Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077982

ARTICLE

Hierarchical Joint Cross-Modal Attention and Gating Mechanism for Multimodal Sentiment Analysis

Shuqiu Tan, Chunsheng Tan and Yahui Liu*

School of Computer Science and Engineering, Chongqing University of Technology, Chongqing, China

*Corresponding Author: Yahui Liu. Email: liuyh@cqut.edu.cn

Received: 21 December 2025; Accepted: 06 March 2026; Published: 08 May 2026

ABSTRACT: Multimodal sentiment analysis aims to accurately identify emotional states by comprehensively utilizing information from multiple sources such as text, audio, and visual data. However, semantic heterogeneity and temporal differences exist between different modalities, limiting the effectiveness of feature fusion. To address this issue, this paper proposes a hierarchical joint cross-modal attention and gating mechanism (HJCAG) for multimodal sentiment analysis. This method introduces a hierarchical structure, dividing modal interactions into bimodal and trimodal layers to progressively model the semantic relevance between modalities. First, deep features are extracted from text, audio, and visual modalities using pre-trained models to obtain high-dimensional representations of semantics, speech, and facial expressions, which are then aligned to a unified feature space. Second, a joint cross-modal attention module is designed at the bimodal and trimodal levels, calculating cross-attention weights based on the correlation between the joint feature representation and individual modal representations. Explicit modeling of multimodal interaction relationships and semantic alignment fully leverages the complementary information of different modalities. Furthermore, this paper introduces a gating mechanism to adaptively control the contribution weights of each modal feature, reducing redundant information interference and improving the discriminativeness of the fused representation. Finally, the fused global features are input into the emotion classifier to identify emotional states. The proposed method achieves 75.47 ± 0.22% and 69.25 ± 0.37% accuracy and 76.84 ± 0.45% and 68.97 ± 0.41% weighted F1 scores on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database and Multimodal EmotionLines Dataset (MELD), respectively, outperforming mainstream multimodal baseline methods, verifying the effectiveness and robustness of the proposed method in multimodal feature fusion and emotion recognition.

KEYWORDS: Multimodal sentiment analysis; cross-modal attention; hierarchical structure; joint features; gating mechanism

1 Introduction

With the rapid development of social media, short video platforms, and human-computer interaction technologies, affective computing [1] has gradually become one of the important research directions in the field of artificial intelligence. Sentiment analysis, as a core task within this field, aims to identify and understand the emotional states conveyed by humans in multimodal signals such as language, speech, and facial expressions. Compared to traditional unimodal sentiment analysis, multimodal sentiment analysis [2] (MSA) utilizes multi-source information such as text, audio, and visual data simultaneously, thus enabling a more comprehensive and accurate characterization of human emotional features. Therefore, it has broad application prospects in areas such as intelligent customer service, mental health monitoring, virtual human interaction [3], and film and television content analysis.

However, multimodal sentiment analysis still faces many challenges. First, data from different modalities differ significantly in their representation and semantic space. Text modality primarily expresses semantic information, while audio and visual modalities more directly reflect emotional intensity and nonverbal features [4]. This modal heterogeneity complicates feature alignment and semantic fusion. Second, issues such as inconsistent time steps, missing information, and noise interference between modalities lead to insufficient cross-modal information interaction, affecting the accuracy of sentiment recognition. Traditional feature concatenation or simple weighted fusion methods [5] are insufficient to effectively capture inter-modal dependencies and semantic complementarity. Therefore, designing a fusion mechanism that can adaptively model inter-modal correlations and suppress redundant information has become crucial for research.

In recent years, attention mechanisms [6] have demonstrated outstanding performance in natural language processing and multimodal learning. Their core idea is to selectively focus information by dynamically assigning weights to highlight key information and weaken irrelevant features. In multimodal sentiment analysis, cross-modal attention is widely used for modeling intermodal interactions, achieving finer semantic alignment by learning the correlations between text, audio, and visual features [7]. However, most existing studies still have two shortcomings in their design: first, multimodal interactions are usually limited to single-layer or symmetrical structures, failing to distinguish and model the relationships between different levels (such as bimodal and trimodal); second, existing methods generally perform bidirectional attention calculations directly on modal pairs (e.g., text-visual, text-audio), lacking modeling of the interactions between joint modal features and single-modal features, thus limiting the model's ability to capture complex semantic dependencies.

To address the aforementioned issues, this paper proposes a multimodal sentiment analysis framework based on Hierarchical Joint Cross-modal Attention with Gating Mechanism (HJCAG). This method is designed to fully consider the hierarchical nature of intramodal feature aggregation and intermodal semantic interaction. Specifically, it introduces a hierarchical modeling strategy, performing feature fusion at both bimodal (A-V, T-A, T-V) and trimodal (T-A-V) layers. Through this hierarchical design, the model first captures the local semantic dependencies between bimodalities, and then aggregates the global semantic information of the trimodal layers at a higher level, thus achieving a progressive information integration from local to global. In the fusion process at each layer, this paper proposes a Joint Cross-modal Attention (JCA) mechanism. Unlike traditional cross-modal attention, JCA uses joint modal features (such as A-V joint or T-A-V joint) as queries and individual modal features as keys/values, calculating the relevance weights between them to achieve dynamic attention and information interaction between joint features and individual modal features. This design enables the model to more effectively capture high-order semantic dependencies between modalities, strengthening semantic consistency and complementarity. Furthermore, to avoid some modal noise or redundant information affecting the overall discrimination performance, this paper introduces a gating mechanism to adaptively adjust the contributions of each modality, ensuring the effectiveness and stability of the fused features in local and global sentiment representations.

In summary, the main contributions of this paper include the following four aspects:

- A hierarchical multimodal fusion framework is proposed, which models the semantic dependencies between modalities step by step from the bimodal level to the trimodal level.
- A joint cross-modal attention mechanism is designed, with the interaction between joint features and single-modal features as the core, to achieve deep semantic alignment and information enhancement.
- A gating mechanism is introduced to achieve adaptive adjustment of modal weights, effectively suppressing the interference of noisy modalities on feature fusion and improving the stability and generalization performance of the model.

- Experimental results on two widely used benchmark datasets show that the proposed method consistently outperforms the mainstream baseline models, improving accuracy by 1.52% and 1.70%, respectively, and weighted F1 scores by 2.76% and 2.67%, respectively, compared to the closest methods.

2 Related Works

Modal interaction and feature fusion are core components of multimodal sentiment analysis, and their performance directly impacts sentiment recognition accuracy. Based on the timing and method of fusion, existing methods are mainly divided into early fusion and late fusion. Early fusion [8] directly concatenates representations of different modalities at the feature layer, resulting in a simple structure. However, due to significant differences in feature distribution between modalities, it is prone to introducing noise and lacks semantic alignment constraints. Late fusion [9] fuses prediction results after each modality independently completes sentiment discrimination. While this alleviates the feature mismatch problem, it struggles to fully exploit collaborative information between modalities. Therefore, in recent years, researchers have proposed more complex and efficient fusion mechanisms.

For example, The Tensor Fusion Network (TFN) introduced by Zadeh et al. [10] employs third-order tensor representations to explicitly model full interaction patterns among modalities. To alleviate the computational burden of high-order tensors, Liu et al. [11] developed the Low-rank Multimodal Fusion (LMF) framework based on tensor decomposition techniques. Building upon Transformer architectures, the MulT model introduced by Tsai et al. [12] models inter-modal dependencies through directional cross-modal attention and incorporates a self-attention mechanism to handle asynchronous multimodal sequences. However, subsequent analysis by Ma et al. [13] pointed out that MulT primarily captures interactions at a single feature scale, without sufficiently exploiting fine-grained semantic information. To address this limitation, they further proposed the MCMulT, enabling semantic modeling across hierarchical feature representations. Considering heterogeneity in multimodal sequences, Sun et al. [14] introduced the EMT-DLFR framework to integrate global contextual information for enhanced global–local interaction modeling. In addition, UniMSE designed by Hu et al. [15] performs multimodal fusion at both syntactic and semantic levels and integrates contrastive learning strategies to refine fused feature representations. Shi and Huang [16] proposed MultiEMO, which effectively integrates multimodal cues by capturing cross-modal mapping relationships between text, audio, and visual modalities through a bidirectional multi-head cross-attention layer.

Furthermore, gating mechanisms play a crucial role in modality weight control. By learning the contribution of each modality to the final emotion decision through gating units, the model can suppress irrelevant or noisy features, thereby achieving adaptive fusion. For example, Arevalo et al. [17] proposed a gated multimodal unit (GMU) to balance the information flow between different modalities through gating functions. Rahman et al. [18] proposed MAG-BERT, which further introduced a modality adaptive gating strategy into the BERT framework, achieving superior fusion performance.

Inspired by the above-mentioned research, this paper proposes the hierarchical joint cross-modal attention and gating mechanism model. This model achieves improvements over existing work in hierarchical structure modeling, cross-modal semantic association capture, and adaptive control of modal weights, providing a more efficient and robust fusion strategy for multimodal sentiment analysis. Details will be presented in the Section 3.

3 Methodology

The overall structure of the HJCAG proposed in this paper is shown in Fig. 1. The design principles and implementation methods of each part will be introduced below.

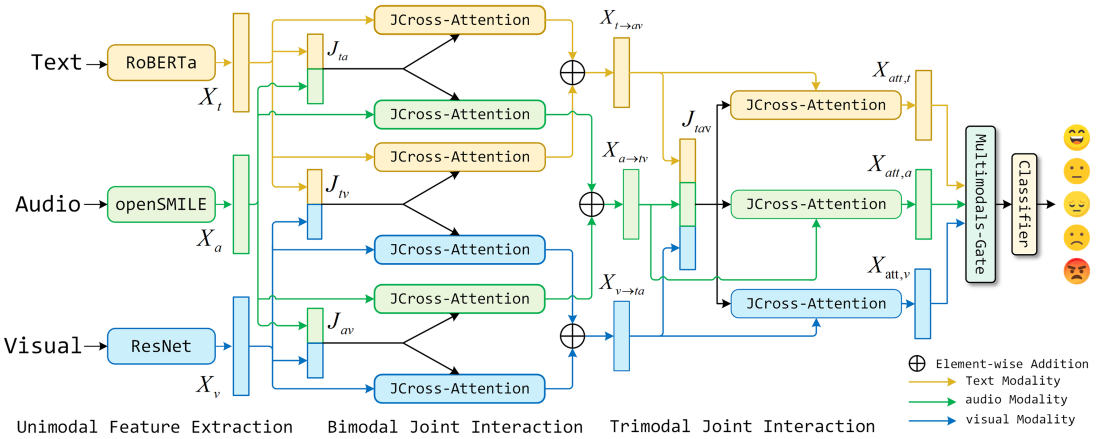

Figure 1: Overview of the proposed framework.

3.1 Unimodal Feature Extraction

3.1.1 Text Modality

This paper uses the pre-trained language model RoBERTa [19] as the feature extractor for text modalities. RoBERTa is based on a multi-layer Transformer encoder structure and is optimized on the basis of the BERT model, which can more fully model the semantic information of the text. To adapt to the sentiment recognition task, we fine-tune RoBERTa end-to-end and use the hidden state corresponding to the [CLS] label in its last layer as the sentiment representation of the text. The final text feature vector has a dimension of 1024.

3.1.2 Audio Modality

In terms of acoustic modal feature extraction, this paper adopts the openSMILE toolkit [20]. openSMILE is a general feature extraction framework for audio signal processing, which supports flexible configuration of various modular feature extraction components through a scriptable command line. Based on the original acoustic features extracted by openSMILE, the dimensionality is further compressed through a fully connected layer. The acoustic feature dimension of the IEMOCAP dataset is reduced to 1582, and the feature dimension of the MELD dataset is reduced to 300.

3.1.3 Visual Modality

For visual modal feature extraction, this paper employs the ResNet-50 network as the visual feature extractor. ResNet-50 [21] is a convolutional neural network based on a deep residual learning mechanism. By introducing residual connections, it effectively alleviates the gradient degradation problem in deep network training, thereby improving feature representation capabilities. This paper utilizes a pre-trained ResNet-50 to encode visual information in video frames and extracts its high-level semantic features as visual representations with a dimension of 512, which are then used for subsequent multimodal fusion and emotion recognition tasks.

Although we adopt widely-used feature extractors such as ResNet-50 and openSMILE to ensure fair comparison with existing multimodal sentiment analysis methods, the proposed HJCAG framework is backbone-agnostic. It can be readily extended to incorporate modern self-supervised representations.

3.2 Hierarchical Joint Cross-Modal Attention

3.2.1 Bimodal Joint Interaction

While bimodal feature fusion can be achieved through unified multimodal training, studies have found that multimodal performance typically degrades compared to single-modal fusion [22]. This is attributed to a number of factors, such as the differences in learning dynamics between T and V modalities [22], different noise topologies, the inclusion of more or less task-related information in certain modal flows, and specialized input representations [23]. Therefore, we train DL models separately for each modality to extract single-modal features, which are then further fed into the HJCA fusion model for pairwise modality fusion. To reliably fuse these modalities by effectively leveraging the complementary relationships between bimodalities, we rely on a cross-attention-based fusion mechanism to efficiently encode intermodal information while preserving intramodal features. While cross-attention is typically applied to features of individual modalities, we explore it within a joint framework. Specifically, our joint feature representation is obtained by concatenating pairwise modal features to focus on joint single-modal features, for example, concatenating A and V features to focus on individual A and V features. By using joint representations, each modality's features focus on itself and other modalities, thus helping to capture semantic intermodal relationships between the two modalities. Furthermore, using combined feature representations in the cross-attention module can significantly reduce heterogeneity between the two modalities, thereby further improving system performance.

Given an input video subsequence S of L frames, the audio, visual, and text data are preprocessed and input into the corresponding encoders, as shown in Fig. 1. The feature representations of the audio, visual, and text modalities are represented as $\mathrm{X}_t = \left\{x_t^1, x_t^2, \ldots, x_t^L\right\} \in \mathbb{R}^{d_t \times L}$, $\mathrm{X}_a = \left\{x_a^1, x_a^2, \ldots, x_a^L\right\} \in \mathbb{R}^{d_a \times L}$, and $\mathrm{X}_v = \left\{x_v^1, x_v^2, \ldots, x_v^L\right\} \in \mathbb{R}^{d_v \times L}$, where d_t, d_a, and d_t are the dimensions of the audio, visual, and text features, respectively. x_t^L, x_a^L, and x_v^L represent the feature vectors of each frame in the audio, visual, and text modalities. Given the audio, visual, and text feature representations X_t, X_a, and X_v, by concatenating the feature vectors of each pair of modalities, to avoid simple information duplication, the concatenated joint features are projected through a fully connected layer with non-linear activation, joint feature representations at three levels—text-audio, text-visual, and audio-visual—can be obtained, as shown below:

$$J_{ta} = \mathrm{FC}\left(\left[X_t; X_a\right]\right) \in \mathbb{R}^{d_{ta} \times L} \tag{1}$$

$$J_{tv} = \mathrm{FC}\left(\left[X_t; X_v\right]\right) \in \mathbb{R}^{d_{tv} \times L} \tag{2}$$

$$J_{av} = \mathrm{FC}\left(\left[X_a; X_v\right]\right) \in \mathbb{R}^{d_{av} \times L} \tag{3}$$

where $d_{ta} = d_t + d_a$, $d_{tv} = d_t + d_v$, $d_{av} = d_a + d_v$ represent the feature dimensions of the joint features, and L represents the number of non-overlapping fixed-size clips uniformly sampled from S.

Although the concatenated features are transformed through a fully connected layer, this operation does not perform dimensional compression. Instead, it serves as a modality alignment transformation in a shared embedding space, The projection matrix preserves the original dimensions of the concatenated features, ensuring that no explicit information reduction is introduced.

Subsequently, the joint features J_{ta}, J_{tv}, J_{av} are input into the joint Cross-Modal attention framework (JCA) for each modality, as shown in Fig. 2, to focus on the feature representations of each modality. This helps to simultaneously encode relationships within the same modality and across modalities to obtain the cross-correlation matrix after a single modality interacts with the other two modalities,it is worth noting that the joint representation is only used as an interaction context to compute attention relevance, rather than being directly injected into the modality features. It is given by the following formula:

$$C_t' = \tanh\left(\frac{X_t^T W_{jt\to a} J_{ta}}{\sqrt{d}}\right) + \tanh\left(\frac{X_t^T W_{jt\to v} J_{tv}}{\sqrt{d}}\right) \tag{4}$$

where $W_{jt\to a} \in \mathbb{R}^{d_t \times d_{ta}}$ represents the learnable weight matrix across X_t and J_{ta}, and $W_{jt\to v} \in \mathbb{R}^{d_t \times d_{tv}}$ represents the learnable weight matrix across X_t and J_{tv}. Similarly, the cross-correlation matrices for the other two modalities can be obtained as follows:

$$C_a' = \tanh\left(\frac{X_a^T W_{ja\to t} J_{ta}}{\sqrt{d}}\right) + \tanh\left(\frac{X_a^T W_{ja\to v} J_{av}}{\sqrt{d}}\right) \tag{5}$$

where $W_{ja\to t} \in \mathbb{R}^{d_a \times d_{ta}}$ represents the learnable weight matrix across X_a and J_{ta}, and $W_{ja\to v} \in \mathbb{R}^{d_a \times d_{av}}$ represents the learnable weight matrix across X_a and J_{av}.

$$C_v' = \tanh\left(\frac{X_v^T W_{jv\to t} J_{tv}}{\sqrt{d}}\right) + \tanh\left(\frac{X_v^T W_{jv\to a} J_{av}}{\sqrt{d}}\right) \tag{6}$$

where $W_{jv\to t} \in \mathbb{R}^{d_v \times d_{tv}}$ represents the learnable weight matrix across X_v and J_{tv}, and $W_{jv\to a} \in \mathbb{R}^{d_v \times d_{av}}$ represents the learnable weight matrix across X_v and J_{av}.

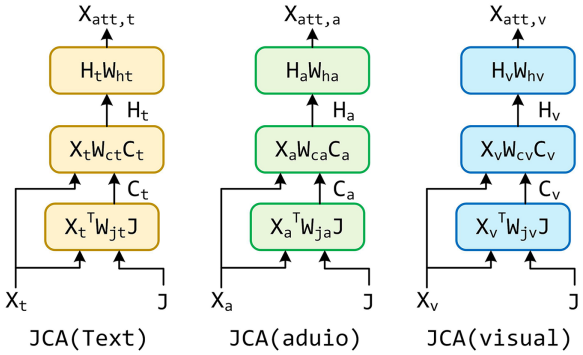

Figure 2: Joint cross-modal attention blocks of text, audio, and visual modalities.

The obtained joint cross-correlation matrix for a single modality is used to calculate attention weights, thereby capturing semantic relevance within the same modality and between different modalities. The joint cross-correlation matrix quantifies the correlation between specific modal feature representations and the global multimodal context. Higher correlation coefficients indicate stronger semantic consistency between corresponding features in the global multimodal context. Based on the joint cross-correlation matrix, the attention map estimates the relative importance of specific modal features. Specifically, for the text modality, the joint correlation matrix C_t' and the corresponding text features X_t are combined using a learnable weight matrix W_{ct}', and then the attention map H_t' is calculated using the ReLU activation function, ensuring that only positively correlated features contribute to the attention weights while suppressing irrelevant or noisy features, as shown below:

$$H_t' = \mathrm{ReLU}\left(X_t W_{ct}' C_t'\right) \tag{7}$$

where $W_{ct}' \in \mathbb{R}^{L \times L}$. Similarly, the attention maps for audio and visual modalities are as follows:

$$H_a' = \mathrm{ReLU}\left(X_a W_{ca}' C_a'\right) \tag{8}$$

$$H_v' = \mathrm{ReLU}\left(X_v W_{cv}' C_v'\right) \tag{9}$$

where $W_{ca}{}' \in \mathbb{R}^{L\times L}$, $W_{cv}{}' \in \mathbb{R}^{L\times L}$. Now we use attention maps to compute the thermal attention features of a single modality, residual connections preserve the original specific modal representations, enabling attention-based refinement without overwriting discriminative unimodal cues:

$$X_{t\to av} = H_t{}' W_{ht}{}' + X_t \tag{10}$$

$$X_{a\to tv} = H_a{}' W_{ha}{}' + X_a \tag{11}$$

$$X_{v\to ta} = H_v{}' W_{hv}{}' + X_v \tag{12}$$

where $W_{ht}{}' \in \mathbb{R}^{L\times L}$, $W_{ha}{}' \in \mathbb{R}^{L\times L}$, $W_{hv}{}' \in \mathbb{R}^{L\times L}$. $X_{t\to av}$, $X_{a\to tv}$ and $X_{v\to ta}$ represent a bimodal joint thermal attention-weighted feature that integrates its own modality and other modalities.

3.2.2 Trimodal Joint Interaction

After bimodal joint interaction, we obtained the relevance and information supplementation of a single modality to the other two modalities. Subsequently, the weighted fusion features were further combined to obtain a text-audio-visual modal joint feature representation containing global semantic information, as shown below:

$$J_{tav} = FC\left(\left[X_{t\to av}; X_{a\to tv}; X_{v\to ta}\right]\right) \in \mathbb{R}^{d\times L} \tag{13}$$

where $d = d_t + d_a + d_v$.

Subsequently, the joint feature J_{tav} is also input into the cross-attention framework of each modality to focus on the feature representations of each modality. This helps to adaptively focus on the key information most relevant to the current emotional state in different modalities, starting from the global joint semantics, to obtain the global attention matrix, which is given by the following equation:

$$C_t = \tanh\left(\frac{X_{t\to av}^T W_{jt\to av} J_{tav}}{\sqrt{d}}\right), C_a = \tanh\left(\frac{X_{a\to tv}^T W_{ja\to tv} J_{tav}}{\sqrt{d}}\right), C_v = \tanh\left(\frac{X_{v\to ta}^T W_{jv\to ta} J_{tav}}{\sqrt{d}}\right) \tag{14}$$

where $W_{jt\to av} \in \mathbb{R}^{d_t\times d}$ represents the learnable weight matrix across $W_{jt\to av}$ and J_{tav}, $W_{ja\to tv} \in \mathbb{R}^{d_a\times d}$ represents the learnable weight matrix across $W_{ja\to tv}$ and J_{tav}, $W_{jv\to ta} \in \mathbb{R}^{d_v\times d}$ represents the learnable weight matrix across $W_{jv\to av}$ and J_{tav}.

Similarly, the joint relevance matrix C_t and the corresponding weighted text features $X_{t\to av}$ are combined using a learnable weight matrix W_{ca}, and then the attention map H_t is computed using the ReLU activation function, as shown below:

$$H_t = \mathrm{Re}LU\left(X_{t\to av} W_{ct} C_t\right), \quad H_a = \mathrm{Re}LU\left(X_{a\to tv} W_{ca} C_a\right), \quad H_v = \mathrm{Re}LU\left(X_{v\to ta} W_{cv} C_v\right) \tag{15}$$

here $W_{ct}, W_{ca}, W_{ct} \in \mathbb{R}^{L\times L}$ are the weight matrices for the learnable text, sound, and visual modalities, respectively. Now, attention maps are used to compute the hot attention features for a single modality:

$$X_{att,t} = H_t W_{ht} + X_{t\to av}, \quad X_{att,a} = H_a W_{ha} + X_{a\to tv}, \quad X_{\mathrm{att},v} = H_v W_{hv} + X_{v\to ta} \tag{16}$$

where $W_{ht} \in \mathbb{R}^{L\times L}$, $W_h \in \mathbb{R}^{L\times L}$, and $W_{hv} \in \mathbb{R}^{L\times L}$ represent the learnable weight matrix of the audio.

This hierarchical design enables the model to first capture the local semantic dependencies between the two modalities, and then fuse the global sentiment information of the three modalities at a higher level, effectively mitigating the negative impact of modal heterogeneity and semantic inconsistency.

3.3 Gating Mechanism

By inputting the features of interest from each modality into the multimodal gating fusion module, the contribution of each modality to the final prediction is dynamically adjusted, preventing the model from being overly biased towards a single modality and ensuring balanced integration of multimodal information. This design enhances the robustness of multimodal fusion and improves the model's generalization performance in fine-grained category sentiment classification tasks. As shown below:

$$[g_t; g_a; g_v] = Soft\max([W_t X_{att,t}; W_a X_{att,a}; W_v X_{att,v}]) \tag{17}$$

$$X_{att} = \sum_{m\in\{t,a,v\}} X_{att,m} \otimes g_m \tag{18}$$

Finally, the multimodal feature representation X_{att} is fed into the classifier to ultimately predict the sentiment category. We use the cross-entropy loss function to estimate the quality of sentiment predictions during training:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\log(\hat{y}_{i,j}) \tag{19}$$

where C represents the number of categories, N represents the number of samples, $y_{i,j}$ represents the true category, and $\hat{y}_{i,j}$ represents the predicted category.

4 Experimental

4.1 Datasets

4.1.1 IEMOCAP Dataset

IEMOCAP [24] is an audiovisual corpus designed for emotion recognition research, comprising approximately 12 h of dyadic conversational recordings. The dataset includes 151 dialogue sessions with a total of 7433 annotated utterances. Each utterance is manually assigned to one of six emotional categories: happy, sad, neutral, angry, excited, and frustrated.

4.1.2 MELD Dataset

MELD [25] is a multimodal, multi-party conversational emotion dataset collected from the television series Friends. It consists of 1433 dialogue segments and 13,708 utterances. Each utterance is annotated with one of seven emotion labels, including anger, disgust, fear, joy, neutral, sadness, and surprise.

The statistics of different emotion types in the IEMOCAP and MELD datasets are shown in Table 1.

Table 1: Emotion label distribution in the IEMOCAP and MELD datasets.

Dataset	Emotion Labels									
	Happy	**Frustrated**	**Excited**	**Neutral**	**Anger**	**Sad**	**Joy**	**Disgust**	**Fear**	**Surprise**
IEMOCAP	648	1849	1041	1708	1103	1084	–	–	–	–
MELD	–	–	–	6436	1607	1002	2308	361	358	1636

Both datasets feature sentiment annotation at the utterance-level. Therefore, we perform utterance-level alignment between the two datasets. For each utterance U, corresponding to a video subsequence S, for the

text modality, we first extract lexical embeddings and then pad or crop them to a fixed sequence length L. For the visual modality, we uniformly sample L frames from S and encode them using a visual backbone network. For the audio modality, we copy or broadcast the extracted descriptors along the temporal dimension to align them temporally with the visual and text sequences to match the length L. This strategy ensures consistency in temporal granularity across modalities.

4.2 Training Details

We adopt the RoBERTa model as the textual feature encoder and fine-tune it using the AdamW optimization strategy. During training, the learning rate is initialized to 2e−5, with a batch size of 16 and a maximum input length of 128 tokens. A linear learning rate schedule with warm-up is employed, where the warm-up phase occupies the first 10% of the total optimization steps. To enhance training robustness and mitigate overfitting, the fine-tuning process is repeated three times, and early stopping is applied based on the validation loss.

Given the pronounced differences in class distribution between the MELD and IEMOCAP datasets, distinct batch sizes are employed to better accommodate their respective imbalance characteristics. Specifically, the batch size is set to 64 for IEMOCAP and 100 for MELD. The model is trained for a total of 200 epochs using the Adam optimizer, with momentum parameters configured as $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and an initial learning rate of 1×10^{-5}. To facilitate stable convergence, the learning rate is reduced by a factor of 0.95 every 10 epochs. Furthermore, L2 regularization is incorporated with a coefficient $\lambda = 1 \times 10^{-5}$. To mitigate overfitting, a dropout rate of 0.1 is applied during training, the length of L is 30, all joint cross-modal attention blocks use a single head and layer normalization is applied after each attention block.

Evaluation metrics: Following previous work [7,26], we report overall accuracy and weighted average F1 score to measure overall performance, and provide F1 score for each emotion category. Furthermore, all reported results were obtained by averaging five-fold cross-validation.

All comparison models used the same backbone architecture and experimental settings.

4.3 Results and Comparison

Experimental evaluations on the IEMOCAP and MELD benchmarks, as reported in Tables 2 and 3, along with comparisons against representative existing approaches. As can be observed, our model obtains the highest weighted-average F1 scores on both datasets and consistently outperforms competing methods across the majority of emotion classes. Such improvements can be mainly due to the proposed architecture, which learns semantic dependencies between modalities in a progressively deeper manner. The bimodal layer focuses on capturing local emotional associations between modal pairs, helping to uncover complementary information from different modalities at a fine-grained level; the trimodal layer further integrates global semantic representations of text, audio, and visual modalities, thereby achieving holistic modeling of complex emotional states and effectively mitigating the impact of modal heterogeneity and semantic inconsistency.

On the IEMOCAP dataset, HJCAG achieves the highest Acc and W-F1 scores, demonstrating overall superior performance. The model shows clear advantages in Sad, Anger, Excited, and Frustrated categories, which are typically minority or semantically overlapping emotions. Although slightly underperforming EmoCaps [31] and SDT [7] in the Happy and Neutral categories, respectively, HJCAG maintains competitive results and exhibits a more balanced performance across emotion classes, indicating reduced bias toward majority categories. On the MELD dataset, HJCAG outperforms all baseline methods in both Acc and W-F1, confirming its robustness under severe class imbalance. Notable gains are observed in low-frequency

emotions such as Fear and Disgust, while strong performance is also maintained for high-frequency categories like Neutral and Surprise. Despite slightly lower performance than CFN-ESA [33] on Joy, HJCAG consistently surpasses other methods across the remaining emotion classes.

Table 2: Performance comparison of different models on the IEMOCAP dataset.

Models	Happy F1	Sad F1	Neutral F1	Anger F1	Excited F1	Frustrated F1	Acc	W-F1
DialogueRNN [27]	32.93	78.04	59.18	63.26	73.62	59.46	63.33	62.83
HAUCL [28]	53.57	82.04	68.61	66.44	75.60	68.23	70.30	70.27
AdalGN [29]	53.04	81.47	71.26	65.87	76.34	67.79	–	70.74
MMTr [30]	–	–	–	–	–	–	72.27	71.91
EmoCaps [31]	74.31	85.47	67.03	65.26	80.14	68.38	73.67	73.01
SMFNM [32]	59.52	81.49	70.06	62.91	74.43	70.15	70.82	70.94
MultiEMO [16]	65.24	83.71	67.47	68.47	76.14	70.26	–	71.98
CFN-ESA [33]	53.67	80.60	71.65	70.32	74.82	68.06	70.78	71.04
SDT [7]	66.19	81.84	74.62	69.73	80.17	68.68	73.95	74.08
HJCAG (Ours)	67.24	86.15	70.43	71.89	81.54	71.67	75.47	76.84

Table 3: Performance comparison of different models on the MELD dataset.

Models	Neutral F1	Surprise F1	Fear F1	Sad F1	Joy F1	Disgust F1	Angry F1	Acc	W-F1
DialogueRNN [27]	76.56	47.64	–	24.65	51.49	–	46.01	58.03	56.98
MMGCN [34]	76.96	49.63	3.64	20.39	53.76	2.82	45.23	–	58.41
DialogueTRM [35]	79.41	55.27	17.39	36.48	60.30	20.18	49.79	65.70	63.80
AdaIGN [29]	79.75	60.53	–	43.70	64.54	–	56.15	–	66.79
HAUCL [28]	–	–	–	–	–	–	–	68.05	66.72
EmoCaps [31]	74.28	64.74	2.14	42.52	62.52	7.05	60.26	64.93	63.88
SMFNM [32]	75.06	57.48	16.83	36.79	62.35	25.04	50.33	62.60	62.42
MultiEMO [16]	79.98	60.28	28.24	41.20	62.86	35.28	53.60	–	66.47
CFN-ESA [33]	80.05	58.78	21.62	41.82	66.50	26.92	54.18	67.85	66.70
SDT [7]	80.19	59.07	17.88	43.69	64.29	28.78	54.33	67.55	66.60
HJCAG (Ours)	81.59	64.88	33.32	46.62	65.13	38.54	58.45	69.25	68.97

We further examine the confusion matrices obtained on both datasets, as illustrated in Fig. 3. The analysis reveals several noteworthy observations. First, the proposed model still encounters difficulties in distinguishing semantically similar emotion pairs, such as happiness vs. excitement and anger vs. frustration in IEMOCAP, as well as surprise vs. joy in MELD. Second, on the MELD dataset, a considerable number of samples from other emotion categories are misclassified as neutral, largely due to the dominance of this class in the dataset. Third, emotions with extremely limited training samples, particularly fear and disgust in MELD, remain challenging to recognize reliably, which hinders effective model learning. Overall, these results indicate that emotion confusion caused by semantic similarity and severe class imbalance continues to be a major obstacle in multimodal emotion recognition.

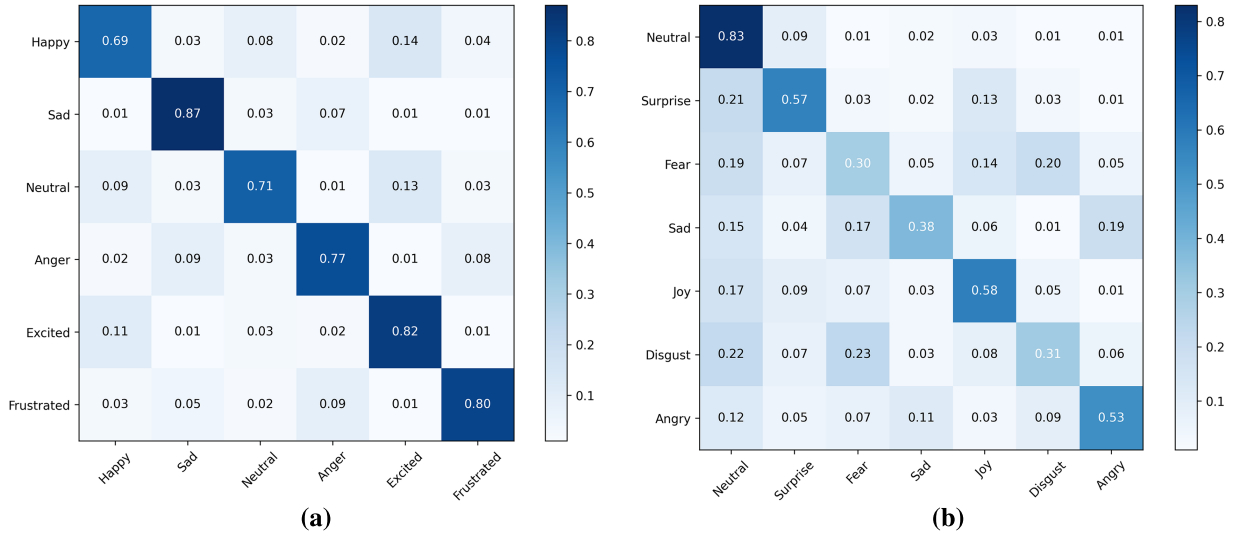

Figure 3: (**a**) Confusion matrix on IEMOCAP dataset (**b**) confusion matrix on MELD.

We observed that the proposed model performs poorly in scenarios involving irony, where the text conveys a positive emotion but its underlying intention is negative. Furthermore, the model occasionally misclassifies samples with subtle emotional differences (e.g., neutral vs. slightly positive), indicating that fine-grained emotional boundaries remain challenging. These observations point to potential directions for future research, such as introducing pragmatic reasoning mechanisms or ambiguity perception fusion strategies.

4.4 Ablation Experiments

To investigate the effectiveness of our proposed hierarchical structure and gating mechanism, we conducted ablation experiments. Table 4 shows the performance under different ablation conditions.

Table 4: Ablation study and modality contribution analysis on the IEMOCAP and MELD datasets.

Models/Modalities	IEMOCAP		MELD	
	Acc	W-F1	Acc	W-F1
HJCAG (Ours)	75.47	76.84	69.25	68.97
w/o bi-modal joint	72.45	72.38	65.34	65.12
w/o tri-modal joint	72.87	73.15	66.57	66.50
w/o multimodal gate	73.14	73.21	67.45	67.63
Text	60.76	61.23	58.82	58.79
Audio	57.77	57.34	48.23	45.81
Visual	40.37	41.72	46.05	43.25
Text + Audio	72.78	72.66	64.45	65.21
Text + Visual	70.01	69.89	64.67	64.83
Audio + Visual	72.35	71.96	63.17	62.64

Bimodal joint layers, trimodal joint layers, and multimodal gating mechanisms are the three key components of our proposed hierarchical joint cross-modal attention. We evaluate the effectiveness of a

component by removing only one component at a time. In Table 4, we observe that: (1) Overall, each ablation method led to a performance decrease to varying degrees, fully demonstrating the necessity of each component of the model. (2) The bimodal joint layer has a significant impact on both datasets. On the IEMOCAP dataset, removing this component leads to a 3.02% decrease in overall recognition accuracy and a 4.46% decrease in the weighted F1 score. On the MELD dataset, removing this component leads to a 3.91% decrease in accuracy and a 3.85% decrease in the weighted F1 score. These results indicate that the feature interactions between the two modalities play a fundamental role in capturing cross-modal correlations and are a core support for subsequent high-level fusion. (3) The trimodal integration layer plays a crucial role. On the IEMOCAP dataset, removing this component leads to a 2.60% decrease in accuracy and a 3.69% decrease in the weighted F1 score. On the MELD dataset, removing this component leads to a 2.68% decrease in accuracy and a 2.47% decrease in the weighted F1 score. This indicates that integrating information from three modalities simultaneously yields more complete emotional cues than any two modalities. (4) Removing the multimodal gating mechanism leads to performance degradation. On the IEMOCAP dataset, accuracy drops by 2.33%, and the weighted F1 score drops by 3.63%. On the MELD dataset, accuracy drops by 1.80%, and the weighted F1 score drops by 1.34%. This verifies that the mechanism can effectively filter noise from different modalities and adaptively adjust its contribution. Therefore, dynamic weight allocation is crucial to ensuring model stability in the presence of noise, inconsistencies in context, or modal conflicts.

Furthermore, removing the bimodal or trimodal joint layer reduces the model to a single-layer structure. As shown in Table 4, the hierarchical structure performs better than the single-layer structure. This is because, although single-layer trimodal attention theoretically possesses strong expressive power, directly modeling the complete trimodal interaction within a single stage may introduce excessive high-order coupling and modality dominance effects into the shared attention space. Although hierarchical interaction uses more interaction attention, increasing model complexity, it decomposes trimodal learning into two progressive stages, achieving local cross-modal alignment before global fusion. This structured design improves representation efficiency and optimization stability.

Effects of different modalities: To demonstrate the effects of different modalities, we remove one or two modalities at a time. From Table 4, we observe: (1) From the results of the single-modal analysis, the text modality performed the best, while the audio and visual single-modal performances were lower, mainly due to the influence of factors such as background noise, changes in speaking style, and blurred facial expressions. (2) Any combination of two modalities significantly outperforms the corresponding single-modal results. For example, both Text+Audio and Audio+Visual show significant improvements, indicating that audio and visual signals can supplement emotional details that are difficult to capture in the text modality. (3) The performance of the bimodal combination is still lower than that of the complete trimodal model, further demonstrating that the three modalities are complementary in emotion expression and that trimodal interaction can provide the most comprehensive emotional information.

The proposed hierarchical joint cross-modal attention and gating (HJCAG) module mainly consists of lightweight fully connected layers, attention projections, and gating networks. Compared with the backbone encoders (e.g., RoBERTa and ResNet-50), the additional parameters introduced by HJCAG are relatively small.

In summary, the ablation experiments fully demonstrate that both bimodal and trimodal joint interaction modules play a crucial role in performance improvement, and the multimodal gating mechanism effectively enhances the model's adaptability to contributions from different modalities. The HJCAG model, through its hierarchical cross-modal attention mechanism and dynamic weight allocation, is able to capture more comprehensive and robust emotional representations.

5 Conclusion

This paper proposes a multimodal sentiment analysis method based on hierarchical joint cross-modal attention and gating mechanisms. This method models the semantic relationships between modalities at bimodal and trimodal levels and explicitly characterizes the high-order interaction relationships between joint features and single-modal features using a joint cross-modal attention mechanism, thereby achieving more effective semantic alignment and information fusion. Simultaneously, the introduced gating mechanism adaptively adjusts the contributions of different modalities, improving the discriminativeness and stability of the fused features. In real-world scenarios, the proposed framework may be challenged by noisy audio, subtle or ambiguous emotional expressions (e.g., sarcasm), and imperfect modality alignment. These factors highlight potential limitations and motivate future work on robustness enhancement and adaptive fusion strategies.

Acknowledgement: The authors would like to thank all contributors and institutions that supported this research.

Funding Statement: This research was supported by the Chongqing Basic Research and Frontier Exploration Project (Chongqing Natural Science Foundation) under Grant No. CSTB2022NSCQ-MSX0918, and the Science and Technology Research Project (Youth) of Chongqing Municipal Education Commission under Grant No. KJQN202301122.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Shuqiu Tan; methodology, Chunsheng Tan; data curation, Chunsheng Tan; writing—original draft preparation, Chunsheng Tan; writing—review and editing, Shuqiu Tan and Chunsheng Tan; visualization, Chunsheng Tan; supervision, Shuqiu Tan and Yahui Liu; funding acquisition, Shuqiu Tan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data presented in this study are openly available in (IEMOCAP) at (https://doi.org/10.1007/s10579-008-9076-6) and (MELD) at (https://doi.org/10.48550/arXiv.1810.02508) reference numbers [24,25].

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion. 2017;37:98–125.
2. Song H, Li J, Xia Z, Yang Z, Du X. Multimodal sentiment analysis based on pre-LN transformer interaction. In: Proceedings of the 6th IEEE Information Technology and Mechatronics Engineering Conference; 2022 Mar 4–6; Chongqing, China. p. 1609–13.
3. Zhang Y, Rong L, Song D, Zhang P. A survey on multimodal sentiment analysis. Pattern Recognit Artif Intell. 2020;33(5):426–38. doi:10.1007/s10462-023-10555-8.
4. Shivappa ST, Trivedi MM, Rao BD. Audiovisual information fusion in human–computer interfaces and intelligent environments: a survey. Proc IEEE. 2010;98(10):1692–715. doi:10.1109/jproc.2010.2057231.
5. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency LP. Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2017 Jul 30–Aug 4; Vancouver, BC, Canada. Stroudsburg, PA, USA: ACL. p. 873–83.
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2017.
7. Ma H, Wang J, Lin H, Zhang B, Zhang Y, Xu B. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. IEEE Trans Multimedia. 2024;26(11):776–88. doi:10.1109/tmm.2023.3271019.

8. Poria S, Chaturvedi I, Cambria E, Hussain A. Convolutional MKL-based multimodal emotion recognition and sentiment analysis. In: Proceedings of the 16th IEEE International Conference on Data Mining; 2016 Dec 12–15; Barcelona, Spain. p. 439–48.
9. Nojavanasghari B, Gopinath D, Koushik J, Baltrušaitis T, Morency LP. Deep multimodal fusion for persuasiveness prediction. In: Proceedings of the ACM International Conference on Multimodal Interaction; 2016 Nov 12–16; Tokyo, Japan. p. 284–88.
10. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2017 Sep 9–11; Copenhagen, Denmark. p. 1103–14.
11. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh AB, Morency LP. Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers); 2018 Jul 15–20; Melbourne, Australia. Stroudsburg, PA, USA: ACL; 2018. p. 2247–56.
12. Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 6558–69.
13. Ma L, Yao Y, Liang T, Liu T. Multi-scale cooperative multimodal transformers for multimodal sentiment analysis in videos. In: Proceedings of the Australasian Joint Conference on Artificial Intelligence; 2024 Nov 25–29; Melbourne, Australia. p. 281–97.
14. Sun L, Lian Z, Liu B, Tao J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Trans Affect Comput. 2023;15(1):309. doi:10.1109/taffc.2023.3274829.
15. Hu G, Lin TE, Zhao Y, Lu G, Wu Y, Li Y. UniMSE: towards unified multimodal sentiment analysis and emotion recognition. arXiv:2211.11256. 2022. doi:10.48550/arXiv.2211.11256.
16. Shi T, Huang S-L. MultiEMO: an attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2023 Jul 9–14; Toronto, ON, Canada. Stroudsburg, PA, USA: ACL; 2023. p. 14752–66.
17. Arevalo J, Solorio T, Montes-y-Gómez M, González F. Gated multimodal units for information fusion. arXiv:1702.01992. 2017. doi:10.48550/arXiv.1702.01992.
18. Rahman W, Hasan MK, Lee S, Bagher Zadeh A, Mao C, Morency LP, et al. Integrating multimodal information in large pretrained transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 2359–69.
19. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019. doi:10.48550/arXiv.1907.11692.
20. Eyben F, Weninger F, Gross F, Schuller B. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia; 2013 Oct 21–25; Barcelona, Spain. p. 835–8.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
22. Wang W, Tran D, Feiszli M. What makes training multi-modal classification networks hard? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 12692–702.
23. Nagrani A, Yang S, Arnab A, Jansen A, Schmid C, Sun C. Attention bottlenecks for multimodal fusion. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2021. p. 14200–13.
24. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval. 2008;42(4):335–59.
25. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: a multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 527–36.
26. Qi Y, Ibrayim M, Tohti T. Contextual xLSTM-based multimodal fusion for conversational emotion recognition. Pattern Anal Applic. 2025;28(3):132. doi:10.1007/s10044-025-01508-8.

27. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E. DialoguERN: an attentive RNN for emotion detection in conversations. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2019 Jan 29–31; Honolulu, HI, USA. p. 6818–25.
28. Yi Z, Zhao Z, Shen Z, Zhang T. Multimodal fusion via hypergraph autoencoder and contrastive learning for emotion recognition in conversation. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, Australia. p. 4341–8.
29. Tu G, Xie T, Liang B, Wang H, Xu R. Adaptive graph learning for multimodal conversational emotion detection. Proc AAAI Conf Artif Intell. 2024;38(17):19089–97. doi:10.1609/aaai.v38i17.29876.
30. Zou S, Huang X, Shen X, Liu H. Improving multimodal fusion with main modal transformer for emotion recognition in conversation. Knowl Based Syst. 2022;258(4):109978. doi:10.1016/j.knosys.2022.109978.
31. Shou Y, Liu H, Cao X, Meng D, Dong B. A low-rank matching attention based cross-modal feature fusion method for conversational emotion recognition. IEEE Trans Affect Comput. 2024;16(2):1177–89. doi:10.1109/taffc.2024.3498443.
32. Yang J, Dong X, Du X. SMFNM: semi-supervised multimodal fusion network with main-modal for real-time emotion recognition in conversations. J King Saud Univ Comput Inf Sci. 2023;35(9):101791. doi:10.1016/j.jksuci.2023.101791.
33. Li J, Wang X, Liu Y, Zeng Z. CFN-ESA: a cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. IEEE Trans Affect Comput. 2024;15(4):1919–33.
34. Mao Y, Liu G, Wang X, Gao W, Li X. DialogueTRM: exploring multi-modal emotional dynamics in a conversation. In: Findings of the Association for Computational Linguistics: EMNLP. Stroudsburg, PA, USA: ACL; 2021. p. 2694–704.
35. Hu J, Liu Y, Zhao J, Jin Q. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Long Papers); 2021 Aug 1–6; Virtual. Stroudsburg, PA, USA: ACL; 2021. p. 5666–75.