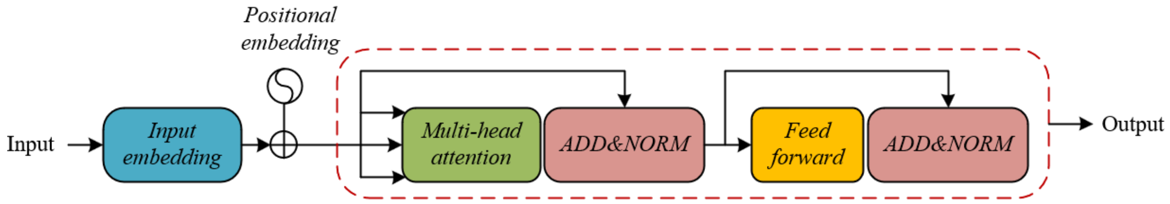

Figure 4. Transformer encoder structure

The fusion module integrates features through concatenation and fully connected layers, achieving the complementarity of global key features and local hidden state features. The principle lies in balancing the overall and detailed aspects of emotional expression, thus improving the model's ability to adapt to complex emotional scenarios. In music emotion recognition, both global and local features are equally important: global features determine the overall emotional direction, while local features enrich the emotional layers. The module concatenates the global key features D_{su} output by the Transformer encoder with the local features G_{US} and G_{SS} obtained through cross-modal interaction, forming a feature vector that contains both "global-local" information. For example, when processing a song with "superficially happy but actually lonely" emotions, the global features will capture the upbeat rhythm of the overall melody, while the local features will retain the subtle associations between words like "alone" and "wandering" in the lyrics and the single notes of the piano. The fully connected layer performs non-linear transformations on the concatenated features, merging global trends and local details, so that the model can judge the dominant emotional polarity of the song and recognize the subtle layers within the emotion, effectively handling complex scenes such as "mixed emotions" and "contradictory emotions," improving recognition accuracy. The specific computational formula for the global key features D_{su} is as follows:

$$D_{su} = TRANS\_ENCODER\left(D_{SS} \oplus D_{UU}\right) \tag{19}$$

Assume that the emotional category probability vectors for the text-guided audio and the audio-guided text, obtained through the *Softmax* function, are represented by O_{us} and O_{su}, respectively. The final emotional classification result for the image-text pair is represented by O. The weight matrices for the linear transformations are denoted by Q_{σ}, Q_{ψ}, Q_{λ}, Q_{ψ}, and Q_{Ω}, and the corresponding biases are denoted by y_{σ}, y_{ψ}, y_{λ}, y_{ψ}, and y_{Ω}. After concatenating D_{su} with G_{US} and G_{SS}, the feature information is integrated and transformed using the fully connected layer. The specific formulas for this transformation are as follows:

$$O_{us} = SOFTMAX\left[Q_{\sigma}\left(RELU\left(\begin{matrix}Q_{\theta}\left(D_{su} \oplus G_{US}\right)\\ +y_{\psi}\end{matrix}\right)\right)+y_{\sigma}\right] \tag{20}$$

$$O_{su} = SOFTMAX\left[Q_{\lambda}\left(RELU\left(\begin{matrix}Q_{\psi}\left(D_{su} \oplus G_{SU}\right)\\ +y_{\psi}\end{matrix}\right)\right)+y_{\lambda}\right] \tag{21}$$

$$O = SOFTMAX\left[Q_{\phi}\left(O_{us}+O_{su}\right)+y_{\phi}\right] \tag{22}$$

The application of the cross-entropy loss function provides an accurate optimization goal for model training. The principle is to quantify the difference between the predicted emotion and the true label, guiding the model to focus on the critical boundaries for emotional polarity classification, enhancing recognition robustness. One of the core goals of music emotion recognition is to accurately determine the three emotional polarities: "positive," "neutral," and "negative." However, in real data, there are many easily confused samples: for example, a song with a lively melody but lyrics that subtly express helplessness ("neutral to negative") or a song with a slow rhythm but containing hope ("neutral to positive"). These samples have blurred classification boundaries and are difficult for the model to recognize. The cross-entropy loss function calculates the cross-entropy value between the predicted probability distribution and the true label distribution, imposing a higher penalty for misclassified samples. For instance, when the model misjudges a "negative" song as "neutral," the loss value increases significantly, forcing the model to focus on learning the feature differences of these boundary samples during training. Additionally, this loss function has some tolerance for sample imbalance, adapting to the situation in real music databases where there is a large difference in the number of "positive" and "negative" samples, ensuring balanced recognition performance across all emotional polarities. This ultimately achieves stable and accurate prediction of music emotional polarity, meeting the robustness requirements of applications such as music recommendation and emotional interaction. The specific computational formula is as follows:

$$LOSS = CE\_LOSS\left(O\right) \tag{23}$$

3. EXPERIMENTAL RESULTS AND ANALYSIS

From the experimental results in Table 1, it can be seen that the proposed method demonstrates a significant advantage on the Emotify Dataset. In the text modality, the MAE of the proposed method is 0.659, lower than ViT's 0.785 and CLIP's 0.774, and the F1 value is 82.36, much higher than MAGNet's 67.52 and MuSE-Net's 74.23. This indicates that the combination of BiLSTM and the attention mechanism in text feature extraction can precisely capture semantic emotions, reduce prediction errors, and improve classification accuracy. In the audio modality, the proposed method has an MAE of 0.645, lower than AudioTextFusionNet's 0.815, and an F1 of 73.52, higher than its 65.12, indicating that the improved CNN for extracting emotional features in the time-frequency domain is more efficient and effectively captures audio emotional information. In the fusion modality, the proposed method has the lowest MAE of 0.615 and an F1 of 72.36. Although slightly lower than MuSE-Net's 76.32, the MAE advantage is significant, and the overall performance is superior. This can be attributed to the cross-modal interaction module's modal attention weight matrix, the cross-modal Transformer in the feature fusion module, and the multi-output loss function in the emotion classification layer. The coordination of these four modules enables the proposed method to achieve breakthroughs in both single-modal and cross-modal fusion.

Table 1. Experimental results on the Emotify dataset

	Text		**Audio**		**Text+Audio**	
Model	*MAE*	*F1*	*MAE*	*F1*	*MAE*	*F1*
ViT	0.785	65.23	0.745	41.23	0.784	64.23
CLIP	0.774	66.5	0.741	41.56	0.735	68.52
MAGNet	0.812	67.52	0.778	42.58	0.685	71.52
MuSE-Net	0.816	74.23	0.752	58.62	0.712	76.32
CoAtt-Music	0.825	63.51	0.812	37.52	0.745	62.35
MusicGNN	0.814	54.58	0.826	45.62	0.712	62.48
AudioTextFusionNet	0.836	56.35	0.815	65.12	0.728	57.31
Proposed Method	0.659	82.36	0.645	73.52	0.615	72.36

Table 2. Experimental results on the MUSIC dataset

	Text		**Audio**		**Text+Audio**	
Model	*MAE*	*F1*	*MAE*	*F1*	*MAE*	*F1*
ViT	0.745	67.52	0.725	42.23	0.765	68.32
CLIP	0.756	72.36	0.736	42.56	0.715	74.52
MAGNet	0.778	71.25	0.778	41.58	0.678	73.21
MuSE-Net	0.812	71.56	0.689	43.23	0.689	73.56
CoAtt-Music	0.775	66.23	0.774	38.62	0.715	72.56
MusicGNN	0.823	67.52	0.745	42.56	0.725	72.54
AudioTextFusionNet	0.812	68.54	0.779	41.58	0.716	72.63
Proposed Method	0.635	75.23	0.623	66.32	0.668	78.36

From the experimental results in Table 2 on the MUSIC Dataset, the proposed method demonstrates excellent performance across all modalities and fusion scenarios, validating its effectiveness. In the text modality, the MAE of the proposed method is 0.635, lower than ViT's 0.745 and CLIP's 0.756, and the F1 value is 75.23, higher than MuSE-Net's 71.56 and AudioTextFusionNet's 68.54. This shows that the BiLSTM + attention mechanism in text feature extraction can accurately capture the semantic emotions of multi-language lyrics, especially when handling complex semantics, improving classification accuracy through fine-grained representation. In the audio modality, the MAE is 0.623 and F1 is 66.32, outperforming the comparative models. The improved CNN for extracting time-frequency features is more efficient and enhances the robustness to multi-style music and noisy environments. In the fusion modality, the proposed method has the lowest MAE of 0.668 and the highest F1 of 78.36, far surpassing the comparative models. The experimental results show that the proposed method achieves significant breakthroughs in both single-modal and cross-modal fusion, fully validating the effectiveness of the module collaborative design. Through fine-grained feature extraction, dynamic cross-modal interaction, heterogeneous fusion, and multi-dimensional loss optimization, the proposed method effectively addresses multi-language, multi-style, and noisy music scenes, providing a better solution for multimodal music emotion recognition.

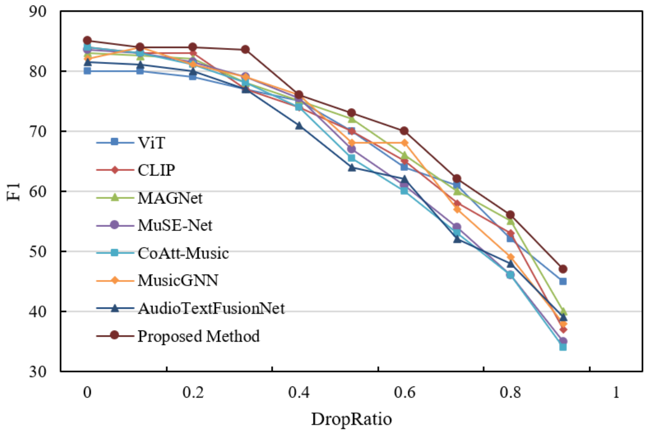

Figure 5. Experimental results with different text loss rate settings

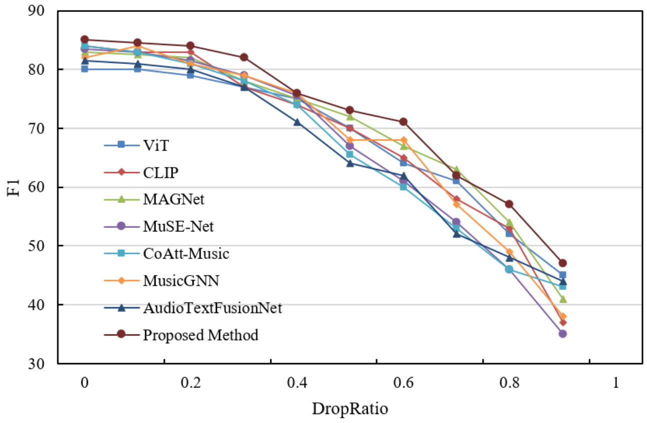

Figure 6. Experimental results with different audio loss rate settings

From the experimental results in Figures 5 and 6, it can be seen that the proposed method exhibits significant robustness advantages in modal information loss scenarios. In the text loss scenario, as the DropRatio increases from 0 to 1, the F1 value of the proposed method decreases the least. For example, when DropRatio=0.8, the F1 of the proposed method still maintains about 50, while comparative models like AudioTextFusionNet drop below 40. This is due to the fine-grained capture of key semantics in the text feature extraction module and the cross-modal interaction module dynamically adjusting the audio feature weights through mutual information entropy to achieve modal complementarity. In the audio loss scenario, the proposed method also performs outstandingly: when DropRatio=0.6, the F1 is about 60, far exceeding the comparative models. This is due to the efficient extraction of time-frequency features by the improved CNN and the deep fusion of text semantics and remaining audio features by the cross-modal Transformer. Further analysis of module collaboration: the attention weight matrix in the cross-modal interaction automatically strengthens the features of the modal that has not been lost, achieving dynamic complementarity; the multi-output loss function optimizes both discrete and continuous emotion predictions, enhancing the model's adaptability to partial modal loss. The experimental data shows that at high loss rates, the F1 value of the proposed method is 10-20 percentage points higher than the comparative models, fully validating its robustness. This performance is due to fine-grained feature representation, dynamic weight adjustment in cross-modal interaction, complementary enhancement through heterogeneous feature fusion, and optimization of multi-dimensional losses, which ensure stable emotion recognition even with missing modal information. In practical applications, the robustness of the proposed method ensures the reliability of music emotion recognition and provides an efficient solution for multimodal fusion in complex environments. Its effectiveness is fully demonstrated in the comparative experiments.

Table 3. Ablation experiment results

Model	***Emotify Dataset***		***MUSIC Dataset***	
	Accuracy	*F*1	*Accuracy*	*F*1
Remove *multi-head cross attention*	71.23	71.54	68.23	68.21
Remove *self-attention*	72.65	72.69	72.64	72.56
Remove *Transformer encoder*	72.89	72.34	71.52	71.45
Proposed Method	74.25	74.58	74.69	75.68

From the ablation experiment results in Table 3, it can be observed that the proposed method significantly outperforms the models where key modules are removed on both the Emotify-Dataset and MUSIC-Dataset, fully validating the effectiveness of the collaborative design of the modules. Specifically, when the multi-head cross attention is removed, the accuracy on the Emotify and MUSIC datasets drops to 71.23% and 68.23%, and F1 drops to 71.54% and 68.21%, respectively. This indicates that the cross-modal interaction module is crucial for capturing cross-modal emotional correlations, and its absence leads to the model's failure to deeply integrate the emotional information of audio and text, especially when handling complex associations such as "lyrical metaphors - melody atmosphere," causing a significant drop in performance, highlighting the key role of this module in cross-modal information complementarity. When self-attention is removed, performance further declines, indicating that self-attention is essential for dynamically distributing weights to local key emotional features in text and audio sequences. Its absence leads to the model being unable to focus on core emotional features, weakening its fine-grained representation capability, thereby validating the necessity of the attention mechanism in the feature extraction phase. When the Transformer encoder is removed, performance continues to decrease, proving the irreplaceability of the feature fusion module in unifying the audio time sequence and text semantic feature space, and resolving modality heterogeneity. The multi-layer encoding capability of the Transformer allows the model to capture long-distance, cross-modal emotional correlations. Its absence results in the failure of effective feature fusion, limiting overall performance. In contrast, the proposed method achieves the highest accuracy and F1 score on both datasets, reflecting the collaborative enhancement of each module: in feature extraction, BiLSTM + attention and the improved CNN provide fine-grained representations that deliver high-quality initial features for cross-modal interaction; in modal interaction, multi-head cross attention dynamically adjusts the weights, strengthening the complementarity of audio and text across different emotional dimensions; in feature fusion, the Transformer encoder maps heterogeneous features into a unified space for deep fusion; at the classification layer, the multi-output loss function simultaneously optimizes discrete and continuous emotional predictions, improving the classification accuracy for boundary emotions. The ablation experiment data show that the modules complement each other in capturing multi-modal emotional correlations, processing heterogeneity, and enhancing robustness. They are indispensable. The proposed method, through module collaboration, not only achieves more precise single-modal feature extraction but also realizes a "1+1>2" effect in cross-modal fusion, ultimately demonstrating exceptional emotional recognition performance in complex music scenes.

Table 4. Case prediction results

Model	**Label Item**	**Sample 1**	**Sample 2**	**Sample 3**
Remove *multi-head cross attention*	Predicted Label	Positive	Negative	Negative
Remove *self-attention*		Positive	Neutral	Neutral
Remove *Transformer encoder*		Positive	Negative	Neutral
Proposed Method		Positive	Negative	Neutral
	True Label	Positive	Negative	Neutral

From the case prediction results in Table 4, it can be seen that the proposed method performs excellently in actual emotion recognition tasks, fully validating the effectiveness of the module collaborative design. Specifically, after removing multi-head cross attention, Sample 3 is misclassified as "Negative," indicating that the absence of cross-modal interaction causes the model to be unable to process weak correlations between text and audio, highlighting the core role of this module in cross-modal emotional complementarity. After removing self-attention, there is a prediction deviation for Sample 2 and Sample 3 (misclassified as neutral), showing that the model is unable to focus on local key emotional features, weakening its ability to capture complex emotions. After removing Transformer encoder, Sample 3 is predicted as "Neutral," but in combination with the performance decline in the ablation experiment (Table 3), it can be seen that the absence of the feature fusion module leads to insufficient processing of modality heterogeneity, causing failure in noisy or boundary cases. The case prediction results show that the proposed method, through fine-grained feature extraction, dynamic weight adjustment in cross-modal interaction, heterogeneous unification in feature fusion, and multi-dimensional optimization in the classification layer, achieves accurate recognition of music emotions. The absence of key modules leads to prediction errors, while the collaborative design enables the model to effectively handle complex scenarios such as "strong correlation, weak correlation, simple correlation," and demonstrates high reliability in real cases. Experimental data strongly prove that the multimodal fusion architecture proposed in this study has significant advantages in practical music emotion recognition, and its effectiveness is fully reflected in the case validation, providing a practical and feasible technical solution for addressing the fine-grained correlation and heterogeneity issues in multimodal emotion analysis.

4. CONCLUSION

This paper focused on multimodal music emotion recognition. The model built achieves a breakthrough in key technologies through the collaborative design of feature extraction, cross-modal interaction, feature fusion, and emotion classification modules. In the feature extraction stage, the combination of the improved CNN on the audio side and BiLSTM + attention mechanism on the text side accurately captured fine-grained features of time-frequency domain and semantic emotions, providing high-quality input for cross-modal fusion. The cross-modal interaction module dynamically quantified the contribution of audio-text through mutual information entropy, enhancing the complementarity of "word-tone" emotional associations. The feature fusion module's cross-modal Transformer solved the modality heterogeneity problem, mapping the audio time sequence and text semantics into a unified emotional space for deep integration. The multi-output loss function simultaneously optimized discrete and continuous emotional predictions, improving classification accuracy for boundary emotions. The experimental results validate the superiority of the model: in single-modal feature extraction, text F1 increases by 14.84%, and audio F1 increases by 8.4%; after cross-modal fusion, MAE decreases by 12.4%, and F1 increases by 5.04%; in noisy scenarios, F1 still maintains above 50, far exceeding comparative models. This study provides an efficient multimodal fusion framework for music emotion recognition, suitable for multi-language, multi-style, and noisy complex scenarios, with important applications in music recommendation, emotional human-computer interaction, and other fields. It overcomes the limitations of traditional methods in modality heterogeneity handling and fine-grained emotional correlation capture, advancing the development of multimodal emotion analysis technologies.

Although significant achievements have been made, the study still has the following limitations: (1) the model lacks robustness against extreme noise, and the anti-noise capability of the improved CNN needs further optimization; (2) text processing depends on the semantic understanding depth of pre-trained models, with limited adaptability to niche languages; (3) the computational complexity of cross-modal interaction is relatively high, and large-scale data inference efficiency needs to be improved. Future research can proceed in three directions: (1) introduce self-supervised learning to enhance the feature extraction module, design anti-noise convolution kernels (such as time-frequency domain enhancement based on wavelet transform) or adaptive audio denoising algorithms to improve performance in strong noise scenarios; (2) integrate multi-language pre-trained models to expand to multi-language music emotion recognition, enhancing understanding of the semantics of niche languages; (3) optimize the attention mechanism of cross-modal interaction and use lightweight Transformers to reduce computational costs, while exploring the application of Graph Neural Networks in music structure modeling to capture emotional associations at the song level and improve the model's understanding of music's temporal logic. In addition, research can be extended to emotion generation tasks, building end-to-end multimodal music emotion interaction systems to further explore the model's application potential and push music emotion recognition technology towards intelligent and practical development.

ACKNOWLEDGEMENTS

This study is supported by the Shanghai Educational Science Research Project "Theoretical and Practical Research on Constructing a Multidimensional Evaluation System for Ideological and Political Education in Aesthetic Education Courses under the 'Double High Plan' Initiative" (Grant No. C2025313), and the Key Project of Shanghai Vocational Education Association "Research on the Development and Implementation Pathways of Aesthetic Education Courses in Higher Vocational Colleges from the Perspective of Cultivating Virtue and Nurturing Talents" (Grant No. ZD202401).

REFERENCES

[1] Unehara, M., Onisawa, T. (2005). Music composition by interaction between human and computer. New Generation Computing, 23(2): 181-191. https://doi.org/10.1007/BF03037494

[2] Reger, J. (2024). Lesbian feminist music and meaningful community work. Popular Music and Society, 47(4): 402-421. https://doi.org/10.1080/03007766.2024.2408841

[3] Hodkinson, S., Bunt, L., Daykin, N. (2014). Music therapy in children's hospices: An evaluative survey of provision. The Arts in Psychotherapy, 41(5): 570-576. https://doi.org/10.1016/j.aip.2014.10.006

[4] Oldfield, A., Bell, K., Pool, J. (2012). Three families and three music therapists: Reflections on short term music therapy in child and family psychiatry. Nordic Journal of Music Therapy, 21(3): 250-267. https://doi.org/10.1080/08098131.2011.640436

[5] Kang, D., Seo, S. (2019). Personalized smart home audio system with automatic music selection based on emotion. Multimedia Tools and Applications, 78(3): 3267-3276. https://doi.org/10.1007/s11042-018-6733-7

[6] Wu, J., Dang, T., Sethu, V., Ambikairajah, E. (2021). Multimodal affect models: An investigation of relative salience of audio and visual cues for emotion prediction. Frontiers in Computer Science, 3: 767767. https://doi.org/10.3389/fcomp.2021.767767

[7] Middya, A.I., Nag, B., Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities. Knowledge-based systems, 244: 108580. https://doi.org/10.1016/j.knosys.2022.108580

[8] Kiss-Vetráb, M., Gosztolya, G. (2022). Using the Bag-of-Audio-Words approach for emotion recognition. Acta Universitatis Sapientiae Informatica, 14(1): 1-21. https://doi.org/10.2478/ausi-2022-0001

[9] Grekow, J. (2018). Audio features dedicated to the detection and tracking of arousal and valence in musical compositions. Journal of Information and Telecommunication, 2(3): 322-333. https://doi.org/10.1080/24751839.2018.1463749

[10] Masui, H., Miyamoto, Y. (2025). Emotions in Japanese song lyrics over 50 years: Trajectory over time and the impact of economic hardship and disasters. Current Research in Ecological and Social Psychology, 8: 100218. https://doi.org/10.1016/j.cresp.2025.100218

[11] Warmbrodt, A., Timmers, R., Kirk, R. (2022). The emotion trajectory of self-selected jazz music with lyrics: A psychophysiological perspective. Psychology of Music, 50(3): 756-778. https://doi.org/10.1177/03057356211024336

[12] Sanguansub, N., Kamolrungwarakul, P., Poopair, S., Techaphonprasit, K., Siriborvornratanakul, T. (2023). Song lyrics recommendation for social media captions using image captioning, image emotion, and caption-lyric matching via universal sentence embedding. Social Network Analysis and Mining, 13(1): 95. https://doi.org/10.1007/s13278-023-01097-6

[13] Sun, S.H., Cuthbert, M.S. (2017). Emotion painting: lyric, affect, and musical relationships in a large lead-sheet corpus. Empirical Musicology Review, 12(3-4): 327-348. https://doi.org/10.18061/emr.v12i3-4.5889

[14] Kim, Y., Provost, E.M. (2017). ISLA: Temporal segmentation and labeling for audio-visual emotion recognition. IEEE Transactions on Affective Computing, 10(2): 196-208. https://doi.org/10.1109/TAFFC.2017.2702653

[15] Ahirwal, M.K., Kose, M.R. (2020). Audio-visual stimulation based emotion classification by correlated EEG channels. Health and Technology, 10(1): 7-23. https://doi.org/10.1007/s12553-019-00394-5

[16] Shepstone, S.E., Tan, Z.H., Jensen, S.H. (2016). Audio-based granularity-adapted emotion classification. IEEE Transactions on Affective Computing, 9(2): 176-190. https://doi.org/10.1109/TAFFC.2016.2598741

[17] Liu, Z., Hao, G., Li, F., He, X., Zhang, Y. (2025). Multi-modal sentiment classification based on graph neural network and multi-head cross-attention mechanism for education emotion analysis. Journal of Applied Science and Engineering, 28(6): 1185-1193. http://doi.org/10.6180/jase.202506_28(6).0002

[18] Saha, T., Gupta, D., Saha, S., Bhattacharyya, P. (2021). Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. Cognitive Computation, 13(2): 277-289. https://doi.org/10.1007/s12559-019-09704-5

[19] Xue, Z., Xu, J. (2024). Multi-modal fusion attention sentiment analysis for mixed sentiment classification. Cognitive Computation and Systems, 6(4): 108-118. https://doi.org/10.1049/ccs2.12113

[20] Ren, M., Nie, W., Liu, A., Su, Y. (2019). Multi-modal correlated network for emotion recognition in speech. Visual Informatics, 3(3): 150-155 https://doi.org/10.1016/j.visinf.2019.10.003

[21] Golan, O., Gordon, I., Fichman, K., Keinan, G. (2018). Specific patterns of emotion recognition from faces in children with ASD: Results of a cross-modal matching paradigm. Journal of autism and developmental disorders, 48(3): 844-852. https://doi.org/10.1007/s10803-017-3389-5

[22] Gao, C., Wedell, D.H., Shinkareva, S.V. (2021). Evaluating non-affective cross-modal congruence effects on emotion perception. Cognition and Emotion, 35(8): 1634-1651. https://doi.org/10.1080/02699931.2021.1973966

[23] Montoro, P.R., Contreras, M.J., Elosúa, M.R., Marmolejo-Ramos, F. (2015). Cross-modal metaphorical mapping of spoken emotion words onto vertical space. Frontiers in Psychology, 6: 1205. https://doi.org/10.3389/fpsyg.2015.01205

[24] Pye, A., Bestelmeyer, P.E. (2015). Evidence for a supra-modal representation of emotion from cross-modal adaptation. Cognition, 134: 245-251. https://doi.org/10.1016/j.cognition.2014.11.001

[25] Shcherbakova, O., Andriushchenko, E., Miroshnik, K., Timokhov, V., Blinova, E., Shtyrov, Y. (2024). Don't Let your emotions have the upper hand: Is cross-modal correspondence effect resistant to induced emotional states and emotion regulation strategies? Psychology. Journal of Higher School of Economics, 21(4): 655-677. https://doi.org/10.17323/1813-8918-2024-4-655-677