BCAF: Bimodal connection attention fusion for robust emotion recognition from speech and text

Jiachen Luo, Huy Phan, Lin Wang, Joshua D. Reiss*

***Abstract*—Bimodal speech emotion recognition is challenging due to the difficulty of capturing subtle emotional cues and effectively modeling complex interactions between audio and text. Existing methods often overlook the importance of modality connections and struggle to filter noisy or misleading cross-modal signals. To address these issues, we propose a novel Bimodal Connection Attention Fusion (BCAF) framework that explicitly models modality-specific representations and their interactions. BCAF includes three key modules: an interactive connection network that learns modality connections through an encoder-decoder structure; a bimodal attention network that captures intra- and inter-modal dependencies; and a correlative attention network that filters cross-modal noise to improve robustness. Extensive evaluations across benchmark datasets show that BCAF achieves superior performance over existing methods, highlighting its strength in modeling detailed modality connections and interactions for emotion recognition.**

***Index Terms*—deep learning, conversational emotion recognition, multi-modal fusion, modality connection, modality interaction, attention**

I. Introduction

THE proliferation of mobile internet and smartphones has led to the widespread use of social networking platforms where users create and share content in various modalities, such as audio, text, and video. Extracting and analyzing emotions from this multimodal content has extensive applications in human-computer interaction, surveillance, robotics, and gaming [1], [2], [3]. However, effectively integrating multiple modalities remains a significant challenge in emotion recognition research.

Previous multi-modal emotion recognition methods have achieved good performance [4], [5]; however, key challenges remain in multi-modal emotion recognition. Different modalities require independent preprocessing and feature extraction designs due to their heterogeneous nature [6], [7], [8]. To develop a model that is both applicable and generalizable across individual modalities and fusion models, it is essential to learn modality connections and interactions for learning discriminative emotional content.

In multi-modal learning, modality connection in multi-modal learning refers to the extent to which information is shared across modalities, shaping a unified representation [9]. Unlike correlation, which quantifies modality dependence, modality connection captures the semantic alignment between modalities. For example, a higher vocal pitch may correlate with excitement, but modality connection determines if the audio tone, facial expressions, and text consistently reinforce that excitement. A strong modality connection ensures meaningful interaction among modalities, whereas a weak connection may lead to ambiguity. Leveraging modality connections enables a more nuanced understanding of emotions, improving the robustness and accuracy of emotion recognition systems.

On the other hand, modality interaction refers to the relationships and dependencies between modalities, which can vary from strong correlations to weak or even conflicting signals [10]. While modality connection emphasizes shared content, modality interaction captures how different modalities influence one another. For example, imagine a meeting scenario where a participant says, "That's great" in a flat, monotone voice while avoiding eye contact and crossing their arms. The verbal content suggests positivity, but the tone and body language convey disengagement or sarcasm. Such conflicting signals highlight the complexity of multi-modal interactions. Accurately interpreting these interactions requires understanding the context and modeling how modalities influence one another.

Two core types of interactions are frequently encountered in multi-modal learning: intra-modal and inter-modal interactions [11], [12], [13]. Intra-modal interactions measure relationships within the same modality, while inter-modal interactions capture relationships between different modalities. Understanding and learning these interactions are essential for building an effective multi-modal emotion recognition system that achieves three key goals: multi-modal integration, robustness, and contextual awareness.

Many multi-modal emotion recognition studies have focused on exploring interactions across different modalities while often overlooking the importance of modality connections. Existing approaches typically concatenate data for joint fusion or aggregate decisions from various modalities through weighted averaging [6], [14], [15], which neglects intra-modal interactions and results in limited modeling of the nuanced dependencies between modalities.

Recently, attention mechanisms have gained attraction due to their effectiveness in modeling modality interactions [16], [17], [18]. One representative work is the Multimodal Transformer (MulT) model [19], which applies cross-modal transformers to process unaligned multimodal sequences. While effective, such approaches primarily focus on global attention-based fusion and often lack a deeper structural understanding of how modalities interact and connect, leading to potential shortcomings such as suboptimal connection strategies, loss

The work does not relate to Huy Phan's position at Meta, Paris, France, e-mail: huyphan@meta.com.
Jiachen Luo, Lin Wang and Joshua Reiss are with the Centre for Digital Music, Queen Mary University of London, London, United Kingdom, UK, e-mail: {jiachen.luo, lin.wang, joshua.reiss}@qmul.ac.uk.
* corresponding author

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3699027

of crucial cross-modal information, and diminished robustness when dealing with conflicting or incomplete modality data.

Compared with MulT and similar frameworks [1], [18], [19], our method introduces a more fine-grained and structured approach to modeling modality interactions. Rather than applying general cross-modal attention across sequences, our method emphasizes the explicit modeling of modality connections via specialized connection-aware modules that capture both intra-modal and inter-modal dynamics. This distinction allows for better interpretability and adaptability, especially in the context of bimodal emotion recognition where the semantic alignment between modalities (e.g., audio and text) is often complex and variable.

Motivated by these observations (see Fig. 1), we propose a novel Bimodal Connection Attention Fusion (BCAF) framework for bimodal emotion recognition, which consists of three key components: the uni-modal representation module, the connection attention fusion module, and the classification module. The main contributions of this paper are as follows:

- We propose an interactive connection network that explicitly models modality-specific features and their connections, enabling more precise analysis of cross-modal relationships between audio and text.
- We introduce a bimodal attention network, which dynamically adjusts weights to comprehensively model both intra-modal and inter-modal interactions, leading to richer and more discriminative fused representations.
- We design a correlative attention network to filter out misleading cross-modal signals and enhance the learning of valid intra- and inter-modal dependencies, improving model robustness and generalization.

Experimental results on several public datasets demonstrate that the proposed BCAF framework significantly outperforms existing methods, validating its effectiveness and superiority in modeling fine-grained modality interactions for emotion recognition tasks.

The remainder of this paper is organized as follows: Section II presents a brief literature review. Section III describes our method in detail. Section IV outlines the experiments conducted. Section V discusses the results. Finally, Section VI provides conclusions based on this work.

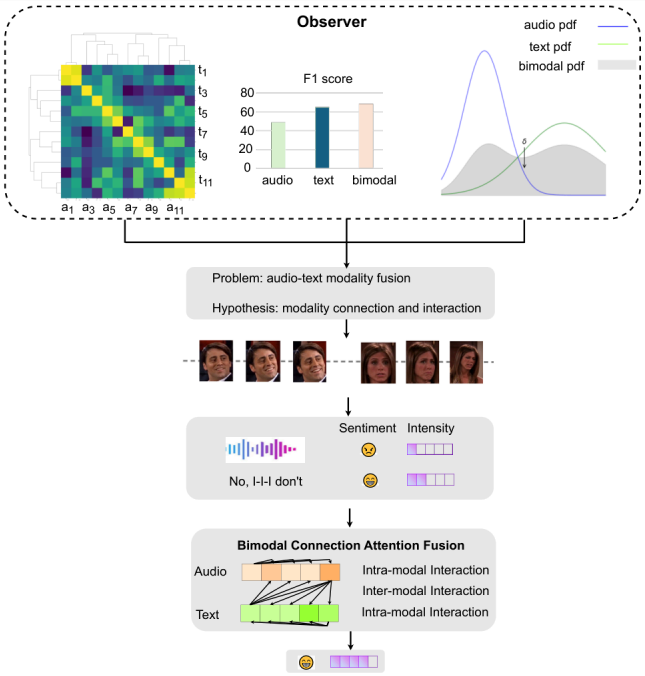


Fig. 1. Illustration of the proposed Correlation-aware Bimodal Attention Fusion (BCAF) framework, inspired by experimental observations on audio-text modality correlation. The top part presents correlation heatmaps across time steps, comparative F1 scores for unimodal and bimodal systems, and modality distribution curves, which motivate our hypothesis on cross-modal interactions. Based on these findings, we define the problem of audio-text modality fusion and propose a hypothesis emphasizing the importance of modality connection and interaction. The middle part of the figure depicts a real dialogue example with aligned audio, and text data, highlighting sentiment intensity analysis. The bottom part shows our BCAF mechanism, which systematically incorporates intra-modal and inter-modal attention fusion to effectively model both within-modality and cross-modality interactions.

II. Related Work

A. Multi-modal Emotion Recognition

Emotion recognition in conversations has found widespread applications across various fields [20]. Humans convey emotions through multiple modalities, including speech and facial expressions [14], [21], [22]. Among these, speech is a critical modality because it expresses emotional cues through both paralinguistic features, such as pitch, tone, and prosody, and linguistic content. Since each modality contributes different types of information, a single modality alone cannot provide sufficient accuracy for emotion recognition [6], [23], [24]. Consequently, combining audio and text data has become a dominant strategy in bimodal speech emotion recognition system.

Multi-modal fusion methods are broadly categorized into early fusion, late fusion, and hybrid fusion [1], [2], [20], [24], [25]. Early fusion combines features from different modalities at the input level but may overlook complex inter-modal dependencies [26], [27]. Late fusion aggregates decision-level outputs from uni-modal classifiers [28], [29], simplifying processing but limiting joint interactions. Hybrid fusion combines intermediate representations from multiple modalities, and this approach achieves a balance between architectural flexibility and representational expressiveness [30], [31].

Recent transformer-based methods have shown promise in modeling multi-modal dependencies. For example, M2FNet [16], HCAM [32], and CFN-ESA [33] use multi-head attention or co-attention mechanisms to learn both intra- and inter-modal interactions. However, such methods often lack hierarchical filtering of noisy or misleading modality signals. MulT [19] employs cross-modal transformers to align unaligned sequences, but fails to capture intra-modal structure or selectively suppress spurious modality features. Similarly, TelME [34], Mamba [35], and AGF-IB [36] introduce fusion constraints and contrastive learning, but still struggle with long-range dependency modeling and fine-grained control.

Other works aim to address modality-specific representation learning. Shan et al. [37] proposed cross-modal transformers that disentangle modality-invariant and modality-specific

components, while Han et al. [38] introduced correlation-controlled fusion between modalities. Zou et al. [39] introduced multi-level acoustic co-attention to preserve hierarchical audio information before fusion. Kakuba et al. [40] provided a comprehensive review of deep learning architectures for bimodal emotion recognition and proposed a multi-learning model to balance between local and global supervision. Ramkrishnan et al. [41] empirically demonstrated the role of acoustic feature importance in determining deep model behavior across emotion categories. In the graph-based domain, Su et al. [42] developed a causal disentanglement framework, while Nie et al. [43], [44] integrated commonsense reasoning and graph convolutional structures to handle long or incremental conversational emotions.

B. Noise Filtering Strategies

Emotional expressions often appear as brief and subtle bursts within long conversational sequences, making it essential to filter out irrelevant or redundant information to prevent dilution of emotional cues. Accordingly, noise filtering has become a crucial component in multi-modal emotion recognition. Existing approaches can be broadly categorized into three strategies: reconstruction-based methods [45], [46], [47], gating-based methods [48], [49], [50], and attention-based methods [51], [52], [53].

1) Reconstruction-based Filters: Reconstruction methods, including auto-encoders [45], variational auto-encoders [46], and encoder–decoder frameworks [47], denoise modality-specific signals by reconstructing them from compressed latent vectors. For example, SMIN [54] integrates intra- and cross-modal modules within an encoder–decoder backbone. Such approaches effectively preserve essential features and provide a self-supervised training paradigm via reconstruction losses. However, they often overemphasize unimodal refinement, insufficiently model cross-modal dependencies, and lack mechanisms for balancing modality contributions, sometimes producing clean but poorly fused representations.

2) Gating-based Filters: Gating mechanisms regulate modality contributions by adaptively weighting features. Early designs used simple context or sigmoid gates [48], later extended by modality-specific modules [49] and mixture-of-experts gating [50]. These methods enhance robustness and interpretability by suppressing irrelevant signals. Yet, scalar gates oversimplify complex dependencies, often bias decisions toward dominant modalities, and fail to explicitly evaluate cross-modal correlations. Moreover, they add computational overhead and may destabilize training without careful regularization.

3) Attention-based Filters: Attention mechanisms provide flexible modeling of intra- and inter-modal dependencies through dot-product, self-, cross-, and multi-head variants [51], [52], [53]. A notable example is MulT [19], which employs cross-modal transformers for sequence alignment. Attention-based filters highlight salient features, capture long-range relationships, and offer interpretability via attention weights. Nevertheless, they may amplify spurious correlations, underemphasize intra-modal structures, and lack explicit balancing between intra- and inter-modal signals.

While each paradigm contributes to noise suppression, none alone is sufficient. Reconstruction excels at denoising but underutilizes cross-modal cues; gating adaptively weighs modalities but oversimplifies dependencies; attention flexibly captures interactions but risks noisy alignments. Current models also tend to underexploit the richness of audio, lack hierarchical modeling of interactions, and remain vulnerable to noisy or misleading modality connections, e.g., from speech recognition errors. To address these limitations, we propose the *Bimodal Connection Attention Fusion (BCAF)* framework, which combines three complementary modules: the interactive connection network for reconstruction-driven refinement of both uni-modal and cross-modal features; the correlative attention network for balanced self-, cross-, and joint attention modeling; and the bimodal correlation evaluation module, which uses correlation-driven gating to suppress noise and balance modality contributions. Together, these components enable robust, interpretable, and semantically aligned fusion, significantly improving the reliability and performance of bimodal speech emotion recognition.

III. Methodology

Our proposed BCAF method simultaneously models modality connections and interactions for bimodal speech recognition system. Fig. 1 illustrates the architecture of BCAF, which consists of the uni-modal representation module, the connection attention fusion module, and the classification module. The core connection attention fusion module comprises of the the interactive connection network, the bimodal attention network and the correlative attention network (see Figs. 3-5).

A. Uni-modal Representation Module

1) Acoustic Encoder: We use the large wav2vec model as the audio modality encoder to obtain a 1024-dimensional utterance-level audio representation from raw audio signals [55]. The large wav2vec model is an advanced self-supervised learning framework designed for speech representation learning. It consists of three key components: a feature encoder, a Transformer-based contextual representation module, and a quantization module. In total, 1024-dimensional utterance-level acoustic features were extracted (H_a).

2) Textual Encoder: We use the RoBERTa model as the text modality encoder to extract a 1024-dimensional utterance-level text representation from raw text [56]. RoBERTa takes the utterance transcript as input and generates rich contextual representations from the final four layers. This process produces four 1024-dimensional vectors for each token in the input. We then average these vectors to obtain a contextual utterance feature vector with a dimension of 1024 (H_l).

B. Connection Attention Fusion Module

We propose the connection attention fusion module to learn modality connections and interactions between audio and text for bimodal speech emotion recognition. The module comprises three main components: the interactive connection network (ICN), the bimodal attention network (BAN), and the

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3699027

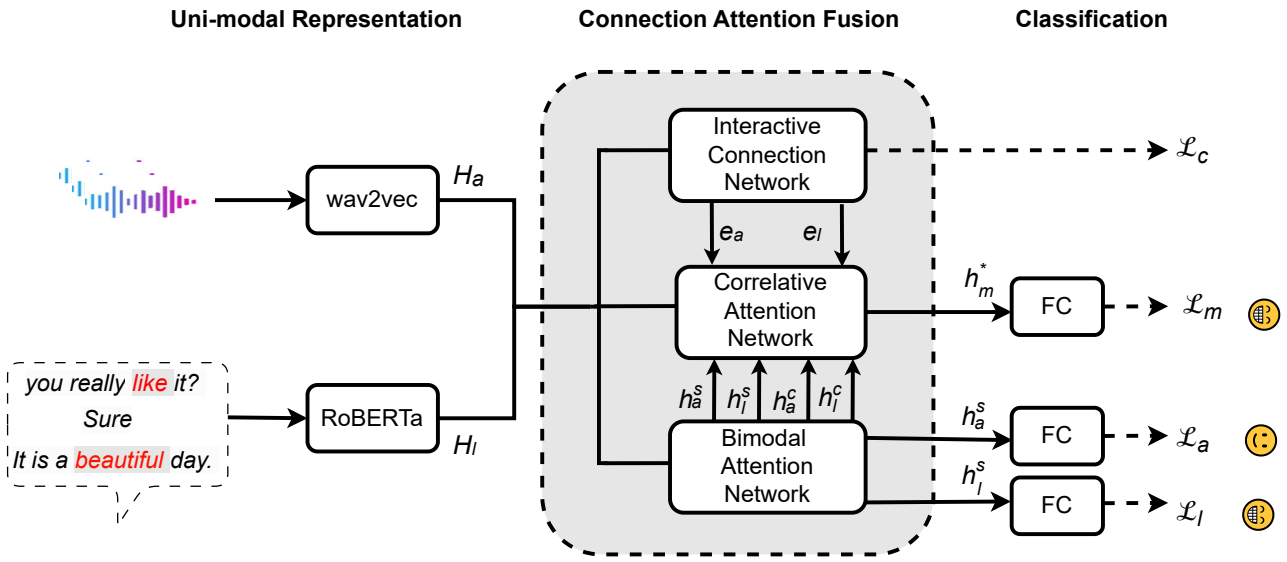


Fig. 2. Architecture of our proposed the Bimodal Connection Attention Fusion method. The method consists of three modules, the uni-modal representation, the connection attention fusion module and the classification module. The uni-modal audio representation H_a and text representation H_l are inputted into all the interactive connection network, correlative attention network and bimodal attention network. The core connection attention fusion module includes the interactive connection network (ICN), the bimodal attention network (BAN) and correlative attention network (CAN), with the details depicted in Figs. 3-5, respectively.

correlative attention network (CAN) (see Figs. 3-5). Detailed explanations of these components are provided in the following subsections.

1) Interactive Connection Network: The interactive connection network uses an encoder-decoder architecture to learn modality connections between audio and text (see Fig. 3). Both the encoder and decoder consist of three fully connected layers, with each layer applying a linear transformation followed by a non-linear ReLU activation function. In the encoder, the fully connected layers progressively reduce the dimensionality of the input uni-modal representation H_m, extracting important features and compressing them into a fixed-size latent vector e_m. This process of hierarchical dimensionality reduction enables the encoder to capture the essential characteristics of the input modality while filtering out irrelevant or redundant information.

The decoder, which also consists of three fully connected layers, reverses this process by gradually increasing the dimensionality of the latent vector e_m. It reconstructs the original uni-modal representation d_m using the latent features as a guide. Each layer in the decoder applies a linear transformation and ReLU activation function to ensure that the reconstructed output closely resembles the original input. This encoder-decoder architecture effectively enables feature compression and reconstruction, facilitating the robust learning of modality connections.

This symmetric encoder-decoder architecture facilitates efficient feature extraction, compression, and reconstruction. By designing the encoder to focus on compact and meaningful representations and the decoder to restore the input features

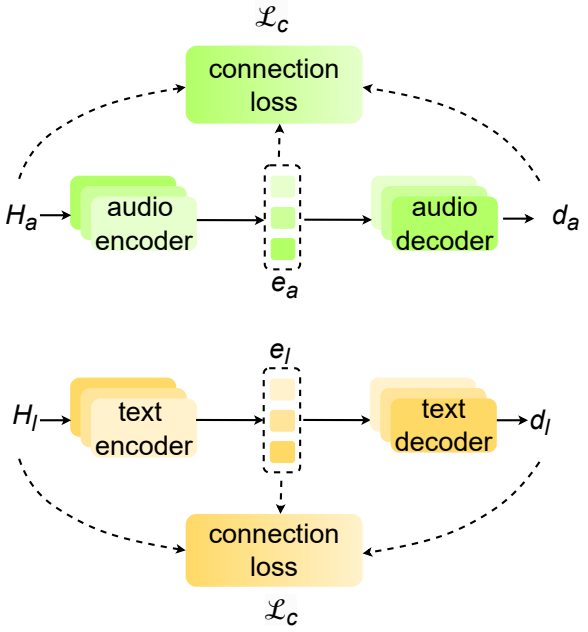


Fig. 3. Update scheme of the interactive connection network (ICN).

accurately, the network ensures that modality-specific information is preserved while enabling cross-modality learning.

The interactive connection network adopts an architecture comprising stacked fully connected layers followed by a dropout layer in both the encoder and decoder (see Fig. 3). For each modality $m \in \{a, l\}$, a simple encoder-decoder is formulated as:

$$e_m = E_m(H_m, \theta_m^e) \tag{1}$$

$$d_m = D_m(e_m, \theta_m^d) \tag{2}$$

where $E_m(\cdot)$ denotes the encoder function for modality m, with θ_m^e as its trainable parameters, and $D_m(\cdot)$ represents the decoder function for modality m, with θ_m^d as its trainable parameters. The dimensions of H_m and d_m are both 1024. The dimension of e_m is 512.

The objective function models the learning problem using the self-supervised spirit. We design the connection loss function to maximize modality connections between audio and text. It is defined as follows:

$$\mathcal{L}_c = \left(\|H_a - d_a\|_F^2 + \|H_l - d_l\|_F^2 + \mu \left(\|I_e - e_a^\top e_l\|_F^2 - \|I_d - d_a^\top d_l\|_F^2 \right) \right) \tag{3}$$

where $\| \cdot \|_F^2$ is the squared Frobenius norm, which calculates the sum of squared elements in a matrix. I_e and I_l represent the identity matrix, and μ is a non-negative hyperparameter that controls the balance between terms.

The first term of the objective function, $\|H_a - d_a\|_F^2 + \|H_l - d_l\|_F^2$, measures the reconstruction error between the original input representations and their reconstructed outputs for both the audio (H_a, d_a) and text (H_l, d_l) modalities. By minimizing this term, the model ensures that the decoder can accurately reconstruct the original input representations, preserving the modality-specific information unique to each modality. This encourages the encoder-decoder mechanism to retain the essential features of each modality while discarding irrelevant or redundant information. As a result, the model achieves high-quality reconstruction for both modalities, which is critical for ensuring robust performance in downstream tasks.

The second term, $\|I_e - e_a^\top e_l\|_F^2 - \|I_d - d_a^\top d_l\|_F^2$, focuses on learning meaningful connectionsbetween audio and text. The first component, $\|I_e - e_a^\top e_l\|_F^2$, aligns the latent representations (e_a, e_l) in the feature space, ensuring strong modality connections during the encoding process. The second component, $\|I_d - d_a^\top d_l\|_F^2$, maintains these connections in the reconstructed space by aligning the outputs (d_a, d_l). Together, these terms promote consistency between the latent and reconstructed spaces, enabling the model to capture robust cross-modal relationships. This ensures that the learned modality connections are both meaningful and preserved across all stages of the network.

2) Bimodal Attention Network: The bimodal attention network introduces self-attention and cross-attention mechanisms to learn intra- and inter-modal interactions between audio and text. Fig. 5 illustrates the architecture of the bimodal attention network. The input to this network consists of queries, keys, and values. The dot product of the query and each key is computed, and a softmax function is applied to generate weights for the values [52]. The bimodal attention network comprises stacked self-attention and cross-attention layers, along with feed-forward layers.

The core idea of this module is to learn intra- and inter-modal interactions between audio and text, then propagate information from both modality-specific patterns and modality associations based on the attention weights. Technically, the self-attention layer aims to learn intra-modal interactions within each modality, such as audio or text. The query, key, and value are derived from the same modality. Given weight matrices W_m^Q, W_m^K, and W_m^V, the modality representation H_m is projected into the query matrix (Q_m), key matrix (K_m), and value matrix (V_m) through linear projections without bias. The self-attention representation can then be summarized as follows:

$$\zeta H_m = \text{softmax}\left(\frac{Q_m K_m^T}{\sqrt{d_k}} \right) V_m \tag{4}$$

where ζH_m represents the self-attention representation with a dimensionality of 1024. d_k represents the dimension of the key vector.

To further enhance the representation capacity, the self-attention representation is passed through a LayerNorm layer followed by an AddNorm layer to obtain the enhanced self-attention representation.

$$H_m^s = \text{LayerNorm}(H_m \oplus \zeta H_m) \tag{5}$$

$$h_m^s = \text{AddNorm}(H_m^s \oplus \text{FeedForward}(H_m^s)) \tag{6}$$

where h_a^s and h_l^s represent the enhanced self-attention representations for audio and text, respectively. The dimensions of h_a^s and h_l^s are both 1024.

Parallel to the self-attention layer, the cross-attention layer captures inter-modal interactions between audio and text. It learns associations between the two modalities and propagates information from one modality to the other based on these associations. Specifically, the cross-attention mechanism follows a similar principle to the self-attention mechanism, with the key difference being that the query, key, and value are derived from different modalities. The enhanced cross-attention representation is summarized as follows:

$$\zeta H_{l-a} = \text{softmax}\left(\frac{Q_l K_a^\mathsf{T}}{\sqrt{d_k}} \right) V_a \tag{7}$$

$$H_a^c = \text{LayerNorm}(H_a \oplus \zeta H_{l-a}) \tag{8}$$

$$h_a^c = \text{AddNorm}(H_a^c \oplus \text{FeedForward}(H_a^c)) \tag{9}$$

$$\zeta H_{a-l} = \text{softmax}\left(\frac{Q_a K_l^\mathsf{T}}{\sqrt{d_k}} \right) V_l \tag{10}$$

$$H_l^c = \text{LayerNorm}(H_l \oplus \zeta H_{a-l}) \tag{11}$$

$$h_l^c = \text{AddNorm}(H_l^c \oplus \text{FeedForward}(H_l^c)) \tag{12}$$

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3699027

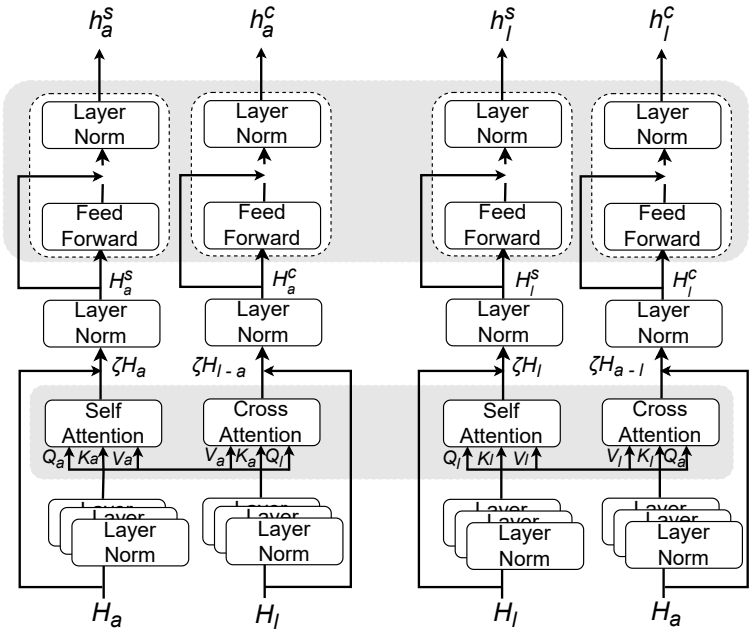


Fig. 4. Update scheme of the bimodal attention network (BAN).

where ζH_{a-l} and ζH_{l-a} represent the propagated information from audio to text and from text to audio, respectively, both with dimensions of 1024. The variables h_a^c and h_l^c denote the enhanced cross-attention representations for audio and text, respectively, and their dimensions are also both 1024. d_k represents the dimension of the key vector.

3) Correlative Attention Network: The correlative attention network is designed to enhance sentiment analysis by explicitly modeling the relationships between uni-modal and bimodal representations. It takes as input the latent uni-modal representations from the interactive correlation network and correlates them with the bimodal representation to capture both intra-modal and inter-modal dependencies. This network is crucial for bridging the gap between individual modality features and their combined representation, ensuring that the interactions between modalities are accurately captured.

The correlative attention network comprises two submodules: the joint attention network and the bimodal correlation evaluation network (see Fig. 5). The joint attention network integrates the uni-modal latent features from the audio and text modalities to produce a comprehensive bimodal representation, allowing the model to focus on the most salient features from each modality. Meanwhile, the bimodal correlation evaluation network assesses the correlations between these modalities by quantifying their interdependencies, offering a deeper understanding of how features from one modality influence those of the other.

By incorporating the correlative attention network, the model achieves a more robust and nuanced representation of the input data, improving its ability to analyze sentiment effectively. This approach ensures that the unique contributions of each modality are not only preserved but also synergistically leveraged to enhance overall performance.

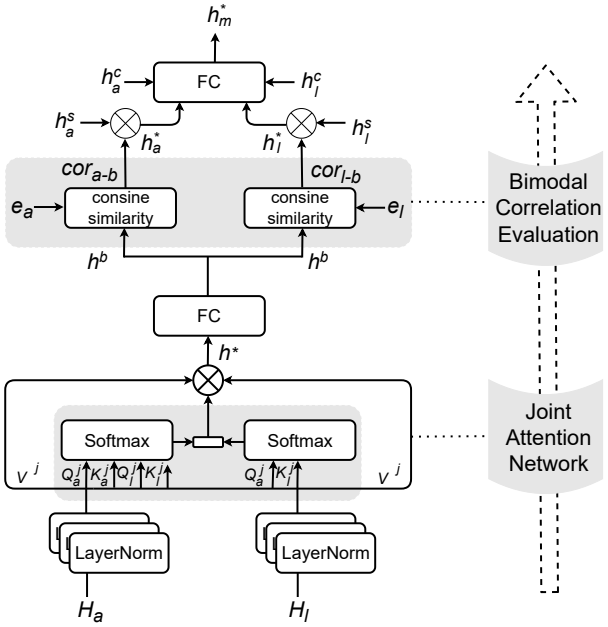


Fig. 5. Update scheme of the correlative attention network (CAN) consisting of the joint attention network and the bimodal correlation evaluation network.

Joint Attention Network: The joint attention network aims to enhance the cross-modal relationship between audio and text by reducing noise and highlighting meaningful interactions. This is achieved using a pair of softmax functions, which help focus on the most relevant features from both modalities while suppressing less informative or noisy elements. The

network effectively refines the cross-modal representation, enabling the model to better capture salient patterns for downstream tasks such as sentiment analysis.

The attention mechanism operates by mapping query, key, and value vectors to outputs. Specifically, it computes attention scores using query and key vectors and applies these scores to calculate a weighted sum of the value vectors. Given the uni-modal representations H_m, the joint attention network first projects these representations into query, key, and value vectors $(Q_a^j, Q_l^j, K_a^j, K_l^j, V_a^j, V_l^j)$ as follows:

$$Q_m^j = h_m^s W_m^{Q^j}, \quad K_m^j = h_m^s W_m^{K^j}, \quad V_m^j = h_m^s W_m^{V^j} \tag{13}$$

The joint attention network computes the joint bimodal representation using the following equation:

$$h^* = \left(\text{softmax}\left(\frac{Q_a^j K_a^{j\top} + Q_l^j K_l^{j\top}}{\sqrt{d_k}} \right) - \lambda \, \text{softmax}\left(\frac{Q_a^j K_l^{j\top}}{\sqrt{d_k}} \right) \right) V^j \tag{14}$$

where h^* denotes the joint bimodal representation with a dimension of 1024. λ is a learnable scalar that adjusts the weight of the second term, which represents cross-modal attention. V^j represents the average of the value vectors V_a^j and V_l^j, encapsulating information from both modalities. d_k represents the dimension of the key vector.

The use of a pair of softmax functions serves a dual purpose. The first softmax term emphasizes intra-modal relationships by combining attention scores from the query and key vectors within each modality. This term ensures that strong intra-modal signals are preserved. The second softmax term focuses on cross-modal relationships, computing the attention scores between the query of one modality and the key of the other. By subtracting the cross-modal scores (weighted by λ), the network suppresses noisy or irrelevant interactions while retaining meaningful cross-modal dependencies. This design improves the robustness of the joint attention mechanism and enhances the overall representation quality for tasks requiring fine-grained cross-modal understanding.

Bimodal Correlation Evaluation: The latent uni-modal representations from the interactive correlation network (e_a for audio and e_l for text), along with the joint bimodal representation from the joint attention network (h^*), are input into the bimodal correlation evaluation network to assess the correlations between audio and text. Inspired by CLIP [57], cosine similarity is employed to measure the correlations between different modalities, resulting in the pairwise cross-modal correlation coefficients, cor_{a-m} and cor_{l-m} (see Fig. 4). The joint multi-modal representation h^* is passed through a linear layer, reducing its dimension to 512, resulting in h^b, which serves as the fused representation used for bimodal correlation evalution.

$$\text{cor}_{a-b} = \left\langle e_a, h^b \right\rangle, \quad \text{cor}_{l-b} = \left\langle e_l, h^b \right\rangle \tag{15}$$

where $\langle \cdot \rangle$ denotes cosine similarity, and cor_{a-b} and cor_{l-b} represent the correlation coefficients for the audio-bimodal and text-bimodal pairs, respectively.

The pairwise cross-modal correlation coefficients are subsequently integrated into the enhanced self-attention representations (h_a^s for audio and h_l^s for text) to produce correlative representations. This process ensures that the resulting representations fully capture inter-modal interactions by utilizing the correlation information to strengthen the interplay between modalities. Through this integration, the model becomes better equipped to align and fuse audio and text features effectively, thereby facilitating a more comprehensive understanding of cross-modal relationships, as described below:

$$h_a^* = h_a^s \times \text{cor}_{a-b}, \quad h_l^* = h_l^s \times \text{cor}_{l-b} \tag{16}$$

where h_a^* and h_l^* denote the correlative uni-modal representations, enhanced by their respective modality correlation coefficients. These coefficients represent the weighted importance of one modality's features in the context of the other, effectively capturing the interdependence between modalities. The dimensions of both h_a^* and h_l^* are 512.

Finally, h_a^c, h_l^c, h_a^*, and h_l^* are concatenated to form the aggregated bimodal representation, which is then passed through four fully connected layers for classification.

$$h_m^* = h_a^c \oplus h_l^c \oplus h_a^* \oplus h_l^* \tag{17}$$

where h_m^* represents the aggregated bimodal representation with a dimension of 3072.

C. Classification

The overall optimization objective consists of the connection loss $\mathcal{L}_c$, the audio loss $\mathcal{L}_a$, the text loss $\mathcal{L}_l$, and the bimodal loss $\mathcal{L}_m$. Specifically, the audio loss, the text loss, and the bimodal loss are processed independently to generate predictions, which are then combined to make a final decision.

D. Training

For the classification task, the audio loss, the text loss, and the bimodal loss are defined as the standard cross-entropy loss. Finally, the overall loss function is expressed as a linear combination of $\mathcal{L}_c$, $\mathcal{L}_a$, $\mathcal{L}_l$, and $\mathcal{L}_m$:

$$\mathcal{L} = \alpha \mathcal{L}_c + \beta(\mathcal{L}_a + \mathcal{L}_l) + \mathcal{L}_m \tag{18}$$

where α, and β are weighting factors that balance the contribution of each loss term.

- **The audio loss** $\mathcal{L}_a$ is used to capture distinctive audio information for emotion prediction:

$$\hat{y}_a^c = \text{softmax}(W_e^a h_a^s + b_e^a) \tag{19}$$

$$\mathcal{L}_a = -\sum_{c=1}^{C} y_a^c \log(\hat{y}_a^c) \tag{20}$$

  where the parameters W_e^a and b_e^a are learnable weights and biases. c represents the emotion categories, and $\hat{y}_a^c$ and y_a^c denote the predicted and true labels, respectively.
- **The text loss** $\mathcal{L}_l$ is used to capture distinctive textual information for emotion prediction:

$$\hat{y}_l^c = \text{softmax}(W_e^l h_l^s + b_e^l) \tag{21}$$

$$\mathcal{L}_l = -\sum_{c=1}^{C} y_l^c \log(\hat{y}_l^c) \tag{22}$$

where the parameters W_e^l and b_e^l are learnable weights and biases. c represents the emotion categories, and $\hat{y}_l^c$ and y_l^c denote the predicted and true labels, respectively.

- **The bimodal loss** $\mathcal{L}_m$ is used to capture interactions between modalities for emotion prediction:

$$\hat{y}_m^c = \text{softmax}(W_e^m h_m^* + b_e^m) \tag{23}$$

$$\mathcal{L}_m = -\sum_{c=1}^{C} y_m^c \log(\hat{y}_m^c) \tag{24}$$

where the parameters W_e^m and b_e^m are learnable weights and biases. c represents the emotion categories, and $\hat{y}_m^c$ and y_m^c denote the predicted and true labels, respectively.

E. Differences between Our Work and Existing Work

Our work distinguishes itself from existing approaches through the design of noise-filtering techniques aimed at improving bimodal emotion recognition.

1) Reconstruction-based Methods: Traditional reconstruction filters, such as SMIN [54], incorporate both intra-modal and cross-modal reconstruction. However, the cross-modal reconstruction in SMIN remains an implicit interaction mechanism that focuses on reconstructing the target modality rather than explicitly aligning the two modalities in the latent space. Consequently, it cannot ensure consistency between modalities or balance their respective contributions, which makes the model more vulnerable to dominant-modality bias, noise, and modality conflicts. In contrast, our method (ICN) introduces an additional connection loss, on top of the reconstruction loss, to explicitly align and balance the two modalities. This design leads to better integration and more stable performance.

2) Gating-Based Methods: Bimodal gating aims to dynamically adjust the contribution of each modality based on their interaction. Prior work has explored a range of gating mechanisms, including both scalar-based and context-based gates [48], [49], [50]. However, existing bimodal gating methods determine the gate values through implicitly learned dominance. This dominance is not derived from explicit cross-modal indicators; rather, it emerges from the network's internal preference formed during training. Consequently, such implicit dominance often leads the model to over-rely on the stronger modality and makes it vulnerable to noise or conflicting modalities, ultimately resulting in suboptimal or unstable fusion decisions. In contrast, our method (BAN) explicitly computes cross-modal correlation $(\text{cor}_{a-b}, \text{cor}_{l-b})$ and uses it to guide the gating process. This explicit correlation-aware gating effectively prevents dominant-modality bias, better captures cross-modal consistency, and improves robustness under noisy or inconsistent multimodal conditions.

3) Attention-Based Methods: Attention mechanisms, such as those used in MulT [19], typically emphasize cross-modal interactions but often overlook intra-modal structures, thereby lacking an explicit balance between intra- and inter-modal signals. In contrast, our method (CAN) introduces a dedicated module for explicit and structured modeling of both intra- and inter-modal dependencies. This design helps preserve modality-internal structure, improves the reliability of cross-modal fusion, avoids dominant-modality bias, and enhances robustness to noise.

Our ablation study in Sec. V-B verifies the improved performance achieved by each of our proposed modules (ICN, BAN, and CAN) over their corresponding counterparts. Moreover, unlike prior approaches that rely on a single noise-filtering paradigm, our BCAF framework not only enhances each individual paradigm but also integrates them into a unified architecture to achieve more robust and reliable emotion recognition. The ablation results further validate the effectiveness of combining these paradigms within our unified framework.

TABLE I
STATISTICS OF TWO BENCHMARK DATASETS: MELD AND IEMOCAP

Dataset	# Conversations		# Utterances	
	Train/Validation	**Test**	**Train/Validation**	**Test**
MELD [26]	1,153	280	11,098	2,610
IEMOCAP [58]	120	31	5,810	1,623

IV. EXPERIMENTS

A. Database and Metrics

We evaluated our proposed BCAF method on the two benchmark datasets: MELD [26] and IEMOCAP [58]. Both these two commonly used public datasets are multi-modal, containing audio, text and video modality for every utterance. Due to the natural imbalance across various emotions, we choose weighted average F1-score as the evaluation metric. Table I shows the distribution of train and validation and test samples for both two datasets.

- MELD is a multi-modal and multi-party dataset for conversational emotion recognition [26]. It consists of 13,708 utterances in 1,433 dialogues collecting from the Friends TV shows. Each utterance is labeled with one in seven emotions: anger, joy, sadness, neutral, disgust, fear and surprise.
- IEMOCAP contains videos of dyadic conversations of ten speakers, spanning 12.46 hours [58]. Each utterance is annotated using the following discrete categories: happy, sad, neutral, angry, excited, and frustrated.

Given the inherent imbalance across different emotion classes, we used the weighted F1-score as our primary evaluation metric. The weighted F1-score calculates the F1-score for each class and applies weights based on the proportion of samples in each class. This approach ensures that the evaluation metric reflects the contribution of each class to the overall performance, addressing the impact of class imbalance effectively.

B. Baseline and State-of-the-Art Methods

To comprehensively evaluate the performance of the proposed method, we compared our results with those of the bidirectional LSTM (bc-LSTM) baseline [26]. This baseline system leveraged an utterance-level LSTM to model context-aware representations from surrounding utterances. Additionally, we compared the proposed BCAF method to various existing state-of-the-art methods:

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3699027

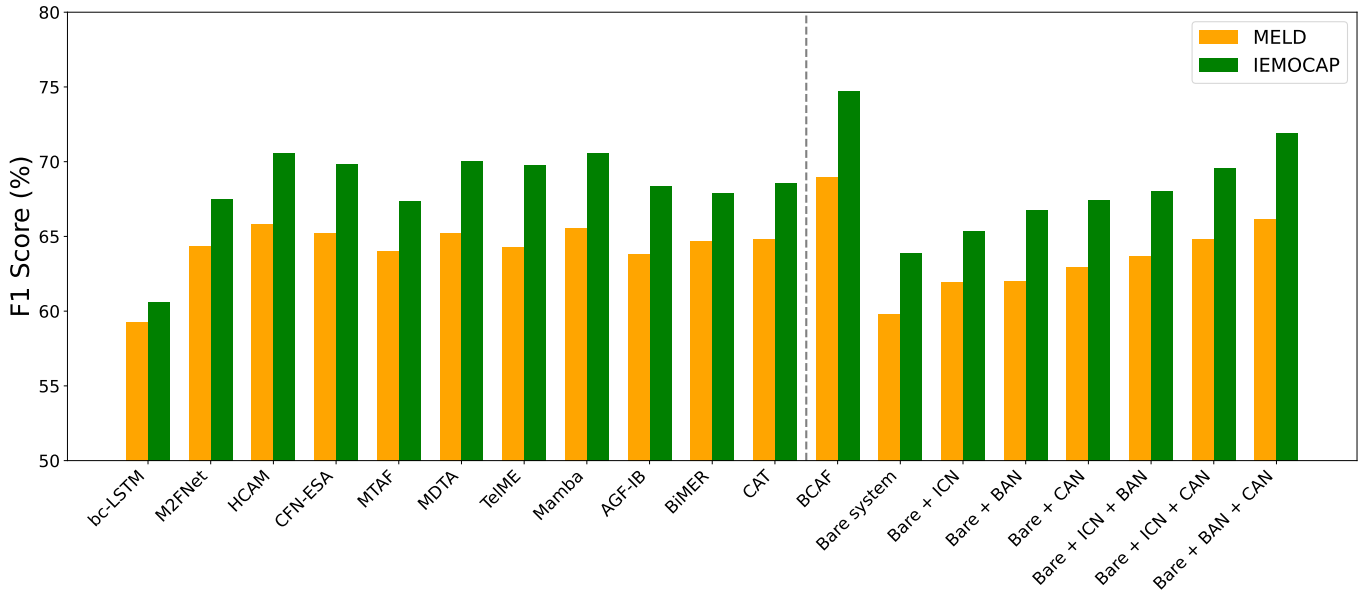


Fig. 6. Overall performance comparison between the BCAF method, the state-of-the-art methods, and the baselines. The bare system configuration excludes the ICN, BAN, and CAN modules.

- **M2FNet** [16] employed a multi-head attention-based fusion mechanism to combine emotion-rich latent representations of emotion-relevant features from visual, audio, and text modalities, learning both intra- and inter-modal relationships.
- **HCAM** [32] used a combination of recurrent and co-attention neural network to capture intra- and inter-modal interactions for emotion classification.
- **CFN-ESA** [33] incorporated a cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition.
- **TelME** [34] used cross-modal knowledge transfer, using a language model (as the teacher) to enhance non-verbal modalities (as the student), thereby optimizing the performance of weaker modalities.
- **MTAF** [59] proposed a multimodal transformer augmented fusion approach that integrated both feature-level and model-level fusion. It employed cross-modal transformer encoders to capture fine-grained interactions between audio and text modalities, significantly enhancing the accuracy of speech emotion recognition.
- **MDAT** [60] constructed a Multimodal Dual Attention Transformer that integrated graph attention and co-attention mechanisms to capture complex dependencies between audio and text modalities, enhancing cross-language speech emotion recognition performance.
- **Mamba** [35] designed a multi-modal fusion strategy based on probability guidance to maximize information consistency across modalities and capture intra- and inter-modal interactions for conversational emotion recognition.
- **AGF-IB** [36] introduced graph contrastive representation learning to capture intra- and inter-modal complementary semantic information, as well as intra-class and inter-class boundary information for emotion categories.
- **BiMER** [61] introduced a bimodal emotion recognition system that leveraged data augmentation techniques to improve the robustness and generalization of emotion recognition models using audio and text inputs.
- **CAT** [62] presented Coordination Attention-based Transformers with a bidirectional contrastive loss, aiming to align and fuse audio and text features effectively for improved multimodal speech emotion recognition.

C. Model Configuration

We implemented our proposed BCAF method using the PyTorch 1.11.0 framework. The model was trained with the Adam optimizer with an initial learning rate of 1e-4 and an early-stopping strategy with a patience of 15 epochs. To aid convergence and improve generalization, we applied L_2 regularization with a weight of 0.0001 and used dropout with a rate of $p = 0.3$ to mitigate overfitting.

V. Results and Discussion

We start by conducting comparative experiments against previous state-of-the-art baselines. Next, we ablate core module to verify the effectiveness of our proposed BCAF method. Following that, we emphasize the importance of bimodal attention network and qualitative analysis. Finally, we conduct case studies and error analysis.

A. Comparison with State-of-the-art Methods

We evaluate our proposed BCAF model on the MELD and IEMOCAP datasets, comparing it to a range of state-of-the-art methods. As shown in Table III, BCAF outperforms all baseline methods in terms of weighted F1-score, achieving 68.93% on MELD and 74.69% on IEMOCAP. Specifically,

TABLE II
PERFORMANCE COMPARISON ON MELD AND IEMOCAP DATASETS. **BOLD VALUES** INDICATE THE BEST PERFORMANCE ON EACH DATASET.

Method	MELD	IEMOCAP
bc-LSTM [26]	59.25	60.59
M2FNet [16]	64.32	67.47
HCAM [32]	**65.78**	70.56
CFN-ESA [33]	65.23	69.80
MTAF [59]	63.98	67.32
MDTA [60]	65.19	70.03
TelME [34]	64.29	69.76
Mamba [35]	65.56	70.58
AGF-IB [36]	63.82	68.36
BiMER [61]	64.65	67.89
CAT [62]	64.79	68.51
Our proposed BCAF	**68.93**	**74.69**

TABLE III
ABLATION STUDY ON THE MELD AND IEMOCAP DATASETS. **BOLD VALUES** INDICATE THE BEST PERFORMANCE ON EACH DATASET.

Method	MELD	IEMOCAP
Bare system	59.76	63.89
Bare + ICN (*Reconstruction*)	61.91	65.32
Bare + BAN (*Gating*)	62.00	66.73
Bare + CAN (*Attention*)	62.90	67.41
Bare + ICN + BAN	63.69	68.01
Bare + ICN + CAN	64.83	69.57
Bare + BAN + CAN	66.12	71.86
Our proposed BCAF	**68.93**	**74.69**
Bare system + SMIN (*Reconstruction*)	60.65	64.96
Bare system + scalar gating (*Gating*)	60.23	64.58
Bare system + MulT (*Attention*)	61.75	65.01

BCAF achieves a 3.15% improvement over HCAM [32] on MELD and a 4.11% improvement over Mamba [35] on IEMOCAP.

We compare BCAF against recent fusion-based models, including MTAF [59], MDTA [60], BiMER [61], and CAT [62], which employ advanced fusion techniques such as transformer augmentation, dual-attention, and contrastive objectives. While these models show promising performance, their weighted F1-scores are notably lower than BCAF's, with BiMER [61] and CAT [62] achieving 64.65% and 64.79% on MELD, respectively, which are 4.28% and 4.14% lower than BCAF.

BCAF also outperforms earlier contextual and transformer-based models, such as bc-LSTM [26], M2FNet [16], and TelME [34]. The performance improvements can be attributed to BCAF's ability to explicitly model modality-specific features, intra-modal structures, and cross-modal correlations in a coordinated and hierarchical manner, rather than relying on implicit fusion or late integration.

Interestingly, BCAF yields more significant improvements on IEMOCAP compared to MELD. This may be due to the difference in dialogue structure and noise levels between the datasets. MELD contains shorter dialogues and more background noise (e.g., honking or crowd chatter), which may hinder the model's ability to capture long-range emotional dependencies. In contrast, IEMOCAP features longer, cleaner recordings, making it easier for the model to capture subtle emotions. Additionally, rare emotion classes like *fear* and *disgust* are often underrepresented and more prone to misclassification in noisy environments, affecting MELD results.

Overall, the experimental results and ablation studies confirm the robustness and generalizability of BCAF, setting a new standard for emotion and intention recognition in conversational contexts.

B. Ablation Study

To validate the effectiveness of the proposed BCAF framework, we conducted a comprehensive ablation study by analyzing the performance gains brought by progressively adding its three core components: the Interactive Connection Network (ICN), the Bimodal Attention Network (BAN), and the Correlative Attention Network (CAN). This cumulative approach allows us to examine not only the individual contributions of each module but also the synergistic effects when combining them. As shown in Fig. 6 and Table III, we designed three categories of experiments: starting from a baseline model with all three modules removed, incrementally adding each module one at a time to evaluate its impact, and evaluating combinations of two or more modules.

The bare system, excluding the ICN, BAN, and CAN modules, achieves a weighted F1-score of 59.76% on the MELD dataset and 63.89% on IEMOCAP, establishing a performance lower bound. In contrast, the complete BCAF model, which integrates all three modules, attains weighted F1-scores of 68.93% on MELD and 74.69% on IEMOCAP—representing absolute improvements of 9.17% and 10.80%, respectively. These results highlight the overall effectiveness of our architectural design.

1) Effect of Adding the Interactive Connection Network: We first examined the effect of adding the ICN to the baseline model. This module was designed to leverage an encoder-decoder architecture to learn modality-specific features and capture modality connections between audio and text.

Adding ICN alone led to a performance improvement to 61.91% on MELD and 65.32% on IEMOCAP, reflecting absolute gains of 2.15% and 1.43%, respectively, over the baseline. In ICN, each modality is processed independently via an encoder-decoder pathway, which transforms the raw input into a latent representation while preserving modality-specific information, such as prosodic variations in speech or syntactic emotional cues in text. The decoder reconstructs the input, ensuring that key structural and emotional features are retained.

Crucially, a connection loss is introduced to regularize the training process by encouraging shared representations between modalities, facilitating inter-modal alignment. Previous works [33], [36], [54] often treated each modality in isolation, whereas ICN provides a foundation for synergistic modeling by promoting semantic complementation.

2) Effect of Adding the Bimodal Attention Network: We then evaluated the effect of the BAN, which was designed to leverage dynamic attention weights to learn intra- and inter-modal interactions between audio and text. When BAN was added to the baseline model, the performance improved to 62.00% on MELD and 66.73% on IEMOCAP, corresponding to gains of 2.24% and 2.84%, respectively.

BAN applies both self-attention and cross-attention mechanisms to uni-modal representations, facilitating interaction within and between modalities. These dynamic attention weights selectively emphasize emotion-relevant signals while suppressing noise, enabling the model to build a more expressive bimodal representation. By capturing long-range dependencies and fusing information across modalities, BAN significantly improves emotion recognition.

BAN enables the model to leverage heterogeneous knowledge from each modality in a high-dimensional space, adaptively capturing complementary content and enhancing cross-modal correlation. This ability to focus on salient interactions is essential for accurate emotion classification.

3) Effect of Adding the Correlative Attention Network: The CAN contributes the largest standalone gain among all modules. Adding CAN to the baseline improves performance to 62.90% on MELD and 67.41% on IEMOCAP, yielding gains of 3.14% and 3.52%, respectively.

CAN is designed to filter out spurious or noisy cross-modal relationships while enhancing the reliable inter-correlations between uni-modal and bimodal representations. In contrast, CAN simultaneously considers both self- and cross-modal information. It utilizes a joint attention mechanism that incorporates both self- and cross-modal attention scores through paired softmax operations. This approach enables the model to refine incorrect correlations and dynamically balance the contribution of each modality.

The performance gains achieved by CAN indicate its importance in the final stages of feature integration, where high-quality fusion and representation refinement are critical.

4) Effect of Combining Modules: To further investigate the interplay between components of the BCAF framework, we conducted a series of experiments where modules were incrementally combined in pairs, as well as in full integration. The primary objective was to determine whether the joint contribution of two modules leads to a super-additive performance gain—i.e., whether their combination yields a larger improvement than the arithmetic sum of their individual contributions.

ICN + BAN: Integrating both ICN and BAN led to a performance of 63.69% on MELD and 68.01% on IEMOCAP, corresponding to absolute gains of 3.93% and 4.12% over the baseline. These values are slightly below the expected additive sums (4.39% on MELD, 4.27% on IEMOCAP), suggesting mild overlap in their functionalities. While ICN and BAN both improve feature learning and alignment, their integration still provides meaningful improvements.

ICN + CAN: When ICN and CAN were combined, the model achieved 64.83% on MELD and 69.57% on IEMOCAP, representing gains of 5.07% and 5.68%, respectively. These fall slightly short of the additive expectations (5.29% on MELD, on IEMOCAP), indicating good complementarity but not strictly synergistic effects.

BAN + CAN: This module pair demonstrates the strongest synergy. The model achieved 66.12% on MELD and 71.86% on IEMOCAP, equating to gains of 6.36% and 7.97%, respectively. These improvements are very close to or slightly above the additive sums (5.38% on MELD, 6.36% on IEMOCAP), suggesting that BAN and CAN reinforce each other effectively. BAN provides dynamic alignment, while CAN fine-tunes cross-modal representations via attention calibration.

All Three Modules (Full BCAF): Integrating the ICN, BAN, and CAN modules yields the highest performance, achieving weighted F1-scores of 68.93% on MELD and 74.69% on IEMOCAP. These results validate the effectiveness of the full architecture, which benefits from the complementary strengths and collaborative learning among the three components. Each module serves a distinct purpose: ICN captures modality-specific representations, BAN enables dynamic cross-modal alignment, and CAN performs high-level filtering and refinement to enhance the final predictions. Their combined operation allows the model to extract, integrate, and optimize multimodal information in a progressively structured manner, from low-level feature encoding to high-level semantic alignment.

Overall, our ablation results lead to two key conclusions. First, each module contributes significantly to model performance when added independently, with CAN contributing the most due to its powerful correlation-aware refinement capabilities. Second, combinations of modules tend to show synergistic behavior, especially the BAN+CAN pair, which exhibits the strongest interaction effect. This synergy reflects the natural pipeline design of the BCAF framework, where BAN enhances joint representations and CAN further calibrates them based on inter-modal relationships. The full integration of ICN, BAN, and CAN forms a coherent and mutually reinforcing architecture that enables robust and generalizable bimodal speech emotion recognition across datasets.

5) Comparison on Individual Noise Filtering scheme: We further compared our individual noise-filtering scheme with existing gating- (scalar-based [47]), reconstruction- (SMIN [54]), and attention- (MulT [19]) based approaches by incorporating them into the bare system. The results in the second part of Table III show that while traditional modules such as SMIN reconstruction, gating, and MulT attention can also improve the performance of the bare system, their gains are smaller than those achieved by our noise-filtering scheme. Specifically, for reconstruction-based approaches, Bare + ICN achieves 61.91% on MELD and 65.32% on IEMOCAP, which is +1.26% and +0.36% higher than Bare + SMIN, respectively. For gating-based approaches, Bare + BAN reaches 62.00% on MELD and 66.73% on IEMOCAP, outperforming Bare + Gate by +1.77% and +2.15%. For attention-based approaches, Bare + CAN obtains 62.90% on MELD and 67.41% on IEMOCAP, surpassing Bare + MulT by +1.15% and +2.40%. These results demonstrate that our proposed modules consistently outperform their traditional counterparts across both datasets.

C. Impact of Attention

The bimodal correlation evaluation explicitly measured similarity by incorporating correlation coefficients into the uni-modal representations. This process aligned the uni-modal features with the joint bimodal space, enhancing the model's ability to capture and leverage complementary information across modalities for more accurate emotion recognition.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2026.3699027

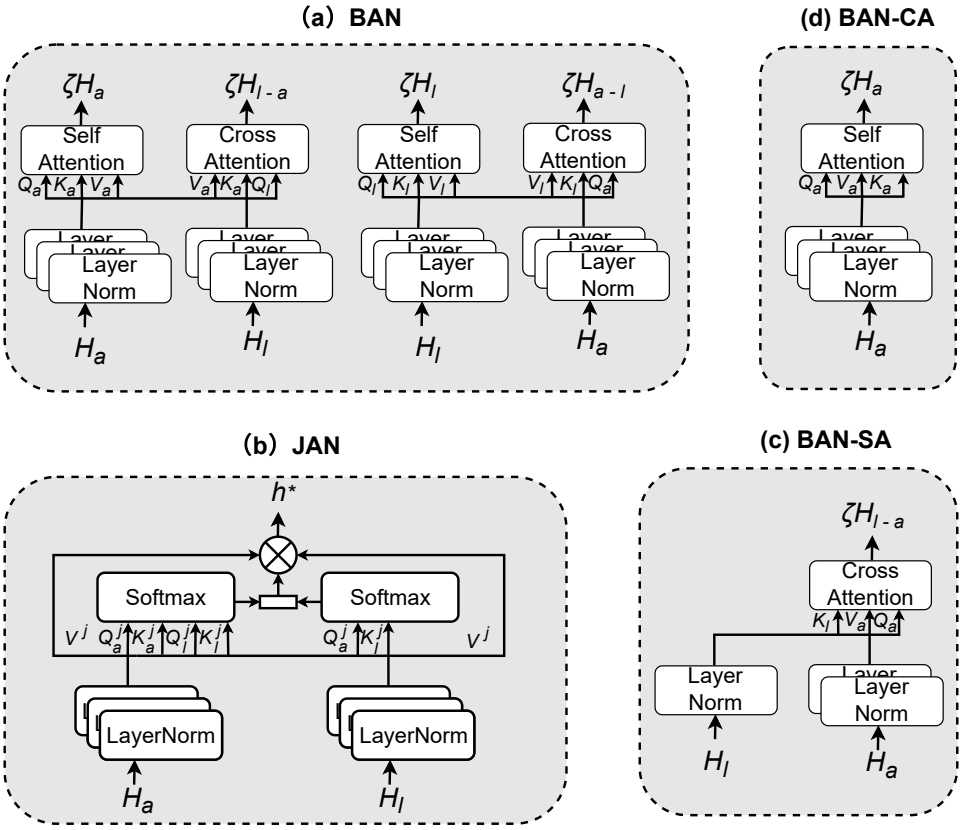


Fig. 7. Update scheme of different attention mechanisms.

TABLE IV
IMPACT OF DIFFERENT ATTENTION MECHANISMS ON THE BCAF MODEL (BOLD FONTS SHOWS BAN ACHIEVES THE BEST F1 SCORE).

Attention	Equation	w/o		Number of Layers	
		MELD	IEMOCAP	MELD	IEMOCAP
BAN	$\text{softmax}\left(\frac{Q_a K_a^\top}{\sqrt{d}}\right) V_a + \text{softmax}\left(\frac{Q_a K_l^\top}{\sqrt{d}}\right) V_a + \text{softmax}\left(\frac{Q_l K_l^\top}{\sqrt{d}}\right) V_l + \text{softmax}\left(\frac{Q_a K_l^\top}{\sqrt{d}}\right) V_l$	**64.83**	**69.57**	5	3
JAN	$\left(\text{softmax}\left(\frac{Q_a K_a^\top + Q_l K_l^\top}{\sqrt{d}}\right) - \lambda\,\text{softmax}\left(\frac{Q_a K_l^\top}{\sqrt{d}}\right)\right) V$	66.08	71.35	3	4
BAN-SA	$\text{softmax}\left(\frac{Q_a K_l^\top}{\sqrt{d}}\right) V_a$	67.21	72.16	7	5
BAN-CA	$\text{softmax}\left(\frac{Q_a K_a^\top}{\sqrt{d}}\right) V_a$	67.52	73.01	7	6

In this section, we examined the impact of different attention variants (see Fig. 7). For each modality, self-attention and cross-attention mechanisms were employed to capture intra- and inter-modal interactions between audio and text. To evaluate the effectiveness of the proposed attention mechanism in modeling these interactions, we implemented four comparison systems. The experimental results are presented in Fig. 6 and Table III.

- **Bimodal Attention Network (BAN):** This network employed separate softmax functions to independently apply self-attention and cross-attention weights, enabling distinct modeling of intra- and inter-modal interactions (see Fig. 7 (a)).
- **Joint Attention Network (JAN):** This network utilized a combined pair of softmax functions to simultaneously apply both self-attention and cross-attention weights, allowing for integrated modeling of intra- and inter-modal interactions (see Fig. 7 (b)).
- **Bimodal Attention Network - Self-Attention (BAN-SA):** This variant used only self-attention, omitting the cross-attention mechanism to focus solely on intra-modal interaction modeling (see Fig. 7 (c)).
- **Bimodal Attention Network - Cross-Attention (BAN-CA):** This variant used only cross-attention, omitting the self-attention mechanism to concentrate exclusively on inter-modal interaction modeling (see Fig. 7 (d)).

We observed that the BAN achieved the best performance on both the MELD and IEMOCAP datasets. As shown in

Fig. 6 and Table IV, the inclusion of the BAN contributed to a performance improvement of 3.6% on the MELD dataset and 6.6% on the IEMOCAP dataset compared to the baseline model without BAN. This comparison highlights the effectiveness of BAN in leveraging both intra-modal and inter-modal interactions between audio and text.

Additionally, the decline in performance followed a consistent trend: BAN $>$ JAN $>$ BAN-SA $>$ BAN-CA. This means that removing certain attention mechanisms resulted in progressively worse performance. Specifically, the absence of JAN caused a performance decrease of 2.85% on the MELD dataset and 3.34% on the IEMOCAP dataset. Similarly, excluding BAN-SA led to a decrease of 1.72% on MELD and 2.53% on IEMOCAP, while removing BAN-CA resulted in the smallest decline of 1.41% on the MELD and 1.69% on the IEMOCAP. These results underscore the relative importance of each attention mechanism in modeling effective intra- and inter-modal interactions between audio and text.

The bimodal attention network independently enabled complementary processing of intra-modal and inter-modal interactions between audio and text. We argue that intra-modal interactions provided the most essential information for accurate predictions, as they captured key modality-specific features directly tied to emotion recognition. Although inter-modal interactions added valuable complementary insights, core emotional cues were typically best represented within each individual modality, making intra-modal processing a primary factor in the model's effectiveness. Moreover, maintaining sufficient layers in the BAN was crucial for optimal performance. Reducing the number of layers negatively impacted the model's ability to process intra- and inter-modal interactions effectively, thereby degrading overall performance.

D. Case Studies

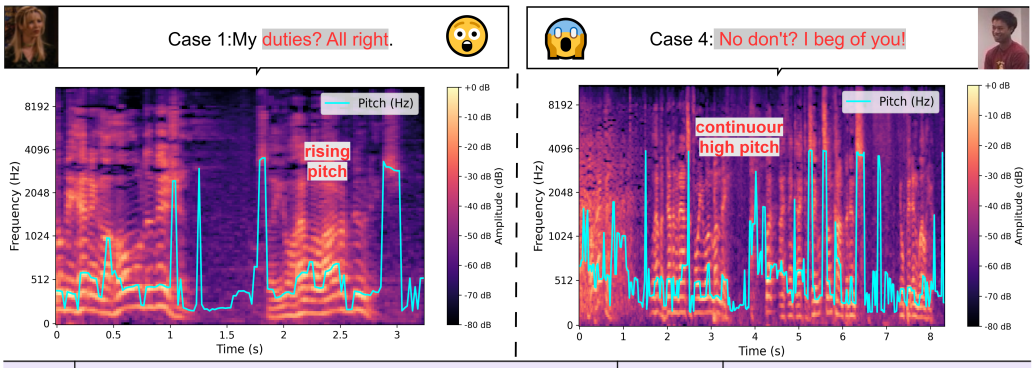
To illustrate the effectiveness of our proposed BCAF model, we selected samples from two datasets and conducted extensive experiments comparing BCAF with baseline models. The results, as shown in Fig. 8, revealed the following insights:

- In most cases, our BCAF model achieved correct predictions, whereas the baseline models failed. This result indicates that BCAF effectively integrates audio information with the text modality, enhancing bimodal speech emotion recognition. By exploring correlations and intra- and inter-modal interactions between audio and text, BCAF provides richer, emotion-relevant insights, enabling the modalities to complement and augment each other.
- We observed challenges in predicting minor emotions when speech with significant background noise was introduced into the text modality for prediction. Specifically, ambiguous expressions, such as the single word *"Um"*, were affected by unwanted and uncontrollable noise, further limiting the model's ability to comprehend human emotional expressions.
- Unexpectedly, while our proposed BCAF method outperformed the baseline models in terms of weighted F1-score, it performed worse than the state-of-the-art HCAM model. Upon further analysis, we attributed this to the emotional cues learned by the graph convolutional network in HCAM, which played a critical role in modeling and leveraging speaker information to capture intra- and inter-speaker dependencies for emotion recognition.

E. Error Analysis and Limitations

We present the confusion matrices for the MELD and IEMOCAP datasets in Fig. 9. As shown, although our BCAF model achieves significant improvements over previous methods (as discussed in Section V.A), it still exhibits several notable limitations, particularly on underrepresented emotional categories such as *disgust*, *fear*, and *happiness*. We further analyzed the causes of these errors and limitations from both data and model structure perspectives.

- Dataset limitations and annotation ambiguity: Similar to prior studies, our model's performance is affected by the subjective nature of emotion perception. In conversational contexts, emotions are often subtle, blended, or context-dependent. The annotations reflect perceived emotion rather than speaker intent, making it difficult for models—even with structured attention—to distinguish closely related emotions like *fear* and *sadness*. This ambiguity limits the ceiling performance for all models, including BCAF.
- Class imbalance and sparsity of rare emotions: The MELD and IEMOCAP datasets exhibit strong label imbalance, with neutral and negative emotions dominating. While BCAF introduces attention-guided fusion and multi-loss supervision, the current architecture does not include explicit mechanisms for addressing class imbalance (e.g., focal loss or reweighting). This results in overfitting to dominant emotion categories and underperformance on rare classes.
- Limitations in connection modeling via encoder-decoder: The interactive connection network is based on a symmetric encoder-decoder architecture to compress and reconstruct modality-specific representations. While effective in capturing high-level modality alignment, it relies on fixed fully connected layers, which may limit flexibility in modeling more nuanced or hierarchical modality interactions. Moreover, the learned modality connections are global and static across samples, potentially missing context-specific or local variations.
- Static attention strategy in the bimodal attention network: Although the bimodal attention network effectively models intra- and inter-modal interactions, its structure is fixed across all instances, using the same attention heads and layers regardless of the emotional content or conversation dynamics. This reduces adaptability to diverse emotional expressions and limits the model's performance in emotionally ambiguous or multi-intent utterances.
- Modal incompleteness and missing visual signals: BCAF focuses solely on audio and text modalities. While this choice simplifies the architecture and aligns with many real-world scenarios (e.g., voice assistants), it inherently limits emotion recognition capability. Visual cues such



Case	Spoken words	Acoustic features	Ground Truth	bc-LSTM	HCAM	BCAF
1	My duties? All right.		surprise	surprise√	surprise√	surprise√
2	Oh. God, crazy Chandler. He spun me...off...the...bed!		happy	natural	happy√	happy√
3	Wow! Spinning that sounds like fun.		happy	happy√	happy√	happy√
4	No don't? I beg of you!		fear	natural	sadness	natural√
5	What is it?		natural	natural	surprise	natural√
6	Um		anger	natural	natural	natural

Fig. 8. Inputs from bimodal data and predictions using bc-LSTM, HACM, and our proposed BCAF method on the MELD and IEMOCAP datasets are analyzed in our case study.

as facial expressions and gestures—which are often critical for disambiguating similar emotions—are excluded, constraining the model's representational richness.

- Independent training of uni-modal and bimodal branches: Our training strategy uses individual losses ($\mathcal{L}_a, \mathcal{L}_l, \mathcal{L}_m$) in addition to the connection loss $\mathcal{L}_c$, which encourages both modality-specific and fused representations. However, this independent supervision can lead to inconsistent optimization goals between branches, and we observe occasional conflicts where unimodal predictions disagree strongly with the fused output.

To address these limitations, future work could explore class-aware sampling or tailored loss functions for imbalanced emotion categories, enhance the encoder–decoder with recurrent or attention layers to capture dynamic dependencies, design adaptive attention mechanisms sensitive to conversational context, incorporate visual modalities to enrich emotional cues, and apply multi-task or consistency-based learning to better align uni-modal and bimodal objectives.

VI. Conclusion

We proposed the Bimodal Connection Attention Fusion (BCAF) method for efficient and robust bimodal speech emotion recognition. BCAF consists of three key modules: the interactive connection network, the bimodal attention network, and the correlative attention network. The interactive connection network enables the model to learn modality connections between audio and text. The bimodal attention network facilitates semantic complementation and optimally leverages intra- and inter-modal interactions between audio and text. Finally, the correlative attention network filters noise from cross-modal relationships and learns inter-correlations between uni-modal and bimodal representations. Experimental results demonstrated that BCAF outperforms state-of-the-art methods in bimodal speech emotion recognition task.

References

[1] W. Wei, B. Zhang, and Y. Wang, "Ts-mefm: A new multimodal speech emotion recognition network based on speech and text fusion," in *Proc. Int. Conf. Multimedia Model. (MMM)*. Doha, Qatar, 2025, pp. 454–467.

[2] D. Peña, A. Aguilera, I. Dongo, J. Heredia, and Y. Cardinale, "A framework to evaluate fusion methods for multimodal emotion recognition," *IEEE Access*, vol. 11, pp. 10 218–10 237, 2023.

[3] S. G. Upadhyay, W.-S. Chien, B.-H. Su, and C.-C. Lee, "Learning with rater-expanded label space to improve speech emotion recognition," *IEEE Trans. Affective Comput.*, vol. 15, no. 3, pp. 1539–1552, 2024.

[4] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 985–1000, 2021.

[5] Z. Lian, H. Chen, L. Chen, H. Sun, L. Sun, Y. Ren, Z. Cheng, B. Liu, R. Liu, X. Peng, *et al.*, "Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models," *arXiv preprint arXiv:2501.16566*, 2025.

[6] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition." in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2020, pp. 374–378.

[7] Z. Zhao, Y. Wang, G. Shen, Y. Xu, and J. Zhang, "Tdfnet: Transformer-based deep-scale fusion network for multimodal emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3771–3782, 2023.

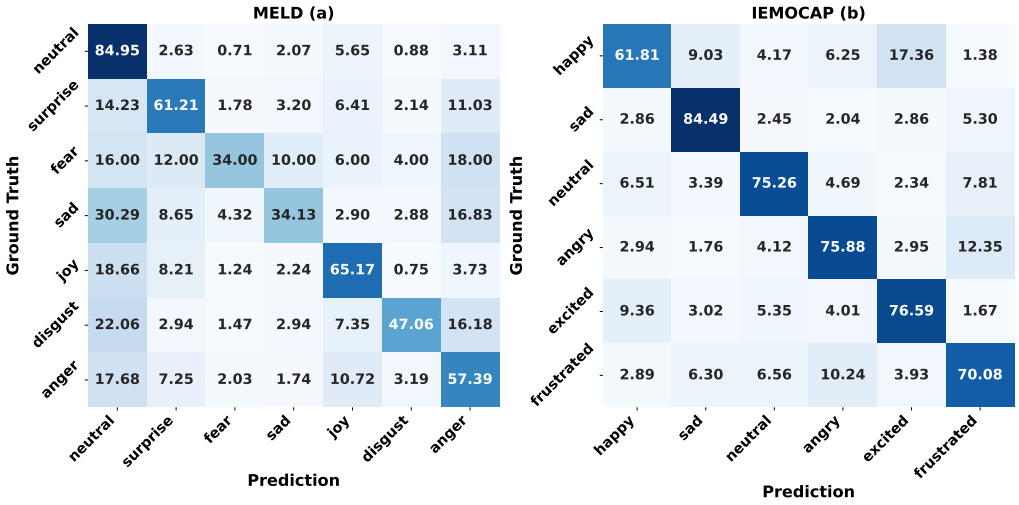


Fig. 9. Visualization of confusion matrices on the test sets of MELD (a) and IEMOCAP (b).

[8] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, "Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2018.

[10] M. Rasenberg, A. Özyürek, and M. Dingemanse, "Alignment in multimodal interaction: An integrative framework," *Cogn. Sci.*, vol. 44, no. 11, pp. 1–29, 2020.

[11] F. Li, J. Luo, L. Wang, W. Liu, and X. Sang, "Gcf2-net: Global-aware cross-modal feature fusion network for speech emotion recognition," *Front. Neurosci.*, vol. 17, pp. 1–14, 2023.

[12] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, and R. Xu, "Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis," in *Proc. 29th Int. Conf. Comput. Linguistics (COLING)*, vol. 29, no. 1. Gyeongju, Republic of Korea, 2022, pp. 7124–7135.

[13] R. Liu, H. Zuo, Z. Lian, H. Yuan, and Q. Fan, "Hardness-aware dynamic curriculum learning for robust multimodal emotion recognition with missing modalities," *arXiv preprint arXiv:2508.06800*, 2025.

[14] K. Ezzameli and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," *Inf. Fusion*, vol. 99, pp. 1–30, 2023.

[15] J. Luo, H. Phan, and J. Reiss, "Cross-modal fusion techniques for utterance-level emotion recognition from text and speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Rhodes Island, Greece, 2023, pp. 1–5.

[16] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: Multi-modal fusion network for emotion recognition in conversation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* New Orleans, Louisiana, USA, 2022, pp. 4652–4661.

[17] S. Zaidi, S. Latif, and J. Qadi, "Cross-language speech emotion recognition using multimodal dual attention transformers. arxiv 2023," *arXiv preprint arXiv:2306.13804*.

[18] Y. Luo, W. Liu, Q. Sun, S. Li, J. Li, R. Wu, and X. Tang, "Triagedmsa: Triaging sentimental disagreement in multimodal sentiment analysis," *IEEE Trans. Affective Comput.*, pp. 1–12, 2025.

[19] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*. Florence, Italy, 2019, pp. 6558–6569.

[20] M. Singh and S. A. Mehr, "Universality, domain-specificity and development of psychological responses to music," *Nat. Rev. Psychol.*, vol. 2, no. 6, pp. 333–346, 2023.

[21] T. Zhu, L. Li, J. Yang, S. Zhao, and X. Xiao, "Multimodal emotion classification with multi-level semantic reasoning network," *IEEE Trans. Multimedia*, vol. 25, pp. 6868–6880, 2022.

[22] S. Zhang, Y. Pan, and J. Z. Wang, "Learning emotion representations from verbal and nonverbal communication," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Vancouver, British Columbia, Canada, 2023, pp. 18 993–19 004.

[23] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, and A. Jabbar, "A review on methods and applications in multimodal deep learning," *ACM Trans. Multimedia Comput. Commun. Appl*, vol. 19, no. 2s, pp. 1–41, 2023.

[24] H. Zou, F. Lv, D. Zheng, E. S. Chng, and D. Rajan, "Large language models meet contrastive learning: Zero-shot emotion recognition across languages," *arXiv preprint arXiv:2503.21806*, 2025.

[25] Y. Li, M. Wang, M. Gong, Y. Lu, and L. Liu, "Fer-former: Multimodal transformer for facial expression recognition," *IEEE Trans. Multimedia*, vol. 27, pp. 2412 – 2422, 2024.

[26] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[27] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, vol. 4. Waikoloa, Big Island, Hawaii, USA, 2005, pp. 3437–3443.

[28] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces (ICMI)*. State College, Pennsylvania, USA, 2004, pp. 205–211.

[29] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Brisbane, Queensland, Australia, 2015, pp. 4749–4753.

[30] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, 2017.

[31] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput*, vol. 31, no. 2, pp. 153–163, 2013.

[32] S. Dutta and S. Ganapathy, "Hcam–hierarchical cross attention model for multi-modal emotion recognition," *arXiv preprint arXiv:2304.06910*, 2023.

[33] J. Li, X. Wang, Y. Liu, and Z. Zeng, "Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition," *IEEE Trans. Affective Comput.*, pp. 1–15, 2024.

[34] T. Yun, H. Lim, J. Lee, and M. Song, "Telme: Teacher-leading multimodal fusion network for emotion recognition in conversation," *arXiv preprint arXiv:2401.12987*, 2024.

[35] Y. Shou, T. Meng, F. Zhang, N. Yin, and K. Li, "Revisiting multi-modal emotion learning with broad state space models and probability-guidance fusion," *arXiv preprint arXiv:2404.17858*, 2024.

[36] Y. Shou, T. Meng, W. Ai, F. Zhang, N. Yin, and K. Li, "Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations," *Inf. Fusion*, vol. 112, p. 102590, 2024.

[37] Q. Shan, X. Wei, and Z. Cai, "Modality-invariant and -specific represen-

tations with crossmodal transformer for multimodal sentiment analysis," *J. Phys.: Conf. Ser*, vol. 2224, no. 1, p. 012024, 2022.

[38] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-P. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interaction (ICMI)*, 2021, pp. 6–15.

[39] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Singapore, 2022, pp. 7367–7371.

[40] S. Kakuba, A. Poulose, and D. S. Han, "Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model," *IEEE Access*, vol. 11, pp. 113 769–113 789, 2023.

[41] P. Ramkrishnan, R. K. Dhaman, and A. Poulose, "Feature importance and model performance in deep learning for speech emotion recognition," in *Proc. Int. Conf. Advances Comput. Commun. (ICACC)*. Kochi, Kerala, India, 2024, pp. 1–6.

[42] Y. Su, Y. Wei, W. Nie, S. Zhao, and A. Liu, "Dynamic causal disentanglement model for dialogue emotion detection," *IEEE Trans. Affective Comput.*, 2024.

[43] W. Nie, Y. Bao, Y. Zhao, and A. Liu, "Long dialogue emotion detection based on commonsense knowledge graph guidance," *IEEE Trans. Multimedia*, vol. 26, pp. 514–528, 2023.

[44] W. Nie, R. Chang, M. Ren, Y. Su, and A. Liu, "I-gcn: Incremental graph convolution network for conversation emotion detection," *IEEE Trans. Multimedia*, vol. 24, pp. 4471–4481, 2021.

[45] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pp. 1096–1103, 2008.

[46] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: https://arxiv.org/abs/1312.6114

[47] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[48] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://openreview.net/forum?id=HJxK5h5lx

[49] Y. Yin, L. Chen, F. Ren, Y. Wang, R. Wang, and Y. Liu, "A novel multimodal emotion recognition approach using multimodal feature fusion and gated recurrent units," *IEEE Access*, vol. 7, pp. 173 055–173 066, 2019.

[50] M. Ma, Z. Meng, J. Xu, J. Zhang, and D. Tao, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 863–872.

[51] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1409.0473

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[53] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Inf. Fusion*, vol. 95, pp. 306–325, 2023.

[54] Z. Lian, B. Liu, and J. Tao, "Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2415–2429, 2022.

[55] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 12 449–12 460, 2020.

[56] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[57] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[58] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[59] Y. Wang, Y. Gu, Y. Yin, Y. Han, H. Zhang, S. Wang, C. Li, and D. Quan, "Multimodal transformer augmented fusion for speech emotion recognition," *Frontiers in Neurorobotics*, vol. 17, p. 1181598, 2023.

[60] S. A. M. Zaidi, S. Latif, and J. Qadir, "Enhancing cross-language multimodal emotion recognition with dual attention transformers," *IEEE Open J. Comput. Soc*, vol. 1, pp. 1–11, 2024.

[61] E. Dikbıyık and Others, "Bimer: Design and implementation of a bimodal emotion recognition system enhanced by data augmentation techniques," *IEEE Access*, vol. 13, pp. 123 456–123 465, 2025.

[62] W. Fan, X. Xu, G. Zhou, X. Deng, and X. Xing, "Coordination attention based transformers with bidirectional contrastive loss for multimodal speech emotion recognition," *Neurocomputing*, vol. 512, pp. 456–467, 2025.