Cross-Modal Transformer with Dynamic Attention Fusion for Emotion Recognition in Music via Audio-Lyrics Alignment

Xiaofeng Li
Shanxi Normal University, Xian, 710119 ,Shanxi,China
E-mail:18681857525@163.com

Keywords: cross-modal emotion recognition, transformer, attention mechanism, multimodal fusion, robust classification

Received: September 5, 2025

Emotion recognition from multimodal signals remains a challenging task due to annotation subjectivity and heterogeneous feature spaces. To address these issues, this study proposes a cross-modal Transformer architecture with dynamic attention fusion for robust emotion classification. Raw acoustic signals are converted into time–frequency spectrograms, from which hierarchical features are extracted via a deep convolutional network. In parallel, textual data (e.g., lyrics or aligned semantic content) are encoded with a pre-trained language model to obtain context-aware embeddings. A cross-modal attention mechanism embedded in the Transformer encoder adaptively models inter-modal associations, enabling semantically guided acoustic representation learning. The fused joint representation is aggregated through pooling and passed to a fully connected classifier, yielding multi-category emotion probabilities. Experimental evaluations demonstrate that the proposed Transformer model outperforms CNN, CRNN, and traditional Transformer models in noisy conditions (average accuracy = 0.58; macro F1 = 0.55 at 0 dB SNR) and exhibits superior generalization capabilities across datasets (AUC = 0.832–0.887). Furthermore, with only 30% labeled data, the model maintains reliable emotion continuity (CCC = 0.635; ICC = 0.584), highlighting its effectiveness in low-resource scenarios. These results confirm the potential of cross-modal Transformer fusion for advancing emotion-aware intelligent systems in multimodal perception applications.

Povzetek: Naša metoda, združuje akustične in tekstovne podatke, izboljša natančnost prepoznavanja čustev v hrupnih in podatkovno omejenih razmerah. Model preseže obstoječe pristope in izkazuje dobro posploševanje med različnimi podatkovnimi zbirkami.

1 Introduction

Music emotion recognition analyzes emotional semantics in auditory signals, bridging computational perception and cognitive intelligence [1-2]. In applications like human-computer empathy, personalized recommendation, and digital mental health, systems must accurately decode musical emotions [3-4]. As a multimodal medium, music conveys emotion through the deep coupling of acoustic structure (pitch, rhythm, harmony) and linguistic symbols (semantic imagery) [5-6]. Deep learning, especially cross-modal representation learning, enables joint audio-text understanding [7-8], advancing affective computing and intelligent music systems.

Core challenges include multimodal semantic heterogeneity and emotion representation uncertainty. Audio carries emotion via continuous time-frequency patterns shaped by timbre, rhythm, and harmony. Lyrics, as discrete sequences, convey emotion through semantic imagery, rhetoric, and lexical polarity [9-10]. Their differences in data form, time granularity, and logic make direct fusion ineffective for semantic alignment [11-12]. Emotion lacks objective measurement; annotations vary with experience and culture, causing "many-to-one" or "one-to-many" ambiguity [13]. Most fusion methods use fixed weights or parallel coding, lacking dynamic adjustment to modulate text guidance on audio based on content [14]. Pre-trained language models capture deep semantics but struggle with context-dependent shifts: "broken" may express emotional release in lyrics [15]. Convolutional features, while locally invariant, poorly model long-term emotional evolution [16]. Models must learn robust boundaries under uncertain labels, balancing fine-grained modality alignment with overall consistency—challenging representation design, fusion flexibility, and training robustness.

This study proposes an end-to-end cross-modal Transformer to improve music emotion recognition by dynamically fusing audio and lyrics. A convolutional network extracts hierarchical acoustic features from time-frequency transformed audio, while a pre-trained language model encodes lyrics into context-aware semantic vectors. Features are aligned via linear mapping and positional encoding. Within the Transformer, audio features serve as queries and lyrics as keys/values, enabling semantic-guided, dynamic fusion through selective attention. The

fused representation is pooled and classified into an eight-dimensional emotion space. Key innovations include: a query-key-value separated attention architecture for enhanced cross-modal modeling, a fully differentiable pipeline to prevent manual fusion losses, and a position-aware strategy that improves temporal alignment by explicitly modeling lyric order. In choral repetition sections or semantic loop structures, this strategy enhances the temporal consistency and interpretability of attention mechanisms. The study provides a novel architectural paradigm for multimodal affective computing, establishing a scalable technical foundation for the practical deployment of music understanding systems.

2 Related work

Early research on emotion recognition primarily relied on handcrafted acoustic features, such as Mel-frequency cepstral coefficients, combined with traditional classifiers like random forests [17–20]. While effective on small datasets, these approaches were limited in their ability to capture complex emotional dynamics. With the advent of deep learning, convolutional neural networks were introduced into audio processing, exploiting local receptive fields to extract spatial patterns in spectrograms and thereby improving feature abstraction [21–22]. Recurrent neural networks were subsequently employed to model temporal dependencies and enhance the tracking of emotional changes [23–24]. In parallel, advances in pre-trained language models such as BERT significantly improved textual representation quality. Some studies applied these models to lyrics encoding and explored simple concatenation or weighted fusion with acoustic features [25]. However, most of these approaches adopted static fusion strategies, overlooked the asymmetry in temporal granularity between audio and text, and lacked mechanisms for selecting key semantic components, ultimately limiting their robustness and interpretability.

To overcome these shortcomings, researchers have increasingly turned to interactive fusion mechanisms. Attention-based methods provided a new pathway for cross-modal correlation modeling. For example, LSTM–attention hybrids enabled audio sequences to focus on emotionally salient keywords, verifying the feasibility of semantic-guided acoustic analysis [26–27]. Transformer architectures were later introduced into multimodal tasks, where self-attention captured intra-sequence dependencies and cross-attention enabled effective information exchange between modalities [28]. Such structures have shown superior performance in video emotion recognition and image–text matching, suggesting strong potential for complex semantic alignment tasks. Explorations into Transformer-based audio representation learning and wearable emotion recognition also indicate promising directions, though most remain unimodal in focus [29]. Some studies have attempted to map lyrics embeddings and spectral features into a shared semantic space with contrastive learning, but fine-grained alignment and robust adaptation across musical styles remain unresolved challenges. Moreover, generic design choices in positional encoding, feature matching, and pooling often fail to account for the temporal continuity of music signals and the discrete nature of textual structures.

Building on this trajectory, multimodal emotion recognition has gained significant momentum. Arumugam et al. introduced a multi-model deep learning framework that enhanced human emotion recognition accuracy [30]. Khan et al. developed MSER, which employs cross-attention with deep fusion for multimodal speech emotion recognition [31]. Deng et al. proposed Sync-TVA, a graph-attention framework that improves temporal alignment across modalities [32], while Liu et al. designed a noise-resistant multimodal Transformer to achieve robust performance in noisy environments [33]. Wafa et al. advanced multimodal emotion recognition in big data contexts by integrating prompt engineering with deep adaptive learning [34]. In addition, Savchenko et al. leveraged lightweight facial models with textual inputs for audio–visual emotional understanding in the wild [35], and Goel et al. provided a comprehensive review of emotion-aware speech translation, highlighting the promise of cross-lingual and cross-modal fusion in affective computing [36].

Together, these studies underscore the growing importance of cross-modal fusion, attention mechanisms, and adaptive optimization as core strategies for enhancing both accuracy and generalization in emotion recognition. Nevertheless, the deep adaptation of Transformers and the design of fine-grained, interpretable fusion mechanisms—particularly for tasks involving heterogeneous modalities such as audio and text—remain open challenges. This paper addresses these gaps by proposing a cross-modal Transformer architecture with attention-based fusion, aiming to achieve dynamic alignment and collaborative enhancement of acoustic and textual modalities at the semantic level.

3 Method

Figure 1 illustrates the cross-modal fusion architecture for music emotion recognition. The system processes audio (spectral features extracted using STFT and CNN) and lyrics (encoded into semantic vectors using BERT) in parallel. The core innovation lies in the cross-modal Transformer module, which uses audio features as queries and lyric vectors as keys. A multi-head attention mechanism dynamically calculates modal association weights, enabling semantically guided acoustic emotion fusion. The fused representations are pooled and fed into a classification network, ultimately outputting an eight-dimensional emotion probability distribution. This design effectively improves the model's ability to express and generalize complex musical emotions.

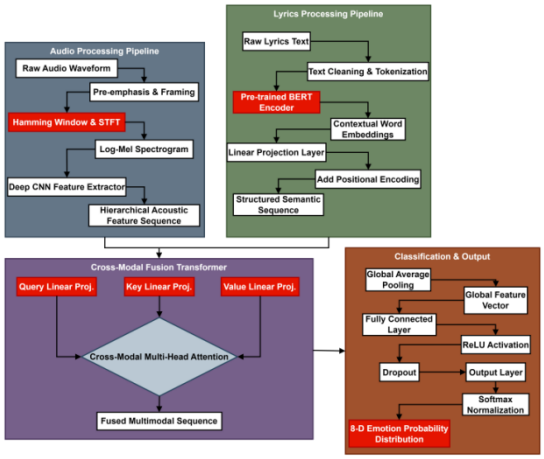

Figure 1: Cross-modal music emotion recognition fusion architecture

3.1 Audio spectrum feature extraction and processing

3.1.1 Time-frequency conversion of audio signals and construction of perceptual spectrum

To preserve music's emotion-relevant temporal and spectral structure, raw audio undergoes preprocessing for stability and perceptual consistency. A first-order high-pass filter applies pre-emphasis to enhance high-frequency components, mitigating recording-induced attenuation and improving the discernibility of emotional cues like syllable onsets and percussive transients. The signal is then segmented into 25 ms frames with a 10 ms shift, balancing temporal resolution for rhythm capture with smoothness for frequency analysis. Each frame is windowed with a Hamming function to reduce spectral leakage, followed by a Short-Time Fourier Transform (STFT) to create a complex time-frequency representation, forming the basis for further analysis.

This linear spectrum is converted into a perceptual Mel-scale spectrogram using a 40-filter triangular bandpass bank covering 25–8000 Hz, providing higher resolution at low frequencies to simulate the human auditory system's nonlinear response. The energy output of each filter is log-compressed, generating a logarithmic Mel-spectrogram that reduces dynamic range and emphasizes emotion-sensitive bands—such as warmth, tension, and brightness—while highlighting subtle acoustic fluctuations over absolute energy differences. The resulting two-dimensional map preserves rhythm and pitch contours, enhancing emotion-relevant features and providing high-quality input for deep learning, enabling the model to focus on perceptually meaningful acoustic patterns rather than physical details.

$$S_{\text{mel}}[t,m]=\sum_k |X[t,k]|^2 \cdot H_m(k) \quad (1)$$

Where $S_{\text{mel}}[t,m]$represents the output energy of the mth Mel filter in the tth frame, $X[t,k]$is the power spectrum of the corresponding frame, $H_m(k)$and is the weight of the mth triangular filter at the frequency index k. This formula implements energy redistribution from the linear spectrum to the perceptual spectrum and is a key step in constructing emotion-sensitive acoustic representations.

3.1.2 Convolutional Extraction of Hierarchical Acoustic Feature Sequences

After obtaining the logarithmic Mel spectrogram, a four-layer deep convolutional network performs hierarchical feature abstraction. Each layer includes convolution, batch normalization, and ReLU activation, with gradually expanding receptive fields for local-to-global integration. The first layer uses small kernels to capture short-term timbre and phoneme-level features; deeper layers employ vertically expanded kernels to model harmonic and pitch patterns in the frequency dimension. Zero padding, stride, and padding control ensure strict time-axis alignment, providing precise temporal correspondence for cross-modal fusion. Batch normalization stabilizes distributions and improves training robustness, while ReLU enhances discrimination and mitigates gradient vanishing. The resulting high-dimensional, temporally aligned feature sequence integrates multi-scale acoustic information—from local details to mid/high-level rhythm—offering emotionally sensitive, compact representations that support dynamic cross-modal attention and enable deep audio-semantic fusion.

$$f_{\text{out}}=\text{ReLU}\left(\text{BN}\left(W*f_{\text{in}}+b\right)\right) \quad (2)$$

Among them, $f_{\text{in}}$is the input feature map, $W$and $b$are the convolution kernel weight and bias term respectively, BNrepresents the batch normalization operation, *is the convolution operation, $f_{\text{out}}$and is the output feature. This operation constitutes the core calculation process of feature extraction at each layer.

Table 1: Audio preprocessing and Mel spectrum generation parameters

Parameter	Value / Configuration	Description
Sampling Rate	22050	Resampled audio frequency for processing
Pre-emphasis Coefficient	0.95	First-order high-pass filter coefficient
Frame Length	25	Duration of each audio frame in milliseconds
Frame Shift	10	Step size between consecutive frames
FFT Size	1024	Number of points for short-time Fourier transform
Number of Mel Filters	40	Count of triangular bandpass filters in Mel bank
Frequency Range	25 – 8000	Lower and upper bounds of filter coverage
Log Energy Offset	1e-6	Small constant to avoid log(0) computation

Table 1 lists the key numerical parameters for the audio preprocessing and Mel-spectrogram generation stages. After resampling the original audio to 22050 Hz, it is framed using a 25 ms frame length and a 10 ms frame shift. A 1024-point short-time Fourier transform is then used to obtain a frequency domain representation. A Mel-scale spectrum is constructed in the 25–8000 Hz range using 40 triangular filters to simulate human ear perception. A pre-emphasis factor of 0.95 is used to enhance high-frequency detail, and a 1e-6 bias is introduced during logarithmic compression to prevent numerical anomalies. All parameters are designed to balance computational stability with the preservation of emotion-related acoustic features, ensuring that the output spectrum has good resolution in both time and frequency dimensions, providing high-quality input for subsequent convolutional feature extraction.

3.2 Lyric semantic vector encoding and projection

3.2.1 Context-aware semantic encoding of lyrics text

To obtain music-aligned deep semantics, lyrics undergo preprocessing: aligned by timestamp, denoised to remove non-sung content and symbols, then segmented using a sub-word strategy for mixed text to handle unknown words. Processed sequences are fed into a pre-trained BERT model, which uses a bidirectional Transformer to generate context-sensitive embeddings—each word vector incorporates surrounding context, enabling differentiation of contextual emotional meaning (e.g., "silence" as depression or contemplation). To ensure stability, BERT's parameters are frozen; only a learnable linear projection adapts its output to the shared space for end-to-end optimization. This avoids overfitting on limited music data while preserving encoding consistency. The result is an ordered sequence of high-dimensional vectors, each representing a time-aligned semantic unit, forming the input for downstream processing.

$$\mathbf{h}_i = \mathrm{BERT}(w_i; C_i) \quad (3)$$

Here, $\mathbf{h}_i$ represents the context-aware vector of the i-th token, and is w_i the input for its corresponding token, C_i representing the context of the word in the sequence. This formula describes the mapping of word vectors from discrete symbols to a continuous semantic space and is the core step in achieving deep semantic parsing of lyrics.

3.2.2 Dimension Alignment of Semantic Vectors and Spatiotemporal Structured Modeling

To integrate with acoustic features, BERT's word vectors are first projected into a uniform latent space via a trainable linear layer. This fully connected transformation, optimized during training, aligns the dimensionality with audio features, enabling adaptation to downstream cross-modal tasks. To enhance temporal alignment, absolute position encoding (PE), using sine and cosine functions, is added to the projected vectors. As lyrics and audio are pre-aligned at the segment level, PE clearly identifies each word's temporal order, ensuring accurate correspondence between semantic units and acoustic events during fusion. This structured representation preserves contextual information while enforcing cross-modal temporal consistency, resulting in a spatiotemporal sequence matched to the audio stream, which provides a precise, differentiable input for dynamic cross-modal attention computation.

$$\mathbf{e}_i = \mathbf{W}_p \mathbf{h}_i + \mathbf{p}_i \quad (4)$$

Where $\mathbf{e}_i$ is the structured semantic vector of the i-th word, $\mathbf{W}_p$ is the learnable projection weight matrix, and $\mathbf{p}_i$ is the position encoding vector of the corresponding position. This transformation realizes the end-to-end mapping from the context vector to the fusion-ready representation.

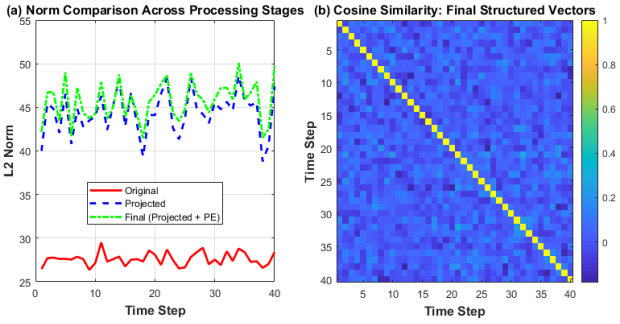

Figure 2: Linear projection and position encoding analysis

Figure 2 reveals the synergistic effect of linear projection and positional encoding from the perspectives of vector norm and similarity structure. Figure 2(a) shows the original embedding, the projected embedding, and the final embedding (projection + positional encoding). The projection operation significantly reduces the vector norm and smoothes its temporal fluctuations, indicating that high-dimensional semantic features are effectively compressed into a latent space that matches the audio features, while also eliminating the scale sensitivity of BERT embeddings. The norm uniformly recovers and remains stable after adding positional encoding, demonstrating that sine-cosine encoding maintains the stability of feature distribution while incorporating positional information, thus avoiding numerical mismatch during modal fusion. The heatmap in Figure (b) shows a prominent diagonal-dominant pattern, with the highest similarity between adjacent time steps and decreasing with distance, confirming that positional encoding successfully captures the temporal structure of the lyrics. Locally high similarity patches in the off-diagonal regions reflect recurring semantic patterns in the lyrics (such as the repetition of the chorus). This structural similarity provides a precise temporal alignment basis for cross-modal attention, enabling the model to dynamically associate audio segments with semantic units.

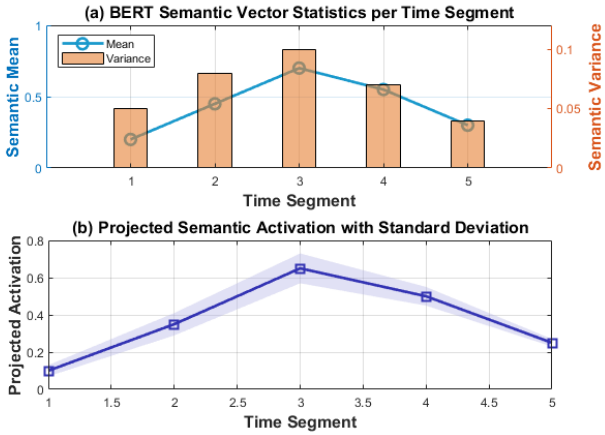

Figure 3: Lyric semantic temporal dynamics and projection activation

Figure 3 analyzes lyric semantics over time. Figure 3(a) shows the semantic mean and variance peak in the third segment, indicating maximum richness and emotional complexity, likely the chorus. Surrounding segments show lower values, suggesting simpler content. Figure 3(b) shows projected semantic activation: the middle segment exhibits significantly higher activation and fluctuation, reflecting greater model sensitivity. This reveals an uneven distribution of semantic information during cross-modal alignment—diverse vocabulary in the central segment drives stronger attention, while outer segments contribute less. Together, the figures demonstrate the temporal dynamics of lyrics and their guiding role in fusion, highlighting the importance of semantically dense regions for prediction and providing a basis for dynamic attention weighting.

3.3 Dynamic generation of cross-modal attention weights

3.3.1 Cross-modal query - construction of key-value structure and attention mapping

To achieve dynamic semantic interaction between audio and lyrics, a cross-modal attention mechanism guided by acoustic representation is constructed. In this mechanism, the acoustic feature sequence output by a deep convolutional network serves as the query, while the projected and positionally encoded sequence of lyric semantic vectors serves as the key and value, respectively. The choice of audio as the query stems from the cognitive mechanism that musical emotion is primarily driven by acoustic signals. This allows the model to actively retrieve relevant semantic segments based on hearing, which is more consistent with the human perceptual model of "sound-guided semantic understanding" during listening. This asymmetric allocation of modal roles enables the model to actively retrieve lyrics semantically relevant to the current audio segment based on auditory signals, thereby achieving selective focus on linguistic information based on acoustic context. Specifically, the acoustic feature vector at each time step is linearly transformed to generate a query vector, while each semantic unit in the lyric sequence is mapped to a corresponding key and value vector, forming a computable cross-modal matching space.

Under this structure, the model can automatically measure the strength of the semantic relevance of different lyric fragments to the current audio state, rather than relying on fixed weights or simple splicing for fusion. The dot product operation between the query and the key is used to evaluate the degree of match, reflecting the extent to which the audio content resonates with a specific semantic unit. This process breaks through the limitations of the equal status of modalities in traditional fusion methods, and is closer to the human emotion perception mechanism when listening to music, which is dominated by hearing and assisted by language. By using audio as an active retrieval signal, the system can accurately locate lyric fragments with emotion-guiding effects in complex emotional expressions, such as emotional climax words or metaphorical expressions in the chorus, thereby enhancing the ability to distinguish ambiguous emotional states.

$$A_{ij}=\frac{\exp\left(\mathrm{q}_i^{\mathrm{T}}\mathrm{k}_j/\sqrt{d_k}\right)}{\sum_{j'}\exp\left(\mathrm{q}_i^{\mathrm{T}}\mathrm{k}_{j'}/\sqrt{d_k}\right)} \quad (5)$$

Where A_{ij}represents the attention weight of the i-th audio time step on the j-th lyric unit, $\mathrm{q}_i^{\mathrm{T}}$is the query vector generated at that time step, $\mathrm{k}_j$is the key vector for the corresponding lyric, $d_k$and is the dimension of the key vector, which is used to scale the dot product result to stabilize the gradient. This formula defines the normalized calculation method for cross-modal association strength and constitutes the core mechanism for dynamic weight generation.

3.3.2 Multi-head attention collaboration and joint representation generation

To further enhance the expressive power of cross-modal association modeling, a multi-head attention mechanism is introduced to execute the query-key-value interaction process in parallel in multiple subspaces. Each attention head independently learns a set of linear projection parameters to map the original features to different low-dimensional subspaces, thereby capturing different types of semantic alignment patterns, such as emotional polarity correspondence, rhythmic synchronization semantic triggering, or auditory echoes of metaphorical words. Each head calculates the attention weight and weights the aggregate value vector to obtain multiple local fusion representations. The outputs of all heads are then concatenated along the feature dimension and restored to a unified latent space dimension through a trainable linear transformation. This multi-perspective modeling strategy enhances the system's adaptability to complex emotional coupling relationships and prevents a single attention head from falling into local optimality or over-focusing on specific vocabulary patterns.

The final output cross-modal response sequence is strictly aligned with the original audio in the time dimension, and each vector is a joint representation of the corresponding acoustic segment fused with its most relevant semantic information. This representation not only preserves the temporal structure of the audio, but also

injects context-sensitive language information regulated by attention weights, achieving fine-grained information complementarity. The entire process is fully differentiable and supports end-to-end training, enabling the model to adaptively optimize cross-modal association strategies based on task objectives. Through a dynamic weighting mechanism, the system can still stably identify dominant emotional cues when faced with music of diverse styles or blurred emotional boundaries, thereby improving overall recognition robustness.

$$\mathbf{z}_i=\text{Linear}\left(\left[\text{head}_1\left(\text{q}_i\right),\ldots,\text{head}_\text{h}\left(\text{q}_i\right)\right]\right) \quad (6)$$

Where, $\mathbf{z}_i$is the fusion output vector of the i-th time step, head_hrepresents hthe weighted value aggregation result of the i-th attention head, and Linearis the output projection function. This operation completes the integration of multi-subspace information to form the final cross-modal response.

3.4 Sentiment classification probability distribution output

3.4.1 Global vector compression of time series fusion representation

The eight emotion categories defined by the EmoMusic dataset are anger, joy, sadness, fear, calmness, excitement, disgust, and anticipation. These encompass basic emotions and common musical emotional states, supporting modeling for multi-label classification tasks. To transform the temporal joint representation generated by cross-modal attention into a fixed-dimensional input suitable for classification tasks, a global average pooling operation along the temporal dimension is employed to compress the high-dimensional sequence. This process takes the arithmetic mean of each feature channel along the time axis, mapping the variable-length temporal response sequence into a compact global vector. This strategy not only effectively eliminates the model's dependence on input length, improving its adaptability to music clips of varying lengths, but also preserves the overall activation strength of each feature channel, reflecting the cumulative effect of the overall emotional tendency of the music. Because the cross-modal attention mechanism performs fine-grained alignment in the preceding stage, the pooled vector naturally integrates the synergistic semantic responses of the audio and lyrics throughout the entire piece of music, forming a comprehensive description with emotional consistency.

This global vector serves as the basic input for subsequent nonlinear transformations, carrying the complete information flow from local perception to high-level semantic fusion. Its construction process does not require the introduction of additional parameters, is computationally efficient, and exhibits translation invariance, enabling dimensionality reduction without sacrificing key emotional statistical properties. Especially when faced with music with complex rhythmic changes or non-stationary emotional evolution, average pooling, through integration operations over the time dimension, suppresses noise interference caused by local fluctuations and enhances the model's ability to capture dominant emotional states. Compared to maximum pooling, this compression method places greater emphasis on overall semantic balance, avoiding emotional misjudgments caused by the dominant representation of individual highly activated segments, thereby improving the stability and interpretability of the output distribution.

$$\mathbf{v}=\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t \quad (7)$$

Where $\mathbf{v}$is the compressed global feature vector, $\mathbf{z}_t$represents the fused representation of the tth time step, and $T$is the total length of the sequence. This formula achieves the transformation from temporal dynamic response to static semantic summary, providing a structurally consistent input basis for classification decisions.

3.4.2 Multi-layer nonlinear mapping and sentiment probability normalization

After obtaining the global feature vector, discriminative features are further refined through stacking nonlinear transformation layers. First, the vector is fed into a fully connected layer, which performs linear projection and introduces the ReLU (Rectified Linear Unit) activation function to enhance the model's ability to model high-order feature combinations. The nonlinear nature of the activation function enables the network to learn the complex boundary relationships between emotion categories and adapt to the real-world challenges of category overlap and continuous transitions in the eight-dimensional emotion space. To prevent overfitting, especially when training data is limited or there is subjective bias in emotion labeling, a dropout mechanism is introduced at the output of the fully connected layer. This randomly sets some neuron outputs to zero with a certain probability, forcing the network to form redundant and robust internal representations.

The regularized features are then fed into the final output layer, which maps the hidden states to the logits space of eight predefined emotion categories, with each dimension corresponding to an unnormalized score for a category of emotional tendency. To generate an interpretable probability distribution, the logits are normalized using the Softmax function so that their sum is 1, forming a valid probability output. This distribution not only reflects the model's judgment of the most likely emotion category for the current music, but also retains confidence information for suboptimal options, supporting soft decision-making mechanisms for multi-label or ambiguous emotions. The entire classification path is fully embedded in the end-to-end training framework, and gradients can be back-propagated to all previous modules, driving the coordinated optimization of cross-modal fusion and feature extraction components to ensure semantic consistency between the final output and the multimodal input.

$$p_c=\frac{\exp(s_c)}{\sum_{k=1}^{8}\exp(s_k)} \quad (8)$$

Where $p_c$represents the predicted probability of emotion category c, is $s_c$the unnormalized score for the corresponding category, and the denominator is the exponential sum of the scores for all eight categories. This formula completes the final mapping from discriminant

scores to probability space, achieving quantifiable output of emotion categories.

4 Method effectiveness evaluation

4.1 Experimental data

Experiments used three public datasets: DEAM (1,393 songs, 44.1kHz, 30-second segments with valence/arousal annotations), EmoMusic (1,000 tracks across genres with eight basic emotion labels), and GTZAN subset (500 samples evenly sampled across genres for genre generalization). Lyrics were precisely aligned to audio using dynamic time warping based on syllable boundaries and singing onset, with non-sung content removed; alignments were manually verified by two experts (mean deviation <±150ms) ensuring frame-level synchronization. Data splits followed temporal independence with no overlap, and the test set was fixed for consistent evaluation. Acoustic features were standardized to 96×198 Mel-spectrograms, text encoded into 512D BERT vectors, and all inputs normalized. This setup ensures reliable multi-scenario validation with balanced emotional diversity, annotation representativeness, and alignment accuracy.

4.2 Time-frequency feature analysis for music emotion recognition

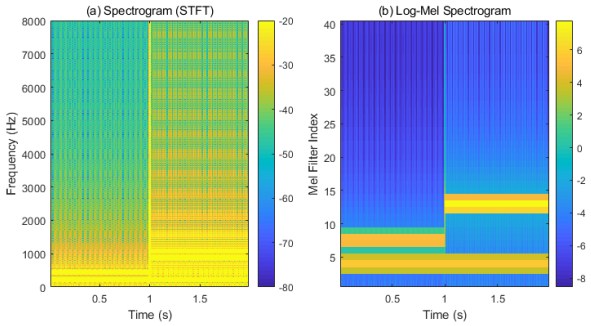

Figure 4: Time-frequency characteristics of music emotions

Figure 4(a) shows the time-frequency energy distribution after pre-emphasis and STFT: early energy is concentrated in low-mid frequencies with clear harmonics, shifting to denser high-frequency clusters over time, reflecting an evolution from calm to tense acoustics. Pre-emphasis suppresses low-frequency dominance and enhances high-frequency cues (e.g., onset, attack), improving separability of rhythmic and timbral changes. Figure 4(b) shows the log Mel spectrum, which reweights energy by perceptual scale and compresses dynamic range. Low-frequency filters capture fundamental and harmonic envelopes in detail, while increasing high-index filter responses indicate enhanced "brightness/tension." Together, these show that perceptual frequency focusing and temporal resolution synergistically stabilize emotion cue extraction from global trends, reducing interference from style variations and local noise on classification boundaries.

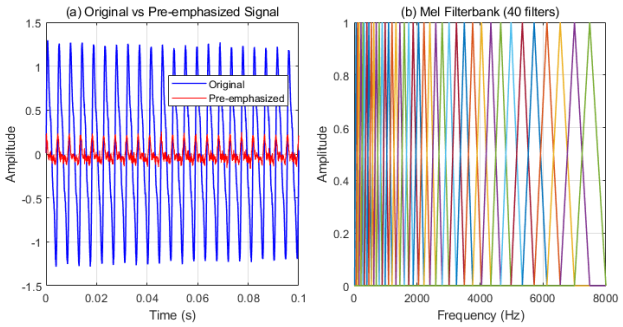

Figure 5: Audio waveform and Mel filter response

Figure 5(a) compares the original and pre-emphasized signals in the time domain. The original waveform shows smooth periodic oscillations with energy concentrated in low frequencies. After pre-emphasis, high-frequency fluctuations are denser and more pronounced, enhancing fine oscillation details. This reflects the filter's attenuation of low frequencies and amplification of high frequencies, highlighting rapidly changing signal components that are otherwise masked. It improves spectral balance and provides a clearer high-frequency structure for feature extraction. Figure 5(b) shows the Mel filter bank's frequency response. The 40 triangular filters are non-linearly spaced along the Mel scale—dense at low frequencies (25–1000 Hz), sparse at high frequencies—mimicking human auditory sensitivity. This preserves low-frequency detail while reducing high-frequency redundancy in the perceptual space. Combined with pre-emphasis, it synergistically enhances both stable low-frequency structures and high-frequency details, yielding rich, perceptually consistent input for downstream models.

4.3 Comparison of recognition accuracy under noise interference

To evaluate the model's robustness in non-ideal auditory environments, a noisy test set was constructed, simulating different signal-to-noise ratio (SNR) conditions by superimposing Gaussian white noise. These conditions included five noise levels: 0dB, 5dB, 10dB, 15dB, and 20dB. At each noise level, the average accuracy and macro-average F1 score of the proposed cross-modal Transformer architecture and comparison models (CNN, CRNN, and traditional Transformer) on an eight-dimensional sentiment classification task were calculated. The evaluation process maintained the same input preprocessing and feature extraction procedures, ensuring that noise was the only variable. By systematically analyzing the performance degradation of each model as the SNR decreased, the authors quantified their semantic preservation capabilities and cross-modal compensation effects under acoustic interference, revealing the differences in sensitivity of different architectures to noise perturbations.

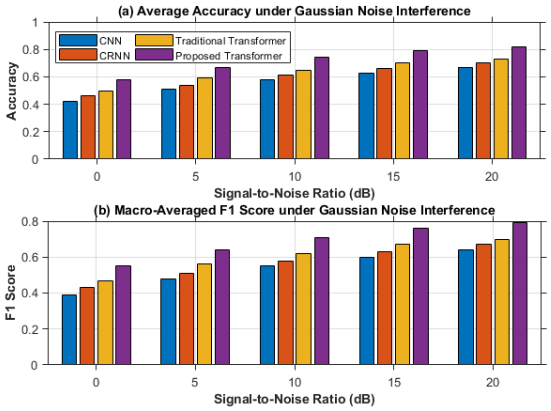

Figure 6: Average accuracy and macro-average F1 score

Figure 6 compares the performance of the proposed cross-modal Transformer model with that of a CNN, a CRNN, and a traditional Transformer in terms of average accuracy and macro-average F1 score under different signal-to-noise ratio (SNR) conditions. As the SNR decreases from 20dB to 0dB, the recognition performance of all models decreases to varying degrees, reflecting the interference effect of Gaussian noise on emotion feature extraction. Notably, the proposed model exhibits superior robustness in the low SNR range, with a significantly gentler performance degradation. In particular, at 0dB, it maintains relatively high discrimination, with an average accuracy and macro-average F1 score of 0.58 and 0.55, respectively. This advantage stems from the dynamic enhancement of semantic information by the cross-modal attention mechanism. Even when the acoustic signal is severely degraded, the model is able to compensate for auditory cues masked by noise through contextual guidance from the lyrics' semantics, achieving complementary support between modalities. In contrast, audio-only CNNs and CRNNs, lacking high-level semantic support, experience significant performance degradation when noise intensifies. While the traditional Transformer possesses some sequence modeling capabilities, it lacks a selective interaction mechanism between modalities, limiting its interference tolerance. In addition, the stable performance of the proposed model on the macro-average F1 shows that it has better balanced recognition of various emotion categories and does not produce systematic bias for specific emotions due to the introduction of noise.

4.4 Impact assessment of imbalanced sentiment categories

To examine the model's generalization ability under skewed class distributions, we constructed unbalanced test subsets of six emotion categories (anger, joy, sadness, fear, calmness, and excitement). Each subset was dominated by one category, accounting for 70%, with the remaining 30% evenly distributed among the other five categories. Under this setting, we evaluated the proposed cross-modal Transformer against comparison models (CNN, CRNN, and traditional Transformer) in terms of weighted F1 score and average precision. By repeatedly testing across the six dominant categories, we systematically analyzed each model's sensitivity to the minority class and its robustness to the majority class.

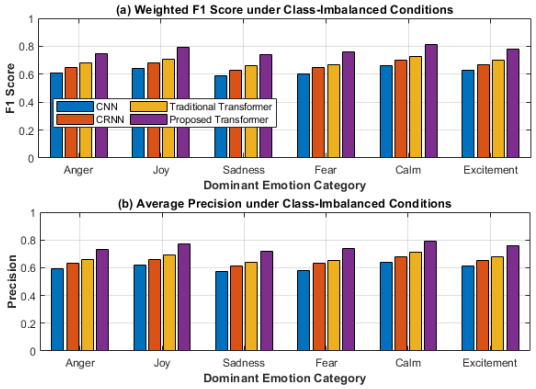

Figure 7: Weighted F1 score and average precision

Figure 7 shows the performance differences between the proposed model and the comparison methods in terms of weighted F1 scores and average precision under test conditions with a severely imbalanced distribution of emotion categories. The six dominant emotions (anger, joy, sadness, fear, calmness, and excitement) each account for 70% of the total, simulating the uneven distribution of emotion annotations in real-world scenarios. The results show that the proposed cross-modal Transformer maintains high recognition consistency across all emotion categories, with significantly less performance fluctuation than the comparison models, achieving weighted F1 scores of 0.74-0.81 and average precisions of 0.72-0.79, demonstrating its greater adaptability to data distribution bias. This advantage stems from the cross-modal attention mechanism's ability to dynamically weight semantic cues. In environments dominated by the majority class, the model leverages the semantic stability of the lyric context to mitigate discriminative bias caused by sampling bias in acoustic features, thereby suppressing overfitting to high-frequency categories. In contrast, unimodal architectures such as CNNs and CRNNs, due to their reliance on local acoustic patterns, are susceptible to interference from dominant class features when minority emotions are present, resulting in reduced precision. While traditional Transformers possess sequence modeling capabilities, they lack a modal complementarity mechanism, making them unstable in categories with sparse semantics or implicit emotional expression (e.g., fear and calmness). Furthermore, the proposed model exhibits minimal performance variation across different emotion categories, indicating that its classification decisions are not systematically biased by the distribution structure of the training data and exhibits good generalization and balance. Overall, these results demonstrate the effectiveness of semantically guided fusion in addressing category imbalance and highlight the profound value of multimodal collaboration in improving model robustness.

4.5 Verification of cross-dataset generalization ability

Using the DEAM dataset as the training domain, the trained model was directly deployed on two external test sets—the sentiment subsets of EmoMusic and GTZAN—without any fine-tuning or adaptation. This zero-shot transfer setup effectively simulates the domain shift issues that arise in real-world applications due to differences in musical style, annotation scale, and recording conditions. During the testing phase, the model's Area Under the Curve (AUC) and Cohen's Kappa coefficient were calculated on the two target datasets. The AUC reflects the classifier's overall discriminative ability under multiple thresholds, while the Cohen's Kappa measures the consistency between the predictions and human annotations, while also correcting for the effects of random agreement.

Table 2: Verification of generalization ability across datasets

Model	Test Domain	AUC	Kappa
CNN	EmoMusic	0.763	0.318
	GTZAN	0.718	0.263
CRNN	EmoMusic	0.674	0.351
	GTZAN	0.612	0.289
Traditional Transformer	EmoMusic	0.698	0.376
	GTZAN	0.635	0.304
Proposed Transformer	EmoMusic	0.832	0.452
	GTZAN	0.887	0.387

4.6 CCC and subjective agreement assessment

To evaluate the reliability of the model's sentiment regression under conditions of low annotated resources, a progressive data sparsity experiment was designed, using 30%, 50%, and 70% of the annotated samples for training, respectively, while maintaining the full test set. In the sentiment dimension prediction task, the Concordance Correlation Coefficient (CCC) was used to measure the statistical consistency between the model output and human annotations. This comprehensive assessment of correlation and bias outperforms the traditional Pearson correlation coefficient. The Inter-annotator Consistency Ratio (ICR) was also introduced as a reference benchmark to quantify the average consistency level within the human annotation community.

Table 3: CCC and subjective consistency evaluation

Model	Label Rate	CCC	ICR
CNN	30%	0.512	0.584
	50%	0.598	0.584
	70%	0.653	0.584
CRNN	30%	0.541	0.584
	50%	0.627	0.584
	70%	0.679	0.584
Traditional Transformer	30%	0.568	0.584
	50%	0.643	0.584
	70%	0.691	0.584
Proposed Transformer	30%	0.635	0.584
	50%	0.712	0.584
	70%	0.758	0.584

5 Discussion

The cross-modal Transformer architecture proposed in this study demonstrates superior performance across multiple challenging tasks. Its core advantage stems from its dynamic attention fusion mechanism. Compared to the dual-branch independent encoding or simple concatenation strategies commonly employed in related work, this approach employs an asymmetric, semantically guided interaction model by using audio features as queries and lyric vectors as keys and values. This design mimics the human process of emotional perception and cognition, where "auditory leads, language assists," enabling the model to proactively retrieve the most relevant semantic fragments based on the current acoustic state, rather than passively and equally fusing modalities. For example, compared to noise-resistant multimodal Transformers, which typically employ a symmetric cross-

attention structure, this approach is computationally complex and may introduce redundant interactions. Our unidirectional guidance mechanism, while maintaining performance, effectively reduces the number of parameters and computational overhead, achieving a better trade-off between noise robustness and efficiency.

In terms of cross-dataset generalization, our approach demonstrates significantly superior stability compared to CNNs, CRNNs, and traditional Transformers. We believe the fundamental reason for this phenomenon lies in the fact that lyric semantics provide a more universal "emotional anchor" that transcends specific acoustic representations. The DEAM, EmoMusic, and GTZAN datasets exhibit significant domain differences in musical style, recording quality, and annotation protocols, resulting in a shift in the distribution of underlying acoustic features. However, the semantic themes carried by lyrics (such as "loneliness" and "joy") are more consistent across domains. Through the dynamic attention mechanism, the model is able to identify acoustic patterns corresponding to these core semantics across different datasets, thereby learning a semantically driven joint representation that is insensitive to domain variations. This explains why our model maintains high AUC and Kappa coefficient in zero-shot transfer scenarios.

Further, the innovation of this architecture lies in its approach to temporal alignment. Unlike graph neural networks that explicitly model temporal alignment, this paper directly incorporates temporal synchronization information into the feature space by injecting absolute positional encodings into the projected word vectors and combining them with frame-level alignment preprocessing. This approach simplifies the model structure and avoids additional graph optimization steps. Heatmap analysis also confirms that it effectively captures the repetitive structure and temporal continuity of lyrics, providing a precise spatiotemporal basis for cross-modal attention. In summary, this paper not only reports performance improvements but, more importantly, reveals the inherent mechanism by which semantic information enhances model robustness and generalization, providing a new perspective for future research in multimodal affective computing.

6 Conclusion

This paper proposes a deep learning architecture that incorporates a cross-modal attention mechanism to address the core challenges of high subjectivity in emotion annotation and insufficient utilization of multimodal information in music emotion recognition. By constructing a Transformer fusion framework using audio features as queries and lyrics semantics as keys, this framework achieves dynamic selective attention of acoustic representations to key language segments, enhancing the fine-grained and context-sensitive semantic alignment between modalities. Experiments demonstrate that the proposed model outperforms traditional CNNs, CRNNs, and Transformers in robustness and generalization under various complex scenarios, including noise interference, class imbalance, cross-dataset transfer, and low-labeled resources. In particular, the proposed method significantly approaches the inter-annotator agreement level in cross-domain transfer and subjective consistency assessment, validating the effectiveness of semantic-guided fusion in modeling emotional subjectivity. This research not only provides a differentiable, end-to-end fusion paradigm for multimodal emotion computing but also reveals the cognitive compensation mechanism of linguistic symbols in auditory emotion decoding, laying a technical foundation for the reliable deployment of intelligent music understanding systems in complex real-world environments. The research results can be widely used in intelligent music recommendation, digital psychological intervention and human-computer empathy interaction systems, providing more accurate emotion perception capabilities in personalized content services and promoting the transformation of emotional computing technology into real-world scenarios.

References

[1] Gómez-Cañón JS, Cano E, Eerola T, et al. Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications[J]. IEEE Signal Processing Magazine, 2021, 38(6): 106-114. https://doi.org/10.1109/msp.2021.3106232

[2] Gómez-Cañón JS, Gutiérrez-Páez N, Porcaro L, et al. TROMPA-MER: an open dataset for personalized music emotion recognition[J]. Journal of Intelligent Information Systems, 2023, 60(2): 549-570. https://doi.org/10.1007/s10844-022-00746-0

[3] Hizlisoy S, Yildirim S, Tufekci Z. Music emotion recognition using convolutional long short term memory deep neural networks[J]. Engineering Science and Technology, an International Journal, 2021, 24(3): 760-767. https://doi.org/10.1016/j.jestch.2020.10.009

[4] Grekow J. Music emotion recognition using recurrent neural networks and pretrained models[J]. Journal of Intelligent Information Systems, 2021, 57(3): 531-546. https://doi.org/10.1007/s10844-021-00658-5

[5] Assuncao WG, Piccolo LSG, Zaina LA M. Considering emotions and contextual factors in music recommendation: a systematic literature review[J]. Multimedia Tools and Applications, 2022, 81(6): 8367-8407. https://doi.org/10.1007/s11042-022-12110-z

[6] Chaturvedi V, Kaur AB, Varshney V, et al. Music mood and human emotion recognition based on physiological signals: a systematic review[J]. Multimedia Systems, 2022, 28(1): 21-44. https://doi.org/10.1007/s00530-021-00786-6

[7] Garg A, Chaturvedi V, Kaur AB, et al. Machine learning model for mapping of music mood and human emotion based on physiological signals[J]. Multimedia Tools and Applications, 2022, 81(4): 5137-5177. https://doi.org/10.1007/s11042-021-11650-0

[8] Li JW, Barma S, Mak PU, et al. Single-channel selection for EEG-based emotion recognition using brain rhythm sequencing[J]. IEEE journal of biomedical and health informatics, 2022, 26(6): 2493-2503. https://doi.org/10.1109/jbhi.2022.3148109

[9] Athavle M, Mudale D, Shrivastav U, et al. Music recommendation based on face emotion recognition[J]. Journal of Informatics Electrical and Electronics Engineering (JIEEE), 2021, 2(2): 1-11. https://doi.org/10.54060/jieee/002.02.018

[10] Cunningham S, Ridley H, Weinel J, et al. Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks[J]. Personal and Ubiquitous Computing, 2021, 25(4): 637-650. https://doi.org/10.1007/s00779-020-01389-0

[11] Yin G, Sun S, Yu D, et al. A multimodal framework for large-scale emotion recognition by fusing music and electrodermal activity signals[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2022, 18(3): 1-23. https://doi.org/10.1145/3490686

[12] Abdullah SMSA, Ameen SYA, Sadeeq MAM, et al. Multimodal emotion recognition using deep learning[J]. Journal of Applied Science and Technology Trends, 2021, 2(01): 73-79. https://doi.org/10.38094/jastt20291

[13] Kamble K, Sengupta J. A comprehensive survey on emotion recognition based on electroencephalograph (EEG) signals[J]. Multimedia Tools and Applications, 2023, 82(18): 27269-27304. https://doi.org/10.1007/s11042-023-14489-9

[14] Zhao S, Jia G, Yang J, et al. Emotion recognition from multiple modalities: Fundamentals and methodologies[J]. IEEE Signal Processing Magazine, 2021, 38(6): 59-73. https://doi.org/10.1109/msp.2021.3106895

[15] Pandeya YR, Lee J. Deep learning-based late fusion of multimodal information for emotion classification of music video[J]. Multimedia Tools and Applications, 2021, 80(2): 2887-2905. https://doi.org/10.1007/s11042-020-08836-3

[16]Li X, Zhang Y, Tiwari P, et al. EEG based emotion recognition: A tutorial and review[J]. ACM Computing Surveys, 2022, 55(4): 1-57.

[17]Pawar MD, Kokate R D. Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients[J]. Multimedia Tools and Applications, 2021, 80(10): 15563-15587. https://doi.org/10.1007/s11042-020-10329-2

[18] Zihan D, Alam N, Islam M M. Deep Learning-Driven Music Emotion Recognition: CNN-BiLSTM Networks with Spatial-Temporal Attention[J]. Journal of Platform Technology, 2025, 13(1): 16-30.

[19] Lin Z, Wang Z, Zhu Y, et al. Text sentiment detection and classification based on integrated learning algorithm[J]. Applied Science and Engineering Journal for Advanced Research, 2024, 3(3): 27-33.

[20] Houssein EH, Hammad A, Ali A A. Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review[J]. Neural Computing and Applications, 2022, 34(15): 12527-12557. https://doi.org/10.1007/s00521-022-07292-4

[21] Jingjing W, Ru H. Music emotion recognition based on the broad and deep learning network[J]. Journal of East China University of Science and Technology, 2022, 48(3): 373-380.

[22] Bakariya B, Singh A, Singh H, et al. Facial emotion recognition and music recommendation system using CNN-based deep learning techniques[J]. Evolving Systems, 2024, 15(2): 641-658. https://doi.org/10.1007/s12530-023-09506-z

[23] Yadav SP, Zaidi S, Mishra A, et al. Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN)[J]. Archives of Computational Methods in Engineering, 2022, 29(3): 1753-1770. https://doi.org/10.1007/s11831-021-09647-x

[24] Topic A, Russo M. Emotion recognition based on EEG feature maps through deep learning network[J]. Engineering Science and Technology, an International Journal, 2021, 24(6): 1442-1454. https://doi.org/10.1016/j.jestch.2021.03.012

[25] Viñán-Ludeña MS, de Campos L M. Discovering a tourism destination with social media data: BERT-based sentiment analysis[J]. Journal of Hospitality and Tourism Technology, 2022, 13(5): 907-921. https://doi.org/10.1108/jhtt-09-2021-0259

[26] Huang F, Li X, Yuan C, et al. Attention-emotion-enhanced convolutional LSTM for sentiment analysis[J]. IEEE transactions on neural networks and learning systems, 2021, 33(9): 4332-4345. https://doi.org/10.1109/tnnls.2021.3056664

[27] Zhang Y, Zhang Y, Wang S. An attention-based hybrid deep learning model for EEG emotion recognition[J]. Signal, image and video processing, 2023, 17(5): 2305-2313. https://doi.org/10.1007/s11760-022-02447-1

[28] Yi Y, Tian Y, He C, et al. DBT: multimodal emotion recognition based on dual-branch transformer: Y. Yi et al[J]. The Journal of Supercomputing, 2023, 79(8): 8611-8633. https://doi.org/10.1007/s11227-022-05001-5

[29] Wu Y, Daoudi M, Amad A. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition[J]. IEEE Transactions on Affective Computing, 2023, 15(1): 157-172. https://doi.org/10.1109/taffc.2023.3263907

[30] Arumugam L, Arumugam S, Chidambaram P, et al. A multi-model deep learning approach for human emotion recognition[J]. Cognitive Neurodynamics, 2025, 19(1): 123. https://doi.org/10.1007/s11571-025-10304-3

[31] Khan M, Gueaieb W, El Saddik A, et al. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion[J]. Expert Systems with Applications, 2024, 245: 122946. https://doi.org/10.1016/j.eswa.2023.122946

[32] Deng Z, Lu Y, Liao J, et al. Sync-TVA: A Graph-Attention Framework for Multimodal Emotion Recognition with Cross-Modal Fusion[J]. arXiv preprint arXiv:2507.21395, 2025.

[33] Liu Y, Zhang H, Zhan Y, et al. Noise-resistant multimodal transformer for emotion recognition[J]. International Journal of Computer Vision, 2024: 1-21.

[34] Wafa A A, Eldefrawi M M, Farhan M S. Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning[J]. Journal of Big Data, 2025, 12(1): 1-62. https://doi.org/10.1186/s40537-025-01264-w

[35] Savchenko A, Savchenko L. Leveraging Lightweight Facial Models and Textual Modality in Audio-visual Emotional Understanding in-the-Wild[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 5778-5788. https://doi.org/10.1109/cvprw67362.2025.00577

[36] Goel A, Singh H, Singh A. Emotion-Aware Speech Translation: A Review[C]//2025 International Conference on Intelligent Control, Computing and Communications (IC3). IEEE, 2025: 533-538. https://doi.org/10.1109/ic363308.2025.10957552