Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Reliable Multimodal Conversational Emotion Recognition Model Based on a Dual-Stage Attention Mechanism

Xinheng Wang
University of Science and Technology Beijing
Lun Xie
xielun@ustb.edu.cn
University of Science and Technology Beijing
Xurong Xie
Chinese Academy of Sciences
Xiaolan Peng
Chinese Academy of Sciences
Zhiliang Wang
University of Science and Technology Beijing

Research Article

Keywords: Emotion Recognition in Conversation, Multimodal fusion, Attention mechanism, Label noise

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10210405/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Reliable Multimodal Conversational Emotion Recognition Model Based on a Dual-Stage Attention Mechanism

Xinheng Wang[1], Lun Xie[1]*, Xurong Xie[2]*, Xiaolan Peng[2], Zhiliang Wang[1]

1. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
2. Institute of Software, Chinese Academy of Sciences & University of Chinese Academy of Sciences, Beijing 100190, China
* Corresponding author.
E-mail address: d202210370@xs.ustb.edu.cn (X. Wang), xielun@ustb.edu.cn (L. Xie), xurong@iscas.ac.cn (X. Xie), xiaolan@iscas.ac.cn (X. Peng), wzl@ustb.edu.cn (Z. Wang)

Abstract

Emotion Recognition in Conversation (ERC) plays a critical role in enhancing the intelligence and naturalness of human-computer interaction systems. However, existing ERC methods still face significant challenges, particularly in the effective integration of multimodal information and the handling of label noise. To address these challenges, this paper proposes a reliable multimodal conversational emotion recognition model based on a dual-stage attention mechanism (DSA-MERC). DSA-MERC is designed to improve the accuracy and robustness of emotion recognition. In the first stage, the model performs the initial fusion of text, audio, and visual modalities using the Bi-Directional MultiAttention (Bi-MultiAttn) cross-attention mechanism. In the second stage, recognizing the importance of the audio modality in emotional expression, we introduce an audio-guided leader-follower attention network (Audio-LFAN) with a Multi-Scale Attention with Unit Force Operated (MAS-UFO) mechanism. The MAS-UFO framework optimizes multimodal emotion features at multi-granularity levels through multi-scale representation learning, significantly enhancing the model's ability to capture emotional cues. To address the issue of label uncertainty, we incorporate both Confidence Learning (CL) and Transfer Learning strategies, which effectively filter noisy labels and further enhance the generalization and stability of the DSA-MERC. Experimental results demonstrate that, when compared to existing baseline models, the proposed method yields substantial performance improvements across two widely recognized benchmark datasets, validating its effectiveness and superiority in addressing emotion recognition in conversation task.

Keywords: Emotion Recognition in Conversation; Multimodal fusion; Attention mechanism; Label noise;

Declaration of Interest statement
Competing Interests: The authors have no competing interests to declare that are relevant to the content of this article
Acknowledgments ＆Funding information
This work is supported by National Science and Technology Major Project (2022ZD0118002).

1 Introduction

Emotion Recognition in Conversation (ERC) represents a significant research focus within Artificial Intelligence (AI) and Natural Language Processing (NLP) [1]. The primary goal of ERC is to enhance the naturalness and effectiveness of human-computer interaction by analyzing conversational content and identifying and interpreting the emotional states of participants [2].

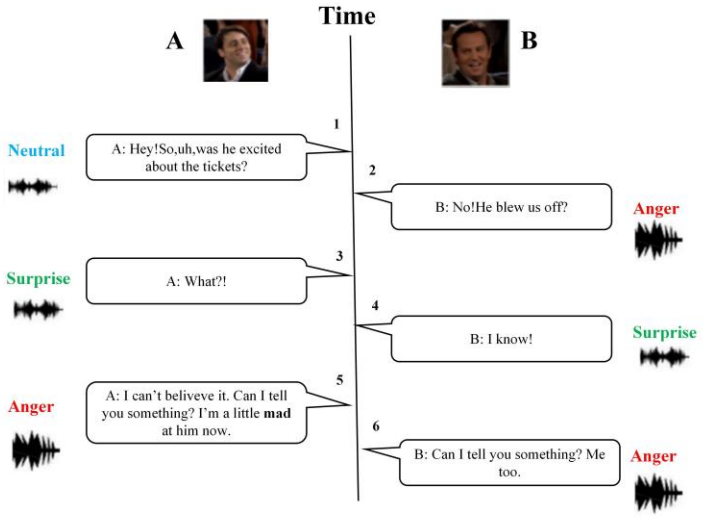


Fig. 1. A multimodal dialog scene with emotional information from three modalities: textual, acoustic, and visual.

Emotion information in a dialogue represented not only by the contextual content of text but also by various modalities during the interaction. These non-textual cues may include expressions or micro-expressions of the interlocutors, tone of voice, speech rate, and more. As illustrated in the multimodal conversation scenario in Fig. 1, affective information may stem from three modalities: textual, acoustic, and visual. Depending on the conversation scenario [3], ERC can be further categorized into unimodal and multimodal scenarios. In the unimodal scenario, the system relies solely on a single modality for emotion recognition. In contrast, in the multimodal scenario, the system captures the emotional states of participants more comprehensively by integrating affective cues from multiple modalities, which enhance the accuracy and robustness of emotion recognition [4][5]. Multimodal ERC now plays a critical role in diverse applications, including opinion mining, user emotional state analysis in social media, and the development of empathic dialog systems to foster emotional understanding. Despite notable advances in multimodal ERC, existing methods still encounter numerous challenges.

First, the effectiveness of multimodal information integration still faces significant limitations. Currently, most ERC methods primarily rely on simple feature concatenation when integrating multimodal information [6]. This basic concatenation approach fails to adequately model the complex associations between different modalities, hindering the generation of high-quality multimodal representations and, consequently, limiting emotion recognition accuracy [7]. In contrast, some advanced ERC

models employ attention-based multimodal fusion strategies, which focus on capturing complementary information across modalities to enhance their interactions. However, while these methods improve modal integration to some extent, they are prone to distraction issues when handling long conversational sequences or high-dimensional modal data, making it challenging to effectively emphasize key emotional cues [8]. Additionally, many ERC methods overly rely on textual modalities, often overlooking crucial emotional cues in speech, such as intonation, speech rate, and volume fluctuations, which are essential for capturing the speaker's emotional state. This limited emphasis on speech cues further reduces the effectiveness of emotion recognition.

Second, existing multimodal emotion recognition methods also suffer from high computational complexity when processing long sequential data. Most current methods employ the Self-Attention (SA) mechanism to capture global dependencies within sequences [9]. However, the computational complexity of the Self-Attention mechanism increases exponentially with sequence length[10]. This significantly raises the model’s training and inference time. It also often results in overly dispersed attention weights, hindering efficient focus on key emotional cues. Consequently, developing an efficient and robust multimodal ERC model within the constraints of limited data and computational resources has become an urgent challenge.

In addition, labeling uncertainty and label noise present a major challenge in multimodal ERC [11]. The process of acquiring multimodal emotion data is complex and costly, resulting in generally small ERC datasets. However, the limited amount of data can limit the training effectiveness and generalization ability of the model. Moreover, the annotation of emotion labels is highly subjective. Different annotators may have substantial differences in interpreting the same sentiment emotion expression, leading to annotation uncertainty and noise [12]. Emotions are inherently ambiguous and complex, and many affective states often exhibit diverse expressions, further exacerbating labeling uncertainty. Although self-assessment by conversation participants may be a preferable labeling method, it is challenging to implement in practice [13]. Therefore, ERC models must address labeling uncertainty by effectively filtering or adapting to noise in the data, thereby enhancing the accuracy of emotion recognition.

To address the above challenges, this paper proposes a reliable multimodal conversational emotion recognition model based on a dual-stage attention mechanism (DSA-MERC), which includes the following key components:

First, to address the lack of fully exploited multimodal complementarity, this paper proposes the Auditory-Guided Dual Attention Multimodal Fusion Mechanism. In the first stage, the Bi-Directional MultiAttention Mechanism (Bi-MultiAttn) enables interactive modeling of textual, visual, and preliminary auditory features to achieve initial multimodal integration. Recognizing the significance of the auditory modality in emotional expression, we introduce an Audio-Guided leader-follower attention network (Audio-LFAN) in the second stage. This mechanism refines auditory features using the Multi-Scale Attention with Unit Force Operated (MSA-UFO Attention) proposed herein, extracting both global and local key emotional features in

multi-granularity, and establishing the auditory modality as the “leader” in multimodal fusion. Correspondingly, textual, visual, and unoptimized auditory features serve as “follower” features, participating in deep multimodal fusion alongside optimized auditory features to improve the model’s capacity to capture emotional cues.

This “Leader-Follower” architecture effectively enhances the prominence of the audio modality in emotion recognition, particularly in capturing emotional cues. By integrating multi-scale learning, deep convolution, and the UFO Attention based on Xnorm module, the MSA-UFO Attention proposed in this paper not only addresses the computational limitations of traditional self-attention mechanisms in processing long sequential data but also optimizes attention weight distribution, enabling the model to capture key affective features with greater accuracy. Specifically, we propose a Unit Force Operation with L2 normalization properties as a replacement for the Softmax function in the self-attention mechanism. This operation reduces the computational complexity associated with the exponential operations of the Softmax function by normalizing input sequences into unit vectors across spatial and channel dimensions, while also mitigating numerical instability issues. In addition, the proposed DSA-MERC leverages Confident Learning (CL) and Transfer Learning strategies to address label noise in the dataset. The CL mechanism identifies and filters mislabels and is combined with Transfer Learning to further train on the cleaned data, significantly enhancing the model’s generalization ability and stability. This strategy ensures that the model remains robust and maintains effective learning when faced with unbalanced and noisy datasets.

Finally, we conducted comprehensive experiments using the MELD and IEMOCAP ERC benchmark datasets. The experimental results show that the proposed DSA-MERC model improves the weighted average w-F1 scores by 1.51%-23.20% and 0.51%-10.09% on the IEMOCAP and MELD datasets, respectively, compared to existing state-of-the-art methods. These results demonstrate the model’s effectiveness and performance benefits in emotion recognition tasks.

The main contributions of this paper are as follows: (1) We propose a reliable multimodal conversational emotion recognition model based on a dual-stage attention mechanism, which enhances multimodal information interaction and fusion. The two-stage multimodal fusion strategy addresses limitations in current multimodal information integration. Additionally, the confident learning strategy effectively mitigates label noise interference, improving emotion recognition accuracy. (2) We systematically study and analyze the model’s key components, including the Auditory-Guided Dual Attention Multimodal Fusion Mechanism, the MSA-UFO Attention mechanism, and the confident learning strategy, validating the roles of these components in enhancing model robustness and performance. (3) Extensive experiments on two public datasets, IEMOCAP and MELD, demonstrate that the proposed model outperforms current state-of-the-art models in performance. Specifically, the w-F1 metrics of the proposed model reach 71.3487 and 65.9877, respectively, demonstrating its effectiveness in emotion recognition tasks.

The rest of the paper is organized as follows: Section II presents an overview of related work; Section III provides a detailed description of the proposed model and methodology; Sections IV and V present the experimental setup and detailed result

analysis; and finally, Section VI summarizes the findings and discusses future research directions.

2 Related Work

2.1 Emotion recognition in conversations

Broadly, research in this ERC field can be divided into recurrence-based, transformer-based, graph-based, and multimodal-based models according to the modeling approach.

Recurrence-based Models: Due to the significant advantages of recursive models in sequential tasks, researchers usually choose Recursive Neural Networks (RNN) to process sequential data in conversations. Recursive models effectively capture temporal dependencies and contextual relationships by processing each utterance in a conversation step-by-step, thus enabling the recognition of affective states. For example, Poria et al. [14] used LSTM to extract the features of each contextual utterance and fused the contextual information through a fully connected layer to enhance dialog emotion recognition. However, the method fails to fully consider the interdependence between sentences, i.e., each utterance is not only affected by its own history information, but also closely related to other utterances in the context. To further optimize the modeling capability of conversation context, Chen et al. [15] proposed a multi-level RNN-based approach based on the study of Poria et al. [14]. The method utilizes a hierarchical recursive structure to encode the utterances and more accurately extract the sentiment information in the dialogue. Majumder et al. [16] further analyzed the main influences on the sentiment of a dialogue through GRU, and pointed out that the sentiment expression is mainly affected by three aspects: the identity of the speaker, the contextual information of the previous sentence, and the sentiment information of the previous sentence. Jiao et al. [17] further pointed out that traditional GRU networks in the literature [16] are difficult to effectively capture remote contextual information, and proposed a hierarchical gated recurrent unit (HiGRU) framework. The framework effectively improves the performance of discourse-level sentiment recognition by modeling word sequences and contextual relations respectively through a hierarchical GRU structure.

Transformer-based Models: Although recursive models excel in capturing contextual dependencies in conversations, their computational efficiency is limited. This has driven researchers to explore alternative models, such as Transformer-based models, to enhance sentiment recognition performance further. Unlike recursive models, Transformer models do not depend on sequential processing; instead, they focus on features at all positions within a sequence simultaneously via an attention mechanism, thereby constructing global contextual relationships more efficiently. For example, Zhu et al. [18] proposes a Topic-Driven Knowledge-Aware Transformer model. The model integrates a topic-enhanced language model with commonsense information extracted from a knowledge base, utilizing the Transformer architecture to fuse both topic and commonsense

information, significantly improving the accuracy and discriminative power of sentiment classification. Li et al. [19] presents EmoCaps, an end-to-end Transformer-based model for conversation sentiment recognition. It extracts sentiment capsules from diverse modal features and efficiently captures sentiment tendencies and contextual information using a multi-head attention mechanism. Liu et al. [20] combine the strengths of both attention-based and recursion-based approaches at the feature extraction level. Global contextual dependencies are captured through the attention mechanism, whereas GRU is employed to extract speaker- and time-aware local contextual information.

Graph-based Models: In recent years, graph-based models have received much attention in dealing with complex relational structure tasks. Such models can effectively retain global structural information and are particularly suitable for application in the field of dialog sentiment recognition. Unlike RNNs, which only model sequential data cyclically, GNNs treat the conversational sentiment recognition task as a connection and message propagation problem between nodes, and capture rich graph structural information by modeling the dependencies between nodes. For example, Ghosal et al. [21] first utilized Graph Convolutional Networks (GCNs) to efficiently capture dependencies and long-distance contextual information between speakers. In this case, the nodes in the graph represent individual discourse, and the edges between the nodes indicate the dependencies between speakers and their relative positions in the dialog. However, Ishiwatari et al. [22] point out that although Ghosal et al.'s approach captures contextual dependencies in a conversation, it ignores the sequential information of the discourse. Therefore, Ishiwatari et al. proposed a Relation-aware Graph Attention Networks (RAGAN) that enhances the model's ability to model both discourse sequential information and speaker dependencies by introducing relational position coding. In addition, some researchers have also combined the attention mechanism with graph networks to obtain significant performance improvements. For example, Li et al. [23] proposed a two-stage multi-source information fusion ERC model based on graph and attention mechanisms, which utilizes a graph network to capture structured affective information and employs a cross-modal fusion network based on a multi-head attention mechanism to achieve intra- and intermodal affective interactions. Lu et al. [24] creatively designed a two-stream graph ERC network, which captures intra-modal and inter-modal affective interactions through two parallel graph networks respectively. long-distance contextual information and inter-modal interactions through two parallel graph networks, which also achieved excellent experimental results.

Multimodal-based Models: Traditional studies on Emotion Recognition in Conversation have primarily focused on textual modalities, aiming to improve emotion recognition accuracy by capturing contextual cues and speaker dependencies. However, these approaches often overlook the potential contributions of other modalities and their essential roles in emotion understanding. With the advancement of research, an increasing number of studies have shifted attention to multimodal ERC, with a focus on exploring intra- and inter-modal interactions and enhancing the representation of salient modalities to improve recognition performance [3]. The core challenge of multimodal ERC lies in the effective fusion of information from different

modalities in order to capture more comprehensive emotional representations. In recent years, advanced multimodal ERC methods have widely adopted attention-based deep learning models for feature fusion, aiming to efficiently integrate multimodal interaction features and enhance both the accuracy and robustness of emotion recognition. For example, Hu et al. [25] proposes Multimodal Dynamic Fusion Networks (MM-DFN) for capturing multimodal contextual dynamics via graph-structured dynamic fusion modules to reduce information redundancy and enhance inter-modal complementarity. Zou et al. [26] proposes the Master Modal Transformer (MMTr) to strengthen weak modal representations and improve multimodal fusion using the multi-head attention mechanism. Ma et al. [27] designs a hierarchical gated fusion strategy to dynamically learn modal weights and capture interaction information among modalities using intra- and cross-modal Transformer structures. Lian et al. [28] proposes CTNet, a multimodal learning framework that uses the Transformer structure to model intra- and cross-modal interactions, combining bi-directional GRUs and speaker embeddings to capture context and speaker dependencies. Shi et al. [29] captures cross-modal mappings of text, audio, and visual modalities through a bidirectional multi-head cross-attention layer, effectively integrating multimodal information to improve emotion recognition accuracy.

2.2 Attention Mechanism

Transformer models have been widely used across various fields to capture global dependencies in sequences through the Self-Attention mechanism. However, the traditional Self-Attention mechanism encounters several challenges when processing long sequences [9]. First, the computational complexity grows exponentially with sequence length, substantially raising computational resource demands. Second, SA attention weights are easily dispersed in long sequences, making it difficult to focus on key local information. Furthermore, although SA captures global dependencies effectively, it lacks the ability to utilize local context, which limits the model's understanding of detailed information.

To address these issues, researchers have proposed various approaches to improve the Transformer model. For instance, Zhang et al. [30] introduces an efficient multi-scale visual Transformer model, ResT. The ResT utilizes deep convolution techniques for memory compression and develops a memory-efficient multi-head self-attention mechanism, enabling effective interactions across attention head dimensions while preserving the diversity of multi-heads. Zhao et al. [31] further proposes a parallel multi-scale attention mechanism, termed MUSE Attention. MUSE Attention processes sequence data across multiple scales in parallel by integrating self-attention and deep convolution techniques. The method effectively captures both global dependencies and local contextual information in the sequence, thereby significantly improving sequence task performance. Additionally, Song et al. [32] proposes the Unit Force Operated Vision Transformer (UFO-ViT), which introduces a cross-normalization-based Xnorm module to replace Softmax in self-attention. This method removes nonlinearity in the original self-attention mechanism, substantially reducing computational resource requirements. These improvements effectively address

computational complexity challenges in long sequence tasks and further expand the applicability of the Transformer model across various tasks

2.3 Addressing Label Uncertainty

The process of annotating emotion labels is challenging because labels often depend on the annotator's subjective judgment [33]. Due to the complexity of emotion labeling and the difficulty of data acquisition, most ERC datasets are small and often accompanied by label noise [34]. In such cases, the model may suffer from overfitting from insufficient data or reduced robustness and stability due to noisy labels.

To address these issues, researchers usually introduce new models or optimize the loss function. For example, Self-Ensemble Label Filtering (SELF) proposed in Nguyen et al. [35] identifies and filters potentially erroneous labels by matching labels to maximum likelihood predictions. The Probabilistic End-to-end Noise Correction for Learning with Noisy Labels (PENCIL) framework proposed in Yi et al. [36] dynamically updates label distributions to correct noisy labels. In addition, Zhang et al. [37] proposes the Generalized Cross Entropy (GCE) loss function, which combines the advantages of Mean Absolute Error (MAE) and conventional Cross Entropy (CCE) to increase the model’s noise robustness. However, these methods mainly address label noise by improving the model or the loss function, without fundamentally solving the dataset’s label noise problem. By comparison, Confident Learning [38] is a denoising method independent of specific data modalities or models, focusing on label quality. It identifies and corrects mislabeling by analyzing the model-predicted joint confusion matrix to generate cleaner and more reliable datasets.

3 Methodology

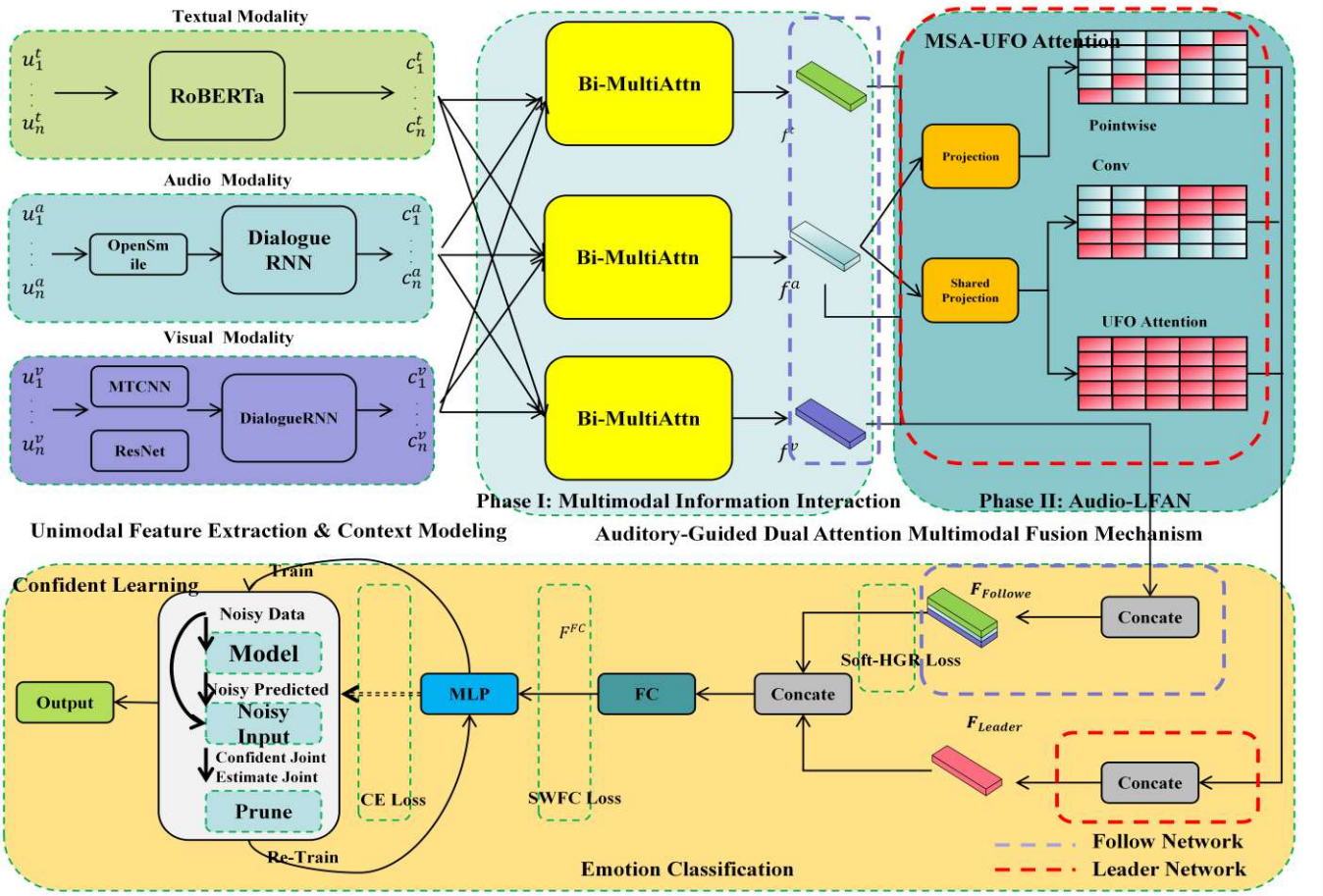


Fig. 2. Schematic of the general framework of the proposed DSA-MERC model, which includes three key parts: Unimodal feature extraction and context modeling, Auditory-Guided dual attention multimodal fusion mechanism, and Emotion classification. And the FC represents Fully-Connected layer, and the MLP stands for Multilayer Perceptron.

3.1 Task Definition

A conversation is composed of n consecutive discourses $\{u_1,\ldots,u_n\}$ and corresponding speakers $\{S_{u_1},\ldots,S_{u_n}\}$. We define the textual, audio, and visual modal sequences of all discourses $\{u_1,\ldots,u_n\}$ in the dialogue as $\{u_1^t,\ldots,u_n^t\}$, $\{u_1^a,\ldots,u_n^a\}$, and $\{u_1^v,\ldots,u_n^v\}$, respectively. The goal of the Emotion Recognition for Conversation (ERC) task is to predict the emotion label of each discourse u from a predefined set of emotion categories K.

3.2 Model architecture

As shown in Fig. 2, the overall architecture of the DSA-MERC model proposed in this paper is introduced and divided into the

following components: Unimodal feature extraction and context modeling, Auditory-Guided dual attention multimodal fusion mechanism, and Emotion classification. Specifically, we propose a reliable multimodal conversational emotion recognition model based on a dual-stage attention mechanism, which enhances multimodal information interaction and fusion. The two-stage multimodal fusion strategy addresses limitations in current multimodal information integration. Additionally, the confident learning strategy effectively mitigates label noise interference, improving emotion recognition accuracy. The following sections detail each component.

3.3 Unimodal Feature Extraction and Context Modeling

Textual Modality: We follow the approach in Kim et al. [39] to jointly capture local discourse and global context by fine-tuning the pre-trained RoBERTa model. The discourse is divided into past, current, and future parts, resulting in a 256-dimensional textual feature representation through [*SEP*] tagging connection technique.

Audio Modality: For audio feature extraction, we use the OpenSMILE toolkit [40] to extract audio feature vectors for each discourse and the DialogueRNN [16] to model audio features in context, resulting in a 256-dimensional feature representation.

Visual Modality: We follow the approach in Shi [29], which uses MTCNN [41] to detect faces in video frames and extracts visual features using a pre-trained ResNet-101 model. Using this method can effectively reduce the interference of background information on the key features of the face, thus enhancing the extraction of key features. These features are then contextually modeled using the DialogueRNN model to generate a 256-dimensional visual feature representation.

3.4 Phase I: Multimodal information interaction

As shown in Fig. 3, in the first stage of multimodal information fusion, we adopt the designed Bi-MultiAttn model to capture the complex associations among text, audio, and visual modalities. The structure of Bi-MultiAttn is inspired by the transformer [6] model. Specifically, Bi-MultiAttn includes Bi-MultiAttntext, Bi-MultiAttnaudio, and Bi-MultiAttnvisual, which have a shared model architecture. Effective integration and enhanced complementarity of multimodal information are achieved through the bi-directional multi-head cross-attention mechanism. In addition, we employ residual connections and layer normalization in the output of each stage to facilitate training and build a deeper and more robust network structure.

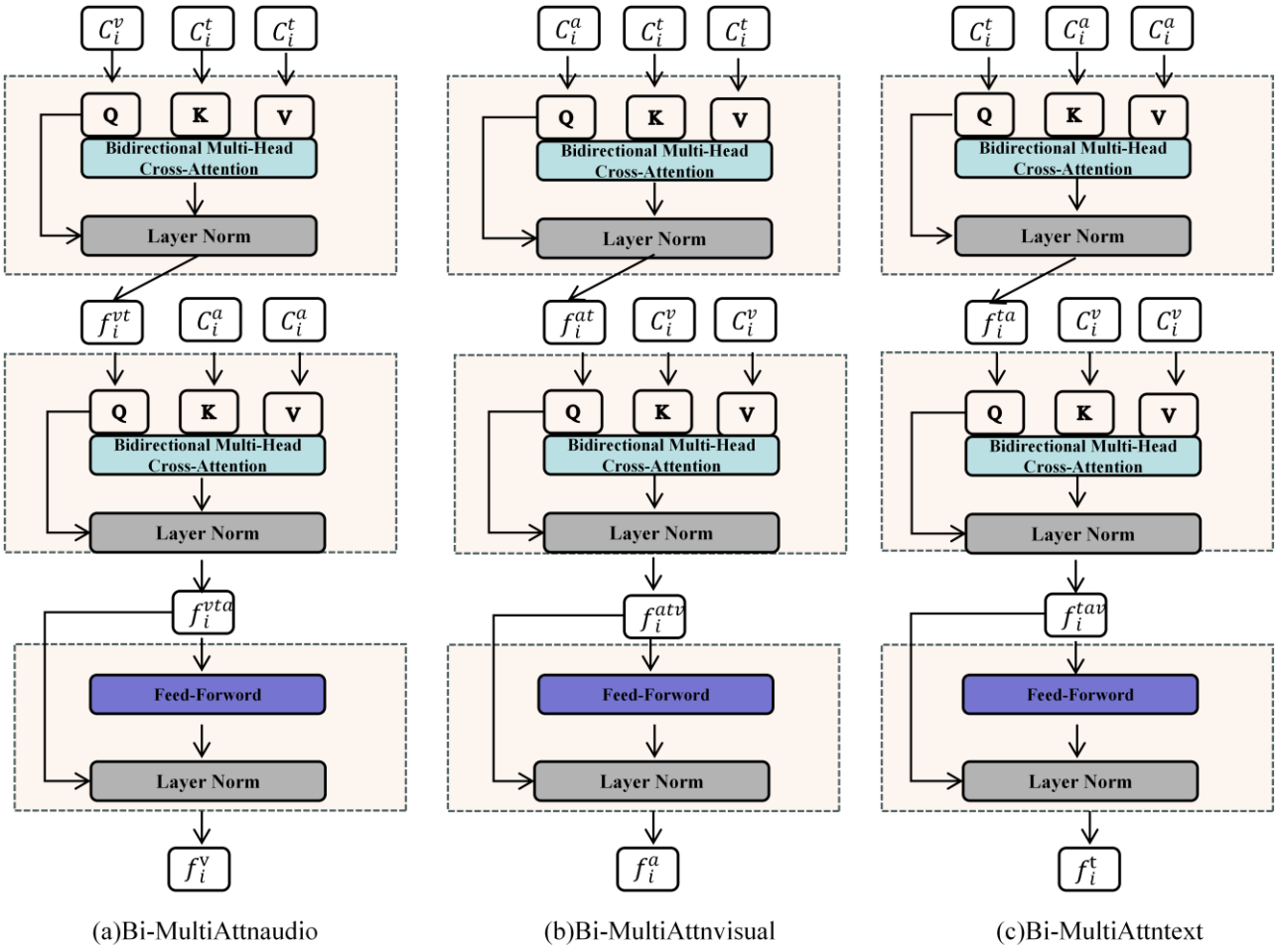


Fig. 3. The first stage of the Auditory-Guided dual attention multimodal fusion mechanism proposed in this paper. This stage consists of the Bi-MultiAttn mechanism, comprising Bi-MultiAttnaudio (a), Bi-MultiAttnvisual (b), and Bi-MultiAttntext (c), which share the same model architecture. The figure shows a schematic of the single-layer architecture, while our model consists of multiple layers stacked together.

Using the audio modality as an example, we illustrate how the Bi-MultiAttnaudio model at layer j combines the audio modality with textual and visual cues to integrate information from different modalities and enhance multimodal complementarity. The process for video and text modalities is similar. The calculations proceed as follows: as shown in (1) and (2), the audio and text modalities are first fused. The query (Q) for the audio modality and the key (K) and value (V) for the text modality are generated through linear transformation, after which weights are calculated and a weighted output is produced using the scaled dot-product attention mechanism. Subsequently, the outputs of all attention heads are spliced using (3)–(5), and preliminary fusion features are obtained by linear transformation and residual concatenation. Next, this fusion result serves as the Q for cross-modal fusion with the K and V from the visual modality to generate the final fusion representation, as shown in (6). In the multimodal fusion process, multihead-attention weights are first computed to generate an output, which is then further processed through residual concatenation and linear transformation. The fused features are then fed into a two-layer fully connected feedforward network with ReLU activation at each layer. Finally, the outputs of the feedforward network are

connected to the inputs through residual connections and normalized with LayerNorm to generate the final multimodal fused features, as in (7)–(9).

$$[Q_{at_h^{(j)}}, K_{at_h^{(j)}}, V_{at_h^{(j)}}] = [F_a^{(j-1)} W_{Q_{at_h^{(j)}}}, C_t W_{K_{at_h^{(j)}}}, C_t W_{V_{at_h^{(j)}}}], h \in \{1,...,H\} \tag{1}$$

$$A_{at_h^{(j)}} = Soft\max(\frac{Q_{at_h^{(j)}} K_{at_h^{(j)}}^T}{\sqrt{d_k}}) V_{at_h^{(j)}}, h \in \{1,...,H\} \tag{2}$$

$$MH_{at^{(j)}} = Cat(A_{at_1^{(j)}},...,A_{at_H^{(j)}}) W_{Oat^{(j)}} \tag{3}$$

$$F_{at^{(j)}} = LayerNorm(F_t^{(j-1)} + MH_{at^{(j)}}) \tag{4}$$

$$[Q_{atv_h^{(j)}}, K_{atv_h^{(j)}}, V_{atv_h^{(j)}}] = [F_{at^{(j)}} W_{Q_{atv_h^{(j)}}}, C_v W_{K_{atv_h^{(j)}}}, C_v W_{V_{atv_h^{(j)}}}], h \in \{1,...,H\} \tag{5}$$

$$A_{atv_h^{(j)}} = Soft\max(\frac{Q_{atv_h^{(j)}} K_{atv_h^{(j)}}^T}{\sqrt{d_k}}) V_{atv_h^{(j)}}, h \in \{1,...,H\} \tag{6}$$

$$MH_{atv^{(j)}} = Cat(A_{atv_1^{(j)}},...,A_{atv_H^{(j)}}) W_{Oatv^{(j)}} \tag{7}$$

$$F_{atv^{(j)}} = LayerNorm(F_{at}^{(j)} + MH_{atv^{(j)}}) \tag{8}$$

$$F_{a^{(j)}} = LayerNorm(F_{atv^{(j)}} + FFN_{a^{(j)}}) \tag{9}$$

where $F_a^{(j-1)}$ is the audio modal output of layer $j-1$, $C_t = \left[c_1^t,...,c_n^t\right]$ is the text modal feature, W is the projection matrix, h is the number of attention heads, d_k is the dimensionality of the key vectors, *LayerNorm* represents layer normalization, $F_{at^{(j)}}$ is the text feature after the audio modal fusion, $C_m, m \in \left[v,t,a\right]$ is the three modal features, and $FFN(\cdot)$ represents the feedforward network.

3.5 Phase II: Audio-Guided Leader-Follower Attention Network

In this paper, we propose the Audio-Guided Leader-Follower Attention Network in the second stage of multimodal fusion. As shown in Fig. 2, the network consists of a leader network and a follower network. The leader network captures local context and global dependencies in audio features by employing multi-scale representation learning through the MSA-UFO Attention mechanism proposed in this paper. The follower network, on the other hand, integrates the audio, text, and visual features generated by the Bi-MultiAttn model in the first stage through feature fusion to achieve deep fusion of multimodal information.

Multi-Scale Attention with Unit Force Operated: The core of Audio-LFAN is the efficient capture of key emotional regions in audio features. Inspired by Busso et al. [34], this paper proposes a method called Multiscale Attention with Unit Force Operated mechanism (MSA-UFO Attention) to better extract emotional features in audio. The composition and working principles of this mechanism are detailed below.

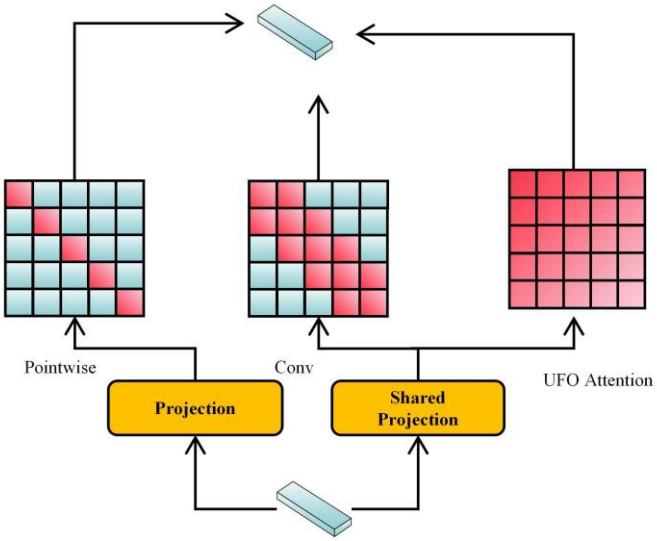


Fig. 4. Attention mechanism of MSA-UFO Attention proposed in this paper.

As shown in Fig. 4, the MSA-UFO Attention mechanism consists of three key components with various granularities: the UFO attention for capturing global features, a depthwise separable convolutional layer for extracting local features, and a location feedforward network to capture token-specific features. The function and implementation of each component are described in detail below.

In the self-attention mechanism, the input sequences are linearly transformed to generate the Q, K, and V matrices, and the attention weights are computed using the softmax function. Due to the high dimensionality of the input matrices, the overall computational complexity increases significantly when processing long sequences in self-attention, particularly in exponential and normalization operations. As a result, the total computational complexity of the self-attention mechanism is $O(N^2 d)$, which greatly limits its efficiency and applicability in long-sequence processing. To address this issue, we utilizes the UFO attention with XNorm mechanism [32] replacing the traditional softmax operation. XNorm not only reduces computational complexity but also enhances computational efficiency, achieving significantly improved performance in long-sequence tasks. As shown in Fig. 5, this process is termed Unit Force Operated Attention.

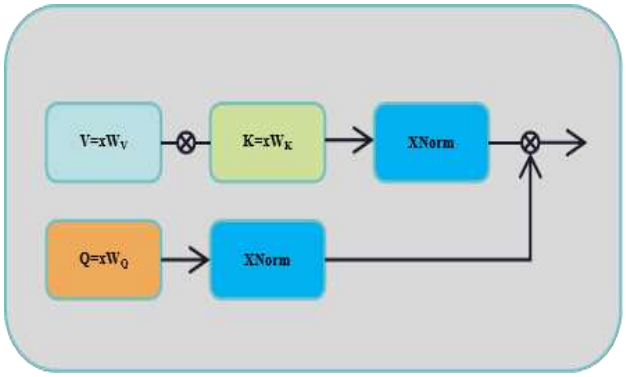


Fig. 5. Schematic flow of the UFO Attention mechanism.

$$XN(a) = \frac{\gamma a}{\sqrt{\sum_{i=0}^{h} \square a \square^2}} \tag{10}$$

$$A(x) = XN_{\dim=filter}(Q)(XN_{\dim=space}(K^T V)) \tag{11}$$

where γ is the learnable parameter and h is the embedding dimension. As shown in (10), the XNorm module is essentially similar to the L2 normalization operation, but its application is extended to two dimensions: the spatial dimension $K^T V$ and the channel dimension Q, and thus the process is also known as Cross-Normalization [32]. By normalizing the input sequence into unit vectors, XNorm ensures consistent feature vector length, effectively avoiding numerical instability. Additionally, XNorm significantly reduces computational complexity, making the self-attention mechanism more efficient for long-sequence processing. The total computational complexity of the UFO Attention mechanism is presented in (12):

$$A(x) = XN_{\dim=filter}(Q)(XN_{\dim=space}(K^T V)) \tag{12}$$

The UFO Attention mechanism significantly improves computational efficiency by replacing the traditional softmax operation with the XNorm module, which reduces the computational complexity from $O(N^2 d)$ to $O(Nd^2)$ when $N > d$. This optimization is especially critical for long-sequence processing. By normalizing feature vectors to unit length, the XNorm module effectively avoids numerical instability resulting from fluctuations in input data amplitude. It also leverages the associative property of matrix multiplication to optimize the calculation order and simplify the overall computation process. This approach not only significantly improves computational efficiency but also enhances the stability and robustness of the model in long-sequence tasks.

Deep convolution operation: It has been shown that deep convolution performs exceptionally well in capturing local contextual information [30]. By decomposing standard convolution into deep convolution and point-by-point convolution, it significantly reduces parameters and computational complexity, especially in long-sequence tasks. It also compensates for the self-attention mechanism's limitations in capturing local information. Therefore, this paper employs multi-scale deep convolution with varying kernel sizes to better capture local contextual information in sequences.

$$Conv_k(X) = Depth_conv_k(XW^V)W^{out} \tag{13}$$

where X represents the input feature matrix, W^V is the linearly varying feature matrix, and $Depth_conv_k(\cdot)$ represents the deep convolution operation applied across each channel's convolution kernel. W^{out} is the weight matrix for point-by-point convolution. Additionally, we introduce a shared projection mechanism [31] and a dynamic convolution kernel selection mechanism [30] to ensure consistent feature representation across stages, while adaptively adjusting each convolution unit’s output through learned

weights, enhancing the model's flexibility and expressiveness.

Pointwise Feed-Forward Network: To learn token-level representations, we combine a UFO Attention network with a pointwise feed-forward network in each layer of the model. The UFO Attention network is responsible for capturing global dependencies in the sequence, while the pointwise feed-forward network applies the same linear transformation at each token's position to extract fine-grained feature representations, as shown in (14) below

$$Po\,\mathrm{int}\,wise(X) = \max(0, XW_1 + b_1)W_2 + b_2 \tag{14}$$

where W_1 and W_2 are the weight matrices, and b_1 and b_2 represent the bias vectors. This structural design allows the model to capture both global and local features, thereby enhancing the model's ability to represent complex data.

Integration network: As shown in (15), the MSA-UFO Attention mechanism proposed in this paper effectively combines the UFO Attention mechanism for capturing global features, the depthwise separable convolutional layer to extract local features, and the positional feed-forward network for capturing token features through multi-scale representation learning.

$$X_i = X_{i-1} + A(x) + Conv(X_{i-1}) + Po\,\mathrm{int}\,wise(X_{i-1}) \tag{15}$$

where X_{i-1} represents the output of the previous layer, $A(x)$ represents the UFO Attention, $Conv(\cdot)$ represents the deep convolutional operation, and $Po\,\mathrm{int}\,wise(\cdot)$ represents the positional feedforward network.

As shown in Fig. 2**,** in our proposed two-stage attentional multimodal fusion mechanism, the model first learns the audio feature F_{Leader} augmented by the leader network and the audio-text-visual multimodal feature F_{Follow} obtained through processing by the follower network. Next, these two features are concatenated to form a unified multimodal joint feature representation. This joint feature representation is then fed into a fully connected layer and a multilayer perceptron comprising two layers for further processing. Finally, after Softmax computation, the probability distribution across emotion categories is obtained, and the emotion label with the highest probability is selected as the final prediction for the i -th discourse $\hat{y}_i$.

$$F_{mutil} = F_{Leader} \oplus F_{Follow} \tag{16}$$

$$Z_i = W^Z F_{Mutil} + b_Z \tag{17}$$

$$l_i = \max(0, W^l z_i + b_l) \tag{18}$$

$$p_i = Soft\max(Wl_i + b) \tag{19}$$

$$\hat{y}_i = \arg\max(p_i) \tag{20}$$

3.6 Confident Learning

Confidence learning is an emerging principled framework [38] for identifying labeling errors, characterizing labeling noise,

and applying it to learning with noisy data. As a branch of weakly supervised learning and learning with noise, confidence learning identifies and addresses labeling errors by estimating the joint distribution between noisy and true labels, instead of simply correcting noisy labels or adjusting loss weights. Confidence learning depends solely on the model's predictive probabilities and artificial labels to detect labeling noise. The Fig. 6 provides a schematic and theoretical derivation of the algorithmic flow in confidence learning. For the detailed mathematical derivation, refer to Literature[38].

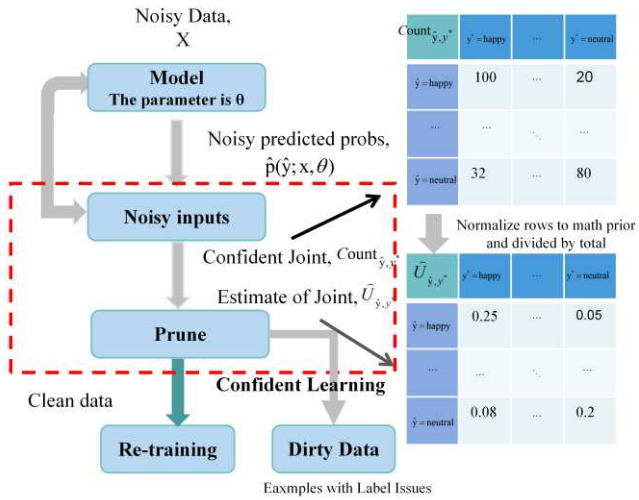


Fig. 6. Example flow diagram of the confidence learning algorithm.

First, we define the samples X containing the noisy labels.

$$X := (x, \tilde{y})^n \in (\mathbb{R}^d, [m]) \tag{21}$$

where x represents the input samples, $\tilde{y}$ represents the noisy labels, m represents the number of classification categories, n represents the total number of samples, and d represents the feature dimensions. The model with parameter θ then calculates the set of predicted probabilities $\hat{p}(\tilde{y}; x, \theta)$ through K-fold cross-validation. As shown in (22) and (23), the core of confidence learning is the construction of the count matrix $Count_{\tilde{y},y^*}$ of the mapping relationship between noise labels and real labels and the estimation of the joint distribution $\hat{U}_{\tilde{y},y^*}$ f statistical noisy labels with the true labels.

$$Count_{\tilde{y},y^*}[i][j] := |\hat{X}_{\tilde{y}=i,y^*=j}| \text{ Where, } \hat{X}_{\tilde{y}=i,y^*=j} := \left\{ x \in X_{\tilde{y}=1} : \tilde{p}(\tilde{y}=j; x, \theta) \geq t_j, j = \underset{l \in [m]; \tilde{p}(\tilde{y}; x, \theta) \geq t_j}{\operatorname{argmax}} \tilde{p}(\tilde{y}=l; x, \theta) \right\} \tag{22}$$

$$\hat{U}_{\tilde{y}=i,y^*=j} = \frac{\dfrac{Count_{\tilde{y}=i,y^*=j}}{\sum_{j \in [m]} Count_{\tilde{y}=i,y^*=j}} \cdot |X_{\tilde{y}=i}|}{\sum_{i \in [m], j \in [m]} \left(\dfrac{Count_{\tilde{y}=i,y^*=j}}{\sum_{j \in [m]} Count_{\tilde{y}=i,y^*=j}} \cdot |X_{\tilde{y}=i}| \right)} \tag{23}$$

Where t_j represents the average predicted probability for all samples in a given category. The formula for the joint distribution

$\hat{U}_{\tilde{y},y^*}$ given a matrix $Count_{\tilde{y},y^*}$ is shown in (23). The marginal distribution of the matrix is made consistent with the sample observations by normalizing each row of $Count_{\tilde{y},y^*}$ so that the sum of each row is 1. The number of samples in each category is adjusted using $\left|X_{\tilde{y}=i}\right|$ to make the marginal distribution consistent with the sample observations. With the six filtering schemes proposed in the literature [38], erroneous samples in noisy labels can be efficiently identified, with adjustable thresholds to control the number of problematic labels returned. Although confidence learning performs well in dealing with label noise, its single-strategy approach may be limited in complex ERC tasks. For this reason, we propose a co-optimization approach that combines confidence learning with additional strategies to improve robustness and generalization in ERC tasks. Specifically, we incorporate a transfer learning strategy to retrain on the corrected dataset, leveraging existing knowledge and higher-quality data to further enhance the model's generalization performance.

As shown in Fig. 7, the overall process of model denoising includes: first, training the DSA-MERC model proposed in this paper on the original dataset to obtain the best test results and the corresponding pre-training model; then, cleaning the dataset using the confidence learning strategy; and finally, applying a pre-trained model to the cleaned dataset to achieve optimal performance.

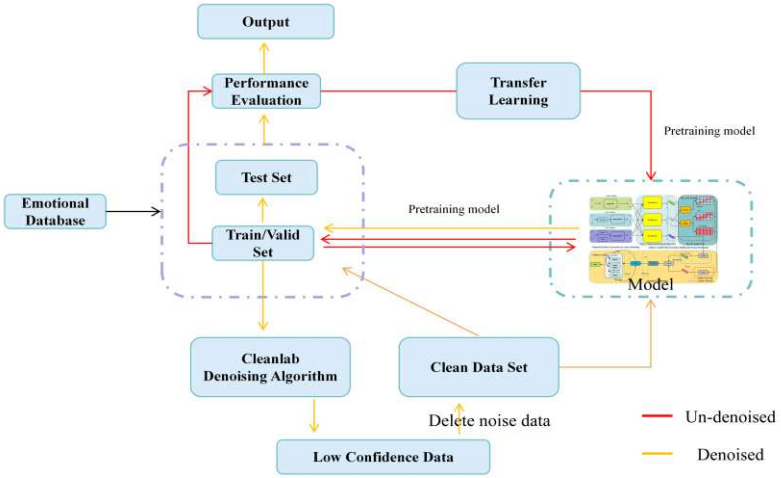


Fig. 7. Overall schematic of model denoising.

3.7 Loss Function Design

As shown in Fig. 2, we introduce multiple loss functions, combining them synergistically to jointly improve the overall performance of the model. Specifically, we use Soft-HGR Loss, SWFC Loss, and Cross-Entropy Loss.

Soft-HGR Loss: To further promote the multimodal fusion effect of the follower network, we introduce the Soft Hirschfeld-Gebelein-Rényi (Soft-HGR) Loss [42]. This loss function aims to enhance the correlation between multimodal features. Specifically, Soft-HGR Loss promotes effective feature fusion by maximizing the mutual information between representations of

different modalities.

$$Loss_{Soft-HGR} = -\sum_{F_A \neq F_B, F_A, F_B \in F} (\mathrm{E}\left[F_A^{\ T} F_B\right] - \frac{1}{2} tr\left(\mathrm{cov}(F_A)\mathrm{cov}(F_B)\right)) \tag{24}$$

Here, we define $F = \{f_t, f_a, f_v\}$, where f_t, f_a, and f_v denote the feature matrices of text, audio, and vision in the follower network, respectively. F_A and F_B are the feature matrices from two different modalities within the multimodal feature set F.

SWFC Loss: We introduce the SWFC Loss proposed in the Shi et al. [29] to further improve the model's ability to distinguish between emotion categories that are semantically similar or less frequent. The SWFC Loss aims to maximize interclass distance and minimize intraclass distance by incorporating the sample weight term $(\frac{N}{n_{y_{i,j}}})^{\alpha}$ and the focusing parameter γ, effectively improving the accuracy of emotion classification. The final mathematical model of SWFC Loss is given in the following equation:

$$s_{j,g}^{(i)} = \frac{\exp(\frac{z_{i,j}^T z_{i,g}}{\tau})}{\sum_{z_{i,s} \in A_{i,j}} \exp(\frac{z_{i,j}^T z_{i,s}}{\tau})} \tag{25}$$

$$Loss_{SWFC} = -\sum_{i=1}^{M}\sum_{j=1}^{C^{(i)}} (\frac{N}{n_{y_{i,j}}})^{\alpha} \frac{1}{|\{z_{i,g} \in A_{i,j} \mid y_{i,j} = y_{i,j}\}|} \sum_{z_{i,g} \in R_{i,j}} (1 - s_{j,g}^{(i)})^{\gamma} \log s_{j,g}^{(i)} \tag{26}$$

Where $(\frac{N}{n_{y_{i,j}}})^{\alpha}$ and α are the sample parameters that control the sample weights and sampling. $s_{j,g}^{(i)}$ is the similarity score between the current sample and the other samples in the batch, and τ is the temperature parameter that controls the penalty strength for negative samples. $z_{i,j}$ is the feature representation of the *j* statement in the dialog i, $A_{i,j}$ is the feature representation of all features in the current batch except $z_{i,j}$, and $R_{i,j}$ represents the set of positive sample pairs.

It's worth noting that cross-entropy loss is employed as the primary loss function for the emotion classification task. The Total Loss function combines Soft-HGR Loss, SWFC Loss, and Cross-Entropy Loss through linear weighting, as shown in the following equation. Where $\lambda_{1,2} \in (0,1)$ represents the weight parameters for each loss function.

$$Loss_{Total} = \lambda_1 Loss_{Soft-HGR} + \lambda_2 Loss_{SWFC} + (1 - \lambda_1 - \lambda_2) Loss_{CE} \tag{27}$$

4 Experimental Settings

4.1 Datasets and Evaluations

We selected two widely used multimodal ERC datasets, IEMOCAP [34] and MELD [43], to evaluate our proposed model. As

shown in Table 1, the emotion distributions of the two datasets are provided.

IEMOCAP: The IEMOCAP dataset contains two-way conversations involving 10 speakers and consists of 153 dialogues and 7,433 discourses. Each discourse is labeled as one of six emotions: happy, sad, neutral, angry, excited, and frustrated.

MELD: The MELD dataset is derived from the TV series *Friends* and includes 1,433 multi-speaker dialogues and 13,708 discourses. Each discourse is labeled as one of seven emotions: neutral, surprise, fear, sadness, joy, disgust, and anger.

Evaluation Indicators: Following previous studies [19][20][23][24][29], we selected weighted average F1 scores to assess overall model performance. Additionally, we report weighted F1 scores for each emotion category to enable more detailed analysis.

4.2 Implementation Details

Model training and inference were conducted on a server equipped with an Intel(R) Core(TM) i5-13600K processor and an RTX 4090 graphics card. The models were implemented in PyTorch and used the Adam optimizer. For the IEMOCAP dataset, the batch size was set to 32, the learning rate to 0.0001, and the L2 regularization parameter to 0.00001. In the MELD dataset, the batch size was set to 250, the learning rate to 0.001, and the weight decay for L2 regularization to 0.0001. Additionally, the 0.1 Dropout rate was applied to mitigate overfitting. For loss parameter settings, the weights of SWFC Loss, HGR Loss, and CE Loss were set to 0.2, 0.4, and 0.4 in both datasets. In the MultiAttn model settings, six MultiAttn modules were stacked in the IEMOCAP dataset, while four were used in the MELD dataset. The temperature parameter, sample weight parameter, and focus parameter in SWFC Loss were set to 0.8, 0.8, and 2, respectively, in IEMOCAP, and to 1.4, 0.8, and 2, respectively, in MELD.

Table 1

The distribution of emotions in the IEMOCAP and MELD datasets

Datasets	Emotions								
	Anger	Happiness/Joy	Sadness	Neutral	Excitement	Frustration	Fear	Surprise	Disgust
IEMOCAP	1103	648	1084	1708	1041	1849	-	107	-
MELD	1607	2308	1002	6436	-	-	358	1636	361

4.3 Baseline model

To better distinguish between methods, we divide the baseline comparison into text-based ERC methods and multimodal ERC methods.

Text-based ERC methods: BC-LSTM(2017)[14]: This is one of the first approaches to explore ERC using a deep RNN-based model. The model does not rely on inter-speaker relationships but captures the contextual relationships in the text through a bidirectional LSTM.

CMN(2018)[44]: The model is the first to explore the interaction between speakers in a dialog. The historical discourse of

speakers is modeled through gated recurrent units (GRUs), and dependencies between speakers are captured through memory fusion based on attentional mechanisms.

DialogueRNN(2019)[16]: This is one of the most commonly used and powerful ERC benchmarking models available. The model utilizes multiple GRU modules to track speaker states, emotional states and global contexts in a conversation to accurately capture emotional context information for more accurate emotion recognition.

DialogueGCN(2019)[21]: This is a graph convolutional neural network-based ERC method that solves the problem of contextual information propagation in traditional RNN-based methods by utilizing self-dependence and inter-speaker dependencies to model conversational context.

IterativeERC(2020)[45]: It was firstly proposed to use iteratively predicted sentiment labels instead of real sentiment labels to model emotional interactions in conversations, which effectively solves the problem of misleading sentiment information interference in the context and demonstrates excellent performance.

Multimodal ERC methods: MM-DFN(2022)[25]: The model designs a graph-based dynamic multimodal fusion module from a multimodal perspective. The module reduces inter-modal redundant information and enhances inter-modal complementarity by capturing contextual dynamics in different semantic spaces, thus providing a comprehensive understanding of multimodal dialog contexts and improving sentiment recognition performance.

CTNet(2021)[28]: CTNet is a multimodal learning framework designed for conversational sentiment recognition, modeling intramodal and cross-modal interactions through Transformer structures, and combining bi-directional GRUs with multiple heads of attention and speaker embeddings to capture context and speaker dependencies.

EmotionIC (2023)[20]: The most important feature of EmotionIC is the integration of attention and recurrent neural structures, modeling global identity perception and local temporal context dependency through IM-MHA and DiaGRU, respectively, and explicitly exploiting emotional inertia and contagion effects in dialogues through SkipCRF, thus enhancing the ability of multilevel emotion modeling.

GA2MIF(2023)[23]: An emotion detection session based on graph and attention fusion of two-stage multi-source information. It is also one of the most advanced multimodal ERC models.

BIGMF(2024)[24]: A two-stream graph learning multimodal fusion approach to effectively mitigate the problem of information conflict and heterogeneity in multimodal fusion by modeling intra-modal and inter-modal information separately.

CBERL(2024)[49] improves multimodal emotion recognition in conversations by enhancing minority-class boundary representations to alleviate class imbalance.

HyFusER(2025)[50] Proposes a hybrid multimodal Transformer that enhances text–speech interaction through dual cross-modal attention for multimodal emotion recognition.

5 Results and discussion

In this section, we analyze the performance of the DSA-MERC model proposed in this paper through detailed comparison with existing work and extensive ablation experiments. First, we analyze the importance of each modality and compare task performance across different modality combinations. Second, we explore the denoising effects of confidence learning strategies, comparing different filtering methods and thresholds. Finally, we analyze the effect of the attention mechanism on model performance, visualizing the attention and convolutional weights in the proposed MSA-UFO Attention mechanism, with specific cases to support the analysis.

5.1 Comparison with existing baseline works

To verify the effectiveness of our proposed method, we compared DSA-MERC with current state-of-the-art baseline methods on the IEMOCAP and MELD datasets, as shown in Tables 2 and 3.

Overall performance comparison: As shown in Tables 2 and 3, the proposed method achieves the best overall performance on both benchmark datasets. On the IEMOCAP dataset, our model obtains a weighted F1 score of 71.38%, outperforming all competing methods. Compared with the weakest baseline, CMN, the proposed method improves the weighted F1 score by 23.20%, while exceeding the previous best-performing model, G2AMIF, by 1.38%. On the MELD dataset, our model achieves a weighted F1 score of 65.99%, slightly surpassing HyFusER (65.76%), MM-DFN (65.48%), EmotionIC (65.36%), and other representative baselines. Compared with the earliest baseline, BC-LSTM, our method improves the weighted F1 score by 10.09 %. These results demonstrate that the proposed framework consistently delivers superior overall performance across both datasets, highlighting its effectiveness and robustness for multimodal emotion recognition in conversations.

Comparison on Performance for Each Class: In Tables 2 and 3, the performance of various benchmark models on the IEMOCAP and MELD datasets across different emotion categories is presented, with bold font indicating the best performance. The experimental results show that although DSA-MERC does not consistently achieve the best performance in every emotion category, it maintains competitive performance across different emotions and obtains the highest overall weighted F1 scores on both datasets. This further validates the robustness and effectiveness of the proposed model. In the IEMOCAP dataset, the multimodal fusion mechanism enables the model to achieve the best performance in the "Happy" and "Neutral" categories, with W-F1 scores of 56.30% and 71.72%, respectively. Although the performance in the "Sad" category (83.40%) is slightly lower than that of MM-DFN (86.07%), and the performance in the "Frustrated" category (70.77%) is only marginally lower than that of CBERL (71.19%), the proposed model consistently achieves competitive results across all emotion categories. These results demonstrate that the proposed multimodal fusion strategy effectively enhances the model's robustness and generalization ability.

In the MELD dataset, the proposed model achieves the best performance in the "Surprise", "Joy", and "Disgust" categories, with W-F1 scores of 58.18%, 65.91%, and 29.87%, respectively, while maintaining competitive performance in the remaining emotion categories. Although several benchmark methods achieve slightly better results on individual emotion classes, the proposed model attains the highest overall weighted F1 score, demonstrating its superior capability in balancing recognition performance across different emotion categories and its effectiveness in complex multimodal emotion recognition tasks.

Table 2

Performance comparison of existing baseline works on the IEMOCAP dataset

Models	IEMOCAP(%)						
	Happy	Sad	Neutral	Anger	Excited	Frustrated	W-F1
BC-LSTM(2017)[14]	34.43	60.87	51.81	56.73	57.95	58.92	54.95
CMN(2018)[44]	29.86	53.83	40.14	52.44	50.09	55.75	48.18
Dialoguernn(2019)[16]	35.54	69.85	55.30	61.85	62.21	59.38	58.76
DialogueGCN(2019)[21]	42.75	83.54	63.54	64.19	63.08	66.99	64.18
IterativeERC(2020)[45]	/	/	/	/	/	/	64.37
MM-DFN(2022)[25]	52.87	**86.07**	71.66	65.04	75.26	62.19	69.87
CTNet(2021)[28]	51.30	79.90	65.80	67.20	78.70	58.80	67.50
EmotionIC(2023)[20]	/	/	/	/	/	/	69.61
BIGMF(2024)[24]	48.56	82.45	66.15	70.39	**77.40**	65.76	69.48
G2AMIF(2023)[23]	46.15	84.50	68.38	70.29	75.99	66.49	70.00
CBERL(2024)[49]	**67.34**	72.84	60.75	**73.51**	70.77	**71.19**	69.27
HyFusER(2025)[50]	/	/	/	/	/	/	68.81
Our	56.30	83.40	**71.72**	65.36	73.84	70.77	**71.38**

Table 3

Performance comparison of existing baseline works on the MELD dataset

Models	MELD(%)							
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	W-F1
BC-LSTM(2017)[14]	73.80	47.70	5.40	25.10	51.30	5.20	38.40	55.90
CMN(2018)[44]	/	/	/	/	/	/	/	/
Dialoguernn(2019)[16]	76.56	47.64	0.00	24.65	51.49	0.00	46.01	57.95
DialogueGCN(2019)[21]	76.02	46.37	0.98	24.32	53.62	1.22	43.03	57.52
IterativeERC(2020)[45]	77.52	53.65	3.31	23.62	56.63	19.38	48.88	60.72
MM-DFN(2022)[25]	79.65	58.17	26.67	35.71	64.89	24.76	52.15	65.48
CTNet(2021)[28]	76.70	52.70	10.00	32.50	56.00	11.20	46.80	64.99
EmotionIC(2023)[20]	/	/	/	/	/	/	/	65.36
BIGMF(2024)[24]	/	/	/	/	/	/	/	58.95
G2AMIF(2023)[23]	76.92	49.08	/	27.18	51.87	/	48.52	58.94
CBERL (2024)[49]	**80.81**	57.91	22.23	**41.36**	65.67	24.65	**55.31**	65.27
HyFusER(2025)[50]	/	/	/	/	/	/	/	65.76
Our	78.95	**58.18**	**25.60**	40.64	**65.91**	**29.87**	53.49	**65.99**

In summary, the experimental results confirm that the proposed model demonstrates superior performance and significant effectiveness in the ERC task. In addition, we observed several consistent and noteworthy phenomena across both datasets:

(1) Superiority of Multimodal Fusion Models. Multimodal fusion models consistently outperform traditional baselines that

rely solely on contextual or speaker-level modeling, indicating that cross-modal information interaction significantly enhances the accuracy of emotion recognition.

(2) Poor Performance on Minority Emotion Classes by Baselines. Experimental results reveal that most baseline models perform poorly on minority emotion categories. For example, the classification performance for emotions such as Fear, Sadness, and Joy typically ranges between 0.00% and 26%, suggesting that many existing methods fail to adequately capture the emotional expressions of minority groups and lack sufficient sensitivity to true emotional labels.

(3) Limited Global Performance Gains Despite Improvements in Minority Classes. Interestingly, although our model significantly outperforms state-of-the-art methods in recognizing minority emotion categories, the overall improvement in weighted F1-score remains relatively modest. As shown in Table 1, the MELD dataset contains a disproportionately large number of Neutral utterances (e.g., 6436 instances), while categories such as Fear and Disgust contain only around 300 samples. This extreme class imbalance likely limits the impact of minority class improvements on the overall W-F1 metric. Nevertheless, the detailed improvements in recognizing minority emotions highlight the model's potential for real-world applications, particularly in scenarios involving complex emotional dynamics and nuanced affective understanding.

5.2 Ablation experiment: the Effectiveness of the Modalities

To examine the importance of each modality, we conducted a multimodal fusion ablation experiment to compare the performance of unimodal and multimodal approaches, as shown in Table 4. Due to the inherent imbalance of the experimental dataset, the weighted average F1 score remains our primary evaluation metric. As shown in Table 4, in unimodal mode, the text and audio modalities perform well, significantly outperforming the video modality, which is consistent with most related studies. Although the text modality demonstrates the highest performance, proving its key role in conversational emotion recognition, the audio modality also exhibits notable advantages. This suggests that the role of audio information in capturing emotional cues is also critical. Audio not only supplements emotional cues that may be absent in text, but also provides a more direct expression of emotion through changes in intonation and speech rate, which plays an important supportive role, especially in complex conversations.

The analysis in Table 4 also shows that bimodal and trimodal fusion methods perform significantly better than the unimodal approach, with the trimodal fusion method proposed in this paper performing best. In the IEMOCAP dataset, trimodal fusion shows an improvement of 12.37%-20.60% over the unimodal approach, while in the MELD dataset, the improvement range is 4.29%-23.59%. This result indicates that multimodal fusion effectively integrates information from multiple modalities, significantly enhancing the accuracy of emotion recognition. In addition, bimodal experiments that incorporate the audio modality also show significant improvements, further confirming that the ERC task relies not only on textual information but

also on non-verbal signals in audio. Through multimodal fusion, the model captures richer and more nuanced emotional cues, leading to significant improvements in emotion recognition performance.

Table 4

Performance comparison of multimodal ablation experiments

Modality	IEMOCAP	MELD
Text(T)	59.01	61.70
Audio(A)	58.33	60.35
Visual(V)	50.78	42.43
A+T	69.15	65.01
A+V	67.73	63.28
Our（A+T+V）	**71.38**	**65.99**

5.3 Ablation experiment: The Effectiveness of Confident Learning

To verify the effectiveness of the confidence learning strategy introduced in this paper, we designed a series of experiments comparing the model's performance with (CL=1) and without (CL=0) the confidence learning strategy. As shown in Table 5, when using the confidence learning strategy, the model exhibits greater robustness in handling data with higher label noise, significantly improving the accuracy of emotion recognition. This result further validates the effectiveness of the confidence learning strategy in enhancing the model's ability to adapt to noisy labels.

Table 5

Comparison of performance of ablation experiments with confidence learning strategies

CL=0/1	IEMOCAP	MELD
CL=0	69.81	65.73
CL=1	**71.38**	**65.99**

Additionally, we compare the effectiveness of different noise filtering methods on the IEMOCAP and MELD datasets, as shown in Table 6. Specifically, we evaluated six noise filtering methods proposed in the literature [38](see [38] for details). The experimental results show that the labels in the IEMOCAP dataset are relatively accurate, while label noise is more pronounced in the MELD dataset. This difference is clearly visible in Table VI through the F1 score and the number of identified noisy labels.

Table 6

Comparison of performance across different denoising filtering methods in the IEMOCAP and MELD dataset

Method	IEMOCAP		MELD	
	F1(%)	Noise labels number	F1(%)	Noise labels number
Pruning by Noise Rate	**71.38**	302	**65.99**	1399
Pruning by Class Noise Probability	70.00	240	65.44	1459
Joint Noise Rate and Class-Based Pruning	70.23	90	65.77	854
Confident Learning	70.96	122	65.56	664
Pruning by Low Normalized Margin	70.75	122	65.82	664
Pruning by Low Self-Confidence	70.78	122	65.64	664

In the IEMOCAP dataset, the Pruning by Noise Rate method achieves the highest F1 score of 71.38%, while Pruning by Class

Noise Probability is the least effective, with an F1 score of only 70.00%. This indicates that relying on class confidence alone is insufficient for effective noise label filtering. In the MELD dataset, Pruning by Noise Rate also achieves a high F1 score of 65.99%, whereas Pruning by Class Noise Probability remains the least effective, with an F1 score of 65.44%.

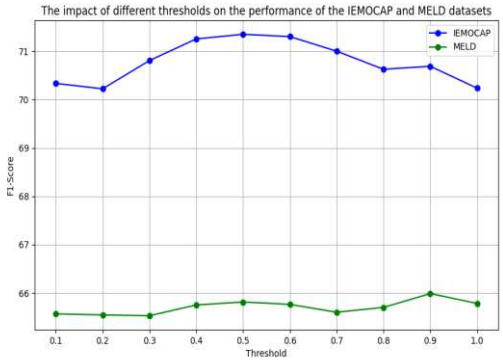


Fig. 8. The impact of different thresholds on performance in the IEMOCAP and MELD datasets.

Additionally, we analyze the impact of Pruning by Noise Rate under different thresholds, as shown in Table 7 and Fig. 8. The experimental results reveal that DSA-MERC's performance initially increases and then decreases as the threshold rises. In the IEMOCAP dataset, the F1 score reaches its peak of 71.38% at a threshold of 0.5; in the MELD dataset, the highest F1 score of 65.99% is achieved at a threshold of 0.9. This further indicates the importance of selecting an appropriate noise filtering strategy and threshold when addressing label noise.

Table 7
The impact of different thresholds on experimental performance in the IEMOCAP and MELD datasets

Thresholds	IEMOCAP(%)	MELD(%)
0.1	70.33	65.56
0.2	70.22	65.54
0.3	70.80	65.53
0.4	71.25	65.75
0.5	**71.38**	65.81
0.6	71.30	65.76
0.7	70.99	65.56
0.8	70.62	65.70
0.9	70.69	**65.99**
1.0	70.23	65.78

5.4 Ablation experiment: The Effectiveness of Audio-LFAN

To further validate the effectiveness of our proposed Audio-LFANin multimodal fusion mechanisms, we conducted several experiments. Specifically, we replaced some current state-of-the-art sequence-based attention mechanisms in the second stage of multimodal fusion, replacing the originally used MAS-UFO Attention to assess whether these substitutions affect the overall performance of the model. The results of the relevant experiments are shown in Table 8.

Among them, SA [9] is a universally applied and powerful attention mechanism. ESA [47] is an external attention mechanism based on an external memory module, with linear complexity and inter-sample correlations, and it also constructs an efficient full

MLP architecture by introducing a multi-head mechanism. AFT [48] significantly improves the computational efficiency of the Transformer model by eliminating dot-product self-attention and introducing efficient localization and spatial weight-sharing mechanisms, thus reducing the memory complexity of the Transformer model while maintaining competitive performance. MUSA Attention [31] combines self-attention and deep convolutional techniques to parallelize the processing of time-series data at different scales, successfully capturing both global dependencies and local contextual information, which substantially improves the model's performance in processing time-series tasks.

The experimental results demonstrate that, compared to using only the first-stage attention mechanism (Bi-MultiAttn), the full model (which includes both the first- and second-stage attention mechanisms) significantly improves performance on the IEMOCAP and MELD datasets, with F1 score increases of 1.05% and 0.62%, respectively. In addition, the state-of-the-art sequence-based attention mechanism is applied in the second stage to extract additional relevant information from the audio, with each model showing varying degrees of performance improvement. These results indicate that MAS-UFO Attention plays a crucial role in capturing both local contextual information and global dependencies in audio during multimodal information fusion. The introduction of MAS-UFO Attention in the second stage effectively enhances multimodal information fusion, further improving the model's overall performance.

Table 8
The impact of different thresholds on experimental performance in the IEMOCAP and MELD datasets

Method	IEMOCAP(%)	MELD(%)
Model+SA	70.55	65.57
Model+ESA	70.79	65.56
Model+AFT	70.38	65.66
Model+MUSA	70.81	65.78
Our (first stage only)	70.33	65.37
Our	**71.38**	**65.99**

5.5 Ablation experiment: The Effectiveness of MAS-UFO Attention

This section provides an in-depth analysis of the effectiveness of the MAS-UFO Attention proposed in this paper. Specifically, we visualize the attention weights and convolutional weights on public conversation data from the IEMOCAP and MELD datasets to demonstrate the performance of MAS-UFO Attention in practical applications. In addition, we introduce a min-max normalization strategy to map the attention weights to the range of [0,1], ensuring that the attention or convolutional weight values at different time steps are on a uniform scale, which facilitates intuitive comparison and analysis.

Visual analysis of attention weight : As illustrated in **Fig. 9**, the MAS-UFO Attention mechanism exhibits a relatively uniform distribution of attention weights, primarily concentrated in the green-to-yellow regions. This distribution demonstrates the model's ability to allocate attention stably and equitably across the input sequences during processing. This is further

supported by the analysis of the MELD dataset, where MAS-UFO Attention exhibits higher feature extraction stability and a relatively even distribution of attention weights.

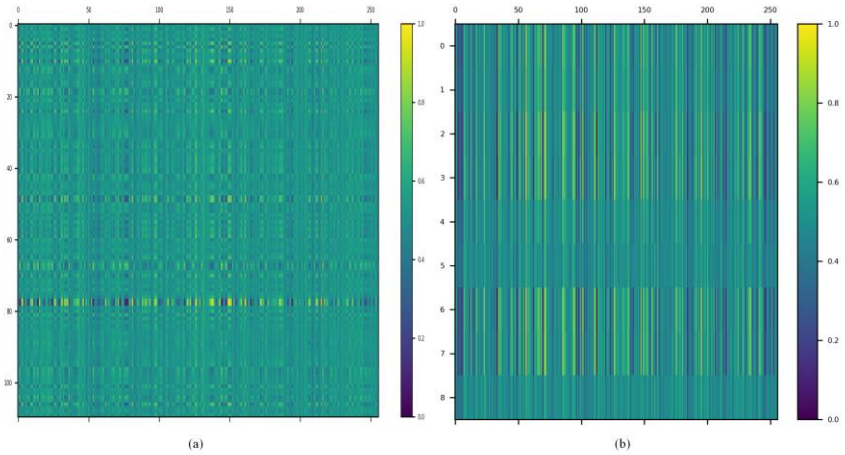


Fig. 9. Attention weight visualization for a specific sequence case in the IEMOCAP dataset (a) and the MELD dataset (b). The case from the IEMOCAP dataset is taken from Ses03M_impro08b, while the case from the MELD dataset is taken from Dialogue_ID_3. The horizontal axis represents the time step number of the audio feature, and the vertical axis represents the time step number of the conversation.

In addition, we further analyzed the changes in attention weights at each time step and used a conversation in the MELD dataset (Dialogue_ID_3) as an example for in-depth discussion. The specific dialog is shown in Table 9. By analyzing the attention weights of each time step sequentially, we find that the model allocates significantly more attention to time steps 0, 2, 3, and 7. The following is a specific analysis of time step 0, including the corresponding emotional expression, changes in attention weights, and the associated audio analysis for this time step.

Taking time step 0 as an example, the textual content “Oh my God, oh my God! Poor Monica!” expresses the character Phoebe's intense surprise. The actor's performance in the raw data further validates this emotion, demonstrating noticeable emotional fluctuations through agitated intonation and an exaggerated tone. As shown in Fig. 10(a), based on the changes in fundamental frequency and RMS energy, it can be inferred that the speaker's emotional fluctuations are more pronounced between approximately 1 and 3 seconds. During this period, there are significant fluctuations in fundamental frequency and energy, indicating that the speaker is experiencing heightened emotion. The disappearance of the rhythm node after 1.8 seconds indicates a change in the speaker's speech rate or rhythm. As shown in Fig. 10(b), the model assigns varying attentional weights at time step 0, which partially reflect the emotional fluctuations.

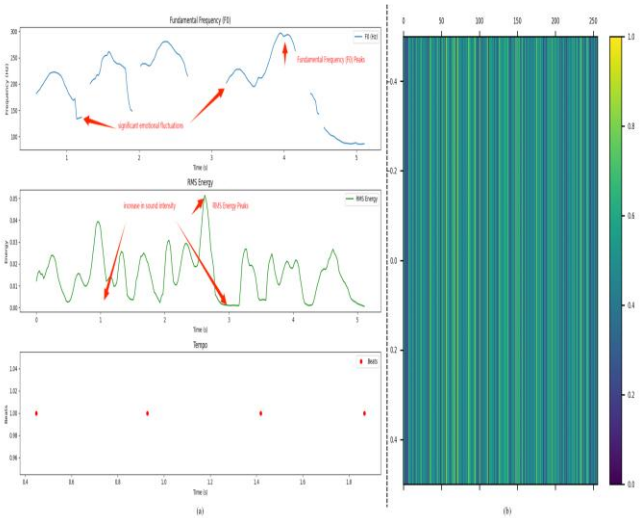


Fig. 10. Audio analysis and weight visualization for specific time steps. (a) shows the visualization of the audio information corresponding to time step 0, and (b) presents the MAS-UFO attentional weight results of the model at time step 0.

In addition, by observing changes in attention weights at multiple time steps in this example, MAS-UFO Attention shows a strong ability to adjust attention at key points in emotion transitions (e.g., from surprise to sadness and then to neutrality). Especially in the transition from sadness to neutrality, changes in attention weights reflect the model's sensitivity to the emotional shift, as shown in Fig. 11. MAS-UFO Attention dynamically adapts to emotional changes and improves the stability and accuracy of feature extraction by focusing on key features.

The excellent performance of MAS-UFO Attention is closely related to its incorporation of the XNorm mechanism. By normalizing the input features, XNorm ensures that features at different time steps maintain the same scale when calculating attention, achieving a balanced allocation of attention. In addition, the associative property of matrix multiplication further enhances the model's flexibility, enabling it to demonstrate higher adaptability and accuracy in complex emotion recognition.

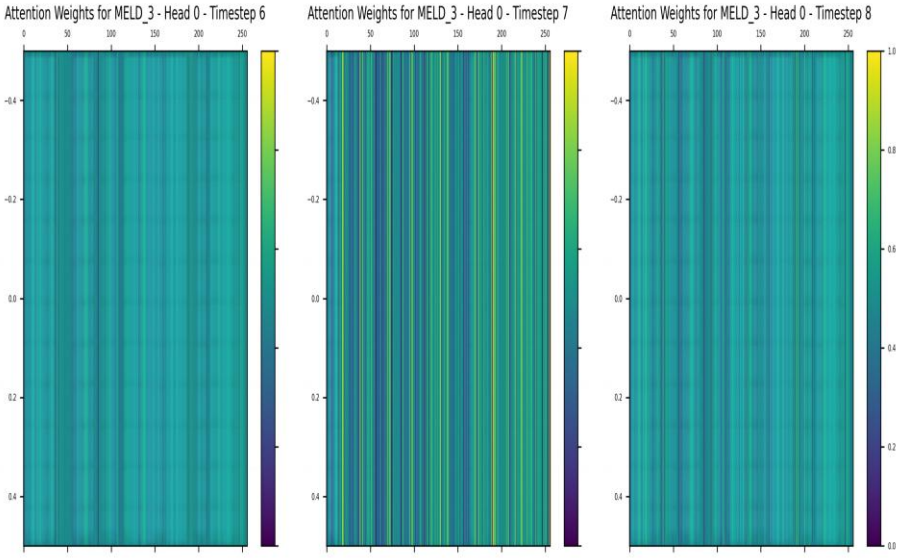


Fig. 11. Attention weight visualization results for MAS-UFO Attention at different time steps. Where, from left to right, are the weight changes at time steps 6, 7 and 8, respectively.

Table 9
Case Study: A Dialogue Example in MELD

Time step	Utterance	Speaker	Emotion
0	Oh my God, oh my God! Poor Monica!	Phoebe	Surprise
1	What, what, what?!	Chandler	Surprise
2	What?!	Phoebe	Surprise
3	He was with her when he wrote this poem.	Phoebe	Neutral
4	Look, 'My vessel so empty with nothing inside.	Phoebe	Neutral
5	Now that I've touched you, you seem emptier still.'	Phoebe	Neutral
6	He thinks Monica is empty, she is the empty vase!	Phoebe	Surprise
7	Oh, totally. Oh, God, oh, she seemed so happy too.	Phoebe	Sadness
8	Done.	Joey	Neutral

Visual analysis of convolution weight： To capture and represent multimodal features more efficiently, we employed convolution kernels of sizes 1, 3, and 5 to capture fine- and coarse-grained feature information at different scales (see Figures 12 and 13). On the IEMOCAP dataset, the weight distribution of the convolution kernels for MAS-UFO Attention is relatively uniform, primarily centered between [-0.008, 0.008]. This pattern indicates stable performance in both local and global feature extraction. This may result from the introduction of the XNorm mechanism, which reduces the risk of gradient vanishing or explosion by normalizing the output, thereby enhancing model stability. A similar trend is observed in the MELD dataset, where the convolutional weights of MAS-UFO Attention remain within a narrow range between [-0.06, 0.06] and are evenly distributed.

Table 10
Quantitative analysis of the convolutional weights of different methods using mean, standard deviation, L1/L2 paradigm, and sparsity quantifiers on the IEMOCAP and MELD datasets

IEMOCAP Dataset						
Index	MAS-UFO Attention			MAS-UFO Attention (Softmax)		
	k=1	k=3	k=5	k=1	k=3	k=5
Mean	0.00087	0.00013	-0.00020	-0.00019	0.01559	-0.00344
Standard deviation	0.02172	0.00179	0.00193	0.036540	0.21023	0.15843
L1 Norm	494.48733	40.67604	44.09956	888.45416	4773.15722	3571.71411
L2 Norm	3.64910	0.30142	0.32716	6.13194	35.37622	26.59284
Sparsity	0.0	0.0	0.0	0.0	0.0	0.0
MELD Dataset						
	k=1	k=3	k=5	k=1	k=3	k=5
Mean	0.0005	0.00018	-4.62625	0.00140	-0.02917	0.009250
Standard deviation	0.02156	0.00196	0.00205	0.03614	0.19332	0.15463
L1 Norm	97.96360	8.69626	9.42160	169.80473	872.39904	699.37945
L2 Norm	1.61858	0.14799	0.15458	2.71437	14.67255	11.62565
Sparsity	0.0	0.0	0.0	0.0	0.0	0.0

As shown in Table 10, we systematically evaluated the weights of the convolutional kernels using quantitative metrics (e.g., mean, standard deviation, L1 norm, L2 norm, and sparsity). In addition, to better compare model performance, we replaced XNorm with Softmax for comparison in our model and analyzed the results in detail. MAS-UFO Attention demonstrates higher stability and effectiveness, with a more concentrated distribution of convolutional kernel weights across the two datasets,

especially for the small-scale convolutional kernel (k=1). In contrast, MAS-UFO Attention (Softmax) exhibits a larger mean and standard deviation, particularly for larger-scale convolutional kernels (k=3 and k=5), with a more dispersed weight distribution, indicating instability in global feature extraction. Additionally, the L1 and L2 norms of MUSE Attention are significantly higher than those of MAS-UFO Attention, especially in the IEMOCAP dataset, indicating that its weights fluctuate more and that its feature extraction process is less stable.

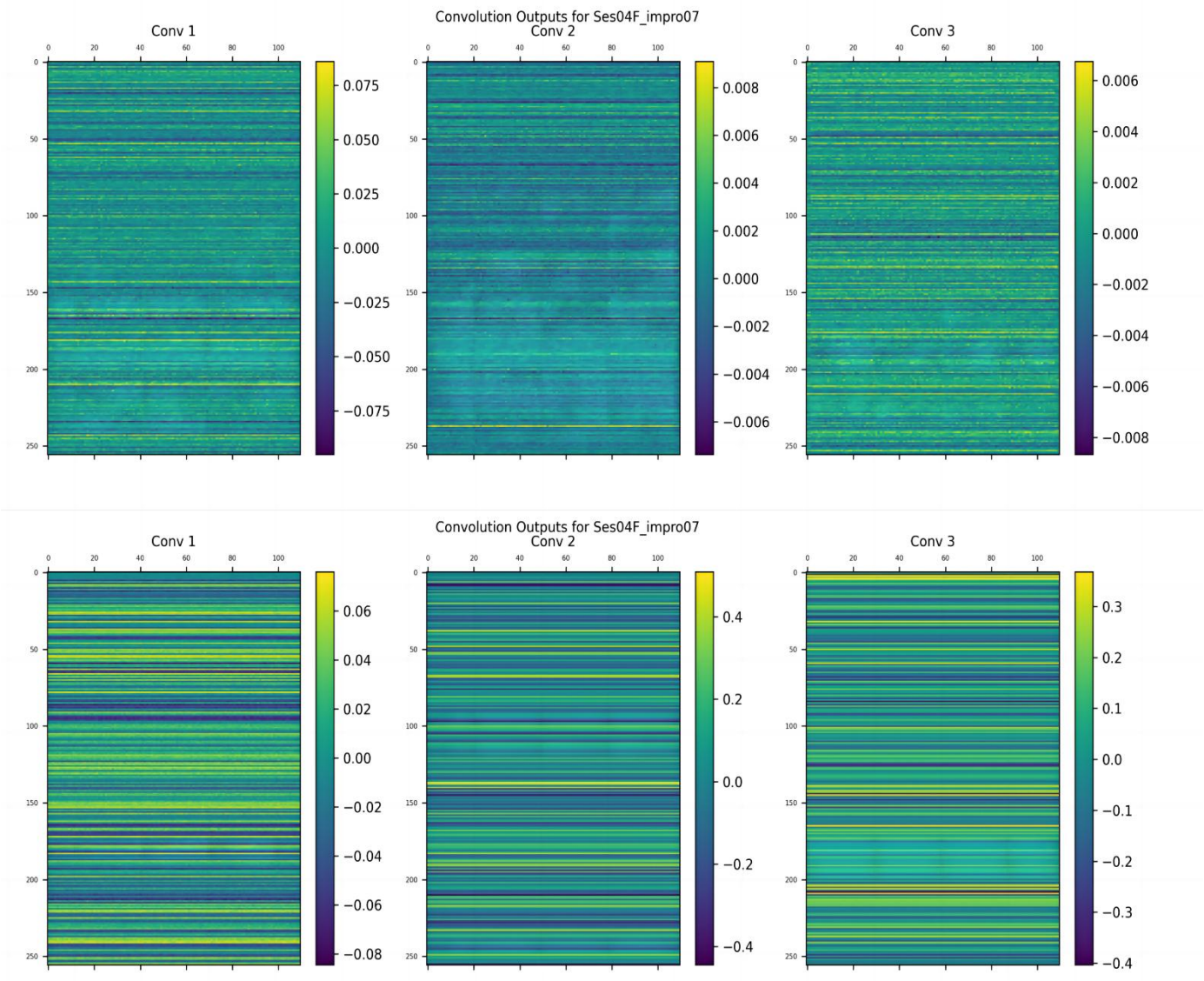


Fig. 12. The visualization results of MAS-UFO Attention (top) and MAS-UFO Attention with Softmax (bottom) on the IEMOCAP dataset for convolutional kernel sizes K = 1, 3, and 5.

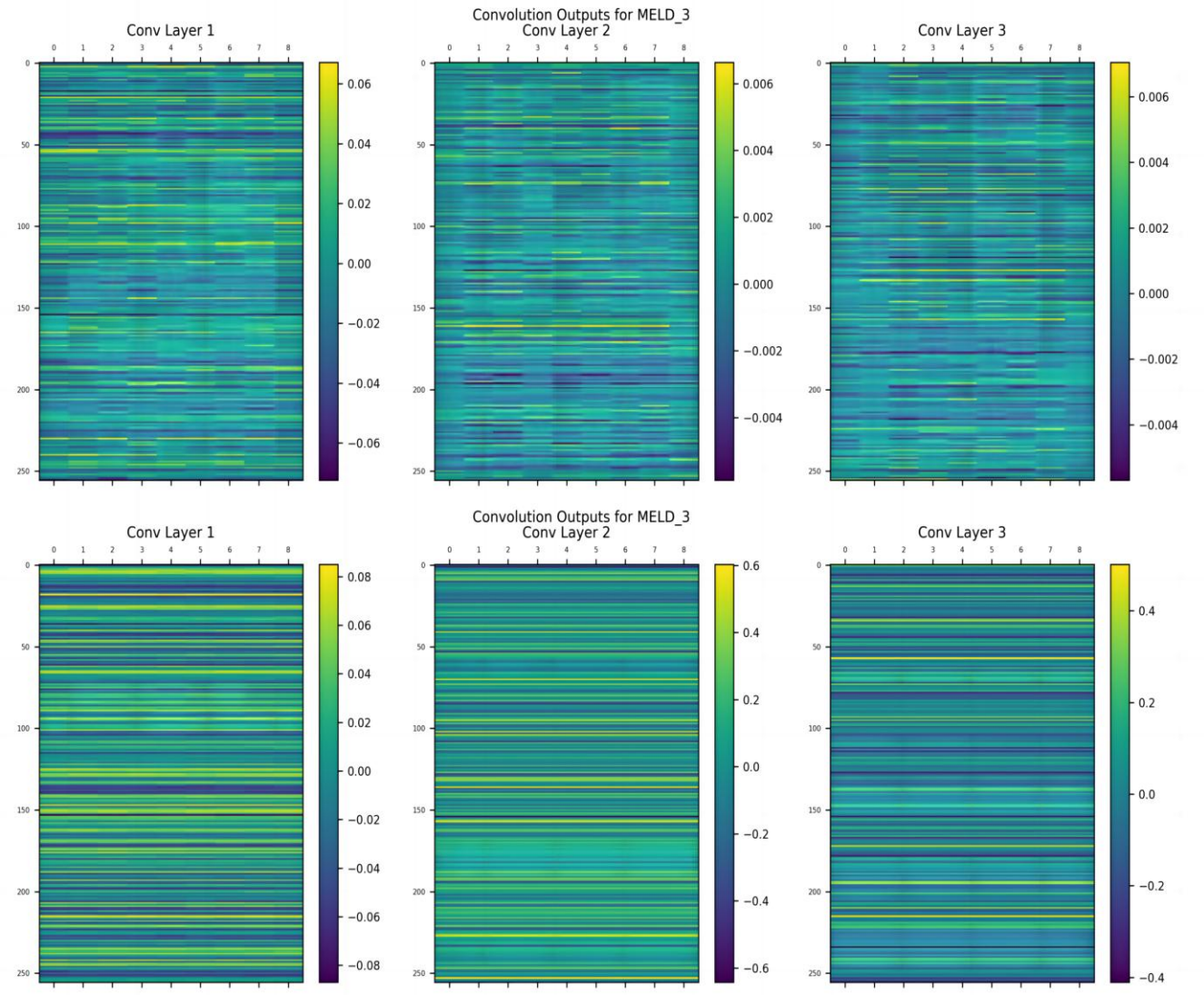


Fig. 13. The visualization results of MAS-UFO Attention (top) and MAS-UFO Attention with Softmax (bottom) on the MELD dataset for convolutional kernel sizes K = 1, 3, and 5.

5.6 Ablation Experiments: The Effectiveness of Loss

We compare the effect of using Confident Learning versus not using it on the change in loss within 100 epochs during the validation phase of the model on the IEMOCAP and MELD datasets. As shown in Figures 14, the model without Confident Learning (CL = 0) performs well in the early stages of training, with a rapidly decreasing loss curve, but flattens out in the later stages, exhibiting slight fluctuations. This suggests a gradual weakening of the learning effect. In contrast, the model using Confident Learning (CL = 1) shows a smoother and continuously decreasing loss curve, particularly for Total Loss and HGR Loss, indicating that Confident Learning effectively filters label noise and enhances model training stability. However, SWFC loss and CE loss still display some fluctuations in the later stages, which could be attributed to several factors. First, as model complexity increases—especially after the introduction of the MAS-UFO Attention mechanism—the model may struggle to fit

complex patterns, leading to instability. Furthermore, Confident Learning alters the data distribution by filtering noisy labels. Despite inheriting the previous training weights, the new data distribution may cause incomplete adaptation of the model in the later training stages, triggering fluctuations in the loss curves. Overall, Confident Learning significantly enhances the learning ability and stability of the model in the early and middle stages. However, in the later stages, overfitting and fluctuations in the loss curves may occur due to the increased model complexity and changes in the data distribution.

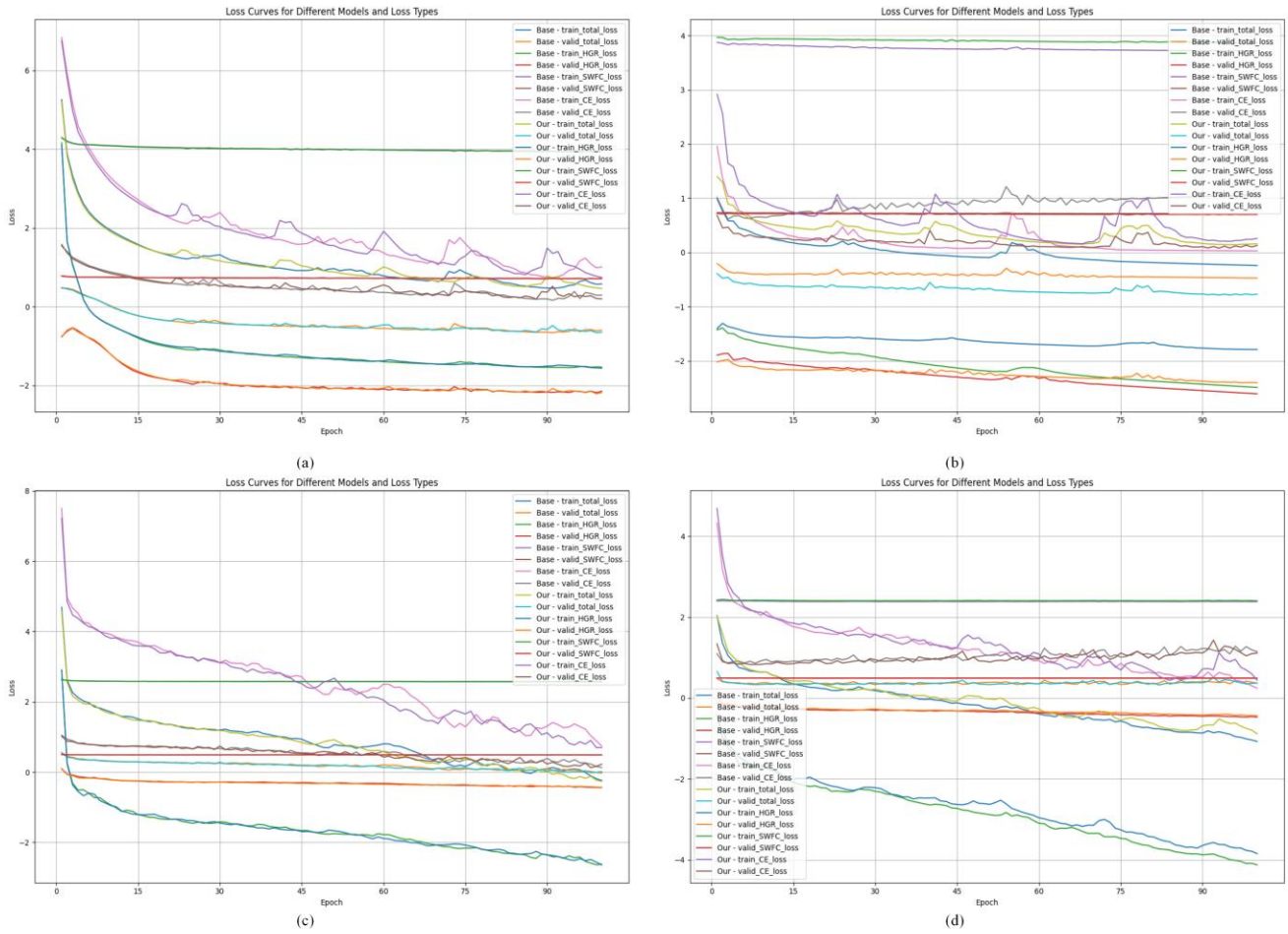


Fig. 14. It shows a comparison of the loss curves of different models on the IEMOCAP and MELD datasets. In the figure, CL=0/1 indicates whether the Confident Learning strategy is used, "Our" represents our proposed method, and "Base" refers to the base model, which uses only the first stage. And the (a) shows the change in loss curves for CL=0 on the IEMOCAP dataset, while the (b) shows the change in loss curves for CL=1 on the IEMOCAP dataset. The (c) shows the change in loss curves for CL=0 on the MELD dataset, and the (d) shows the change in loss curves for CL=1 on the MELD dataset.

6 Conclusion

In this paper, we propose a Reliable Multimodal Conversational Emotion Recognition Model Based on a Dual-Stage Attention Mechanism, addressing key challenges in ERC, such as insufficient multimodal integration, the complexity of processing long sequential data, and the issue of label noise.

In future work, we plan to explore the application of the model to other related tasks, such as multimodal emotion analysis and

emotion computation in real-time human-computer interaction. Additionally, we will investigate the model's potential in handling more complex and diverse multimodal datasets, particularly in low-resource environments where label noise and data scarcity pose significant challenges.

References

[1]Tao, J., & Tan, T. (2005). Affective computing: A review. In J. Tao, T. Tan, & R. W. Picard (Eds.), Affective Computing and Intelligent Interaction (pp. 981–995). Springer. https://doi.org/10.1007/11573548_125

[2]Wang, M., Ma, H., Wang, Y., & Sun, X. (2024). Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion. Applied Acoustics, 218, 109886. https://doi.org/10.1016/j.apacoust.2024.109886

[3]Wang, F., Ding, Z., Xia, R., Li, Z., & Yu, J. (2023). Multimodal emotion-cause pair extraction in conversations. IEEE Transactions on Affective Computing, 14(3), 1832–1844. https://doi.org/10.1109/TAFFC.2022.3226559

[4]Li, C., Xie, L., Shao, X., Pan, H., & Wang, Z. (2024). A multimodal shared network with a cross-modal distribution constraint for continuous emotion recognition. Engineering Applications of Artificial Intelligence, 133, 108413. https://doi.org/10.1016/j.engappai.2024.108413

[5]Li, C., Xie, L., & Pan, H. (2022). Branch-fusion-net for multi-modal continuous dimensional emotion recognition. IEEE Signal Processing Letters, 29, 942–946. https://doi.org/10.1109/LSP.2022.3160373

[6]Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In M. Palmer, R. Hwa, & S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1103–1114). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1115

[7]Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L.-P. (2018). Memory fusion network for multi-view sequential learning. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), Article 1. https://doi.org/10.1609/aaai.v32i1.12021

[8]Wu, C.-H., Lin, J.-C., & Wei, W.-L. (2014). Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. APSIPA Transactions on Signal and Information Processing, 3, e12. https://doi.org/10.1017/ATSIP.2014.11

[9]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[10]Tang, G., Müller, M., Rios, A., & Sennrich, R. (2018). Why self-attention? A targeted evaluation of neural machine translation architectures. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4263–4272). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1458

[11]Lukasik, M., Bhojanapalli, S., Menon, A. K., & Kumar, S. (2020). Does label smoothing mitigate label noise? Proceedings of the 37th International Conference on Machine Learning, 119, 6448–6458.

[12]Yang, Z., Li, X., Cheng, Y., Zhang, T., & Wang, X. (2024). Emotion recognition in conversation based on a dynamic complementary graph convolutional network. IEEE Transactions on Affective Computing, 15(3), 1567–1579. https://doi.org/10.1109/TAFFC.2024.3360979

[13]Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Access, 7, 100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050

[14]Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In R. Barzilay & M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers) (pp. 873–883). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1081

[15]Hsu, C.-C., Chen, S.-Y., Kuo, C.-C., Huang, T.-H., & Ku, L.-W. (2018). EmotionLines: An emotion corpus of multi-party conversations. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA). https://aclanthology.org/L18-1252/

[16]Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019, July). Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 6818-6825). Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). DialogueRNN: An attentive RNN for emotion detection in conversations. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), Article 01. https://doi.org/10.1609/aaai.v33i01.33016818

[17]Jiao, W., Yang, H., King, I., & Lyu, M. R. (2019). HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers) (pp. 397–406). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1037

[18]Zhu, L., Pergola, G., Gui, L., Zhou, D., & He, Y. (2021). Topic-driven and knowledge-aware transformer for dialogue emotion detection. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (volume 1: Long Papers) (pp. 1571–1582). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.125

[19]Li, Z., Tang, F., Zhao, M., & Zhu, Y. (2022). EmoCaps: Emotion capsule based model for conversational emotion recognition. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022 (pp. 1610–1618). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.126

[20]Liu, Y., Li, J., Wang, X., & Zeng, Z. (2024). EmotionIC: Emotional inertia and contagion-driven dependency modeling for emotion recognition in conversation. Science China Information Sciences, 67(8), 182103. https://doi.org/10.1007/s11432-023-3908-6

[21]Ghosal, D., Majumder, N., Poria, S., Chhaya, N., & Gelbukh, A. (2019). DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 154–164). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1015

[22]Ishiwatari, T., Yasuda, Y., Miyazaki, T., & Goto, J. (2020). Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7360–7370). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.597

[23]Li, J., Wang, X., Lv, G., & Zeng, Z. (2024). GA2MIF: Graph and attention based two-stage multi-source information fusion for conversational emotion detection. IEEE Transactions on Affective Computing, 15(1), 130–143. https://doi.org/10.1109/TAFFC.2023.3261279

[24]Lu, N., Han, Z., Han, M., & Qian, J. (2024). Bi-stream graph learning based multimodal fusion for emotion recognition in conversation. Information Fusion, 106, 102272. https://doi.org/10.1016/j.inffus.2024.102272

[25]Hu, D., Hou, X., Wei, L., Jiang, L., & Mo, Y. (2022). MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7037–7041. https://doi.org/10.1109/ICASSP43922.2022.9747397

[26]Zou, S., Huang, X., Shen, X., & Liu, H. (2022). Improving multimodal fusion with main modal transformer for emotion recognition in conversation. Knowledge-Based Systems, 258, 109978. https://doi.org/10.1016/j.knosys.2022.109978

[27]Ma, H., Wang, J., Lin, H., Zhang, B., Zhang, Y., & Xu, B. (2024). A transformer-based model with self-distillation for multimodal emotion recognition in conversations. IEEE Transactions on Multimedia, 26, 776–788. https://doi.org/10.1109/TMM.2023.3271019

[28]Lian, Z., Liu, B., & Tao, J. (2021). CTNet: Conversational transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 985–1000. https://doi.org/10.1109/TASLP.2021.3049898

[29]Shi, T., & Huang, S.-L. (2023). MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers) (pp. 14752–14766). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.824

[30]Zhang, Q., & Yang, Y.-B. (2021). ResT: An efficient transformer for visual recognition. Advances in Neural Information Processing Systems, 34, 15475–15485.

[31]Zhao, G., Sun, X., Xu, J., Zhang, Z., & Luo, L. (2019). MUSE: Parallel multi-scale attention for sequence to sequence learning (No. arXiv:1911.09483). arXiv. https://doi.org/10.48550/arXiv.1911.09483

[32]Song, J. (2021). UFO-ViT: High performance linear vision transformer without softmax (No. arXiv:2109.14382). arXiv. https://doi.org/10.48550/arXiv.2109.14382

[33]Li, W., Fan, L., Shao, S., & Song, A. (2024). Generalized contrastive partial label learning for cross-subject EEG-based emotion recognition. IEEE Transactions on Instrumentation and Measurement, 73, 1–11. https://doi.org/10.1109/TIM.2024.3398103

[34]Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[35]Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., & Brox, T. (2019). SELF: Learning to filter noisy labels with self-ensembling (No. arXiv:1910.01842). arXiv. https://doi.org/10.48550/arXiv.1910.01842

[36]Yi, K., & Wu, J. (2019). Probabilistic end-to-end noise correction for learning with noisy labels. 7017–7025.

[37]Zhang, Z., & Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 8792–8802.

[38]Northcutt, C., Jiang, L., Chuang, I., 2021. Confident learning: estimating uncertainty in dataset labels. J. Artif. Intell. Res. 70, 1373–1411. https://doi.org/10.1613/jair.1.12125

[39]Kim, T., & Vossen, P. (2021). EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa (No. arXiv:2108.12009). arXiv. https://doi.org/10.48550/arXiv.2108.12009

[40]Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia, 1459–1462. https://doi.org/10.1145/1873951.1874246

[41]Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

[42]Wang, L., Wu, J., Huang, S.-L., Zheng, L., Xu, X., Zhang, L., & Huang, J. (2019). An efficient approach to informative feature extraction from multimodal data. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), Article 01. https://doi.org/10.1609/aaai.v33i01.33015281

[43]Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations (No. arXiv:1810.02508). arXiv. https://doi.org/10.48550/arXiv.1810.02508

[44]Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.-P., & Zimmermann, R. (2018). Conversational memory network for emotion recognition in dyadic dialogue videos. Proceedings of the Conference. Association for Computational Linguistics. North American Chapter. Meeting, 2018, 2122–2132. https://doi.org/10.18653/v1/n18-1193

[45]Lu, X., Zhao, Y., Wu, Y., Tian, Y., Chen, H., & Qin, B. (2020). An iterative emotion interaction network for emotion recognition in conversations. In D. Scott, N. Bel, & C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics (pp. 4078–4088). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.360

[46]Jieying, X., Minh, N. P., Blake, M., & Minh Le, N. (2023). Accumulating word representations in multi-level context integration for ERC task. 2023 15th International Conference on Knowledge and Systems Engineering (KSE), 1–6. https://doi.org/10.1109/KSE59128.2023.10299463

[47]Guo, M.-H., Liu, Z.-N., Mu, T.-J., & Hu, S.-M. (2023). Beyond self-attention: External attention using two linear layers for visual tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5), 5436–5447. https://doi.org/10.1109/TPAMI.2022.3211006

[48]Zhai, S., Talbott, W., Srivastava, N., Huang, C., Goh, H., Zhang, R., & Susskind, J. (2021). An attention free transformer (No. arXiv:2105.14103). arXiv. https://doi.org/10.48550/arXiv.2105.14103

[49]Meng, T., Shou, Y., Ai, W., Yin, N., & Li, K. (2024). Deep Imbalanced Learning for Multimodal Emotion Recognition in Conversations. IEEE Transactions on Artificial Intelligence, 5(12), 6472–6487. https://doi.org/10.1109/TAI.2024.3445325

[50]Yi, M.-H., Kwak, K.-C., & Shin, J.-H. (2025). HyFusER: Hybrid multimodal transformer for emotion recognition using dual cross modal attention. Applied Sciences, 15(3), 1053. https://doi.org/10.3390/app15031053

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DeclarationofInterestStatement.docx
- Fundinginformation.docx