Revisiting Multi-modal Emotion Learning with Broad State Space Models and Probability-guidance Fusion

Yuntao Shou[1], Tao Meng[1] (✉), Wei Ai[1], and Keqin Li[2]

[1]College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, Hunan 410004, China
[2]Department of Computer Science, State University of New York, New Paltz, New York 12561, USA
shouyuntao@stu.xjtu.edu.cn, {mengtao, aiwei}@hnu.edu.cn, and lik@newpaltz.edu

Abstract. Multi-modal Emotion Recognition in Conversation (MERC) has received considerable attention in various fields, e.g., human-computer interaction and recommendation systems. Most existing works perform feature disentanglement and fusion to extract emotional contextual information from multi-modal features. After revisiting the characteristic of MERC, we argue that long-range contextual semantic information should be extracted in the feature disentanglement stage and the inter-modal semantic information consistency should be maximized in the feature fusion stage. Inspired by recent State Space Models (SSMs), Mamba can efficiently model long-distance dependencies. Therefore, in this work, we fully consider the above insights to further improve the performance of MERC. Specifically, on the one hand, in the feature disentanglement stage, we propose a Broad Mamba, which does not rely on a self-attention mechanism for sequence modeling, but uses state space models to compress emotional representation, and utilizes broad learning systems to explore the potential data distribution in broad space. Different from previous SSMs, we design a bidirectional SSM convolution to extract global context information. On the other hand, we design a multi-modal fusion strategy based on probability guidance to maximize the consistency of information between modalities. Experimental results show that the proposed method can overcome the computational and memory limitations of Transformer when modeling long-distance contexts, and has great potential to become a next-generation general architecture.

Keywords: Multi-modal Emotion Recognition · State Space Models · Multi-modal Fusion.

1 Introduction

Emotion recognition in conversation [43,36,38,37] has received considerable research attention and has been widely used in various fields, e.g., emotion analysis [14] and public opinion warning [44], etc. Recently, research on Multi-modal Emotion Recognition in Conversation (MERC) has mainly focused on multimodality, i.e., text, video and audio [25,3]. As shown in Fig. 1, MERC aims to identify emotion labels in sentences with text, video, and audio information. Unlike previous work [17] that only uses text information for emotion recognition, MERC improves the model's emotion understanding

capabilities by introducing audio and video information. The introduction of audio and video alleviates the limitation of insufficient semantic information caused by relying solely on text features.

Many existing works [24,40,35] improve the performance of MERC by effectively extracting contextual semantic information of different modalities and fusing inter-modal complementary semantic information. By revisiting the characteristics of MERC, we argue that the core idea of MERC includes a feature disentanglement step and a feature fusion step.

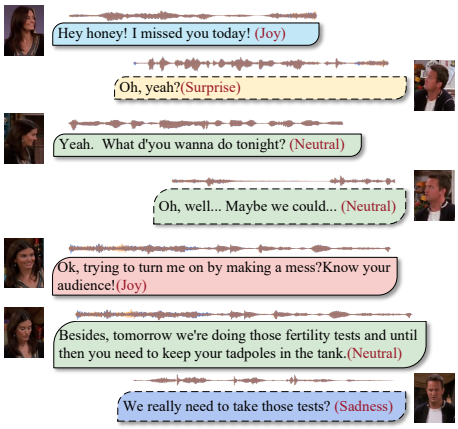

Fig. 1: An illustrative example of multi-modal emotion recognition in conversation. For each given sentence, it contains three modal information about the speaker, i.e., text, video and audio. The task of MERC is to identify the emotional labels contained in the three modal information.

Specifically, the goal of feature disentanglement is to extract the contextual semantic information most relevant to emotional features in multi-modal features [42]. Recent work on Transformers [21] has achieved great success in modeling long-range contextual semantic information. Compared with traditional Recurrent Neural Networks (RNNs) [27,20], the advantage of Transformer is that it can effectively provide global contextual semantic information through the attention mechanism in parallel. However, the quadratic complexity of the self-attention mechanism in Transformers poses challenges in terms of speed and memory when dealing with long-range context dependencies. Inspired by the state space models, Mamba with linear complexity is proposed to achieve efficient training and inference. Mamba's excellent scaling performance shows that it is a promising Transformer alternative for context modeling. Therefore, to efficiently extract long-distance contextual semantic information, we designed the broad Mamba, which incorporates the SSMs for data-dependent global emotional context modeling, and a broad learning system to explore the potential data distribution in the broad space. Different from previous SSMs, we design a bidirectional SSM convolution to extract global context information. In addition, we also introduce position encoding information to improve SSMs' ability to understand sequences at different positions.

After completing feature disentanglement, the model needs to perform feature fusion to maximize the consistency of information between different modalities. The core idea of feature fusion is to assign different weights by determining the importance of different modal features to downstream tasks. Many cross-modal feature fusions have been proposed in existing MERC research, e.g., tensor fusion network [45], graph fusion network [46,34], attention fusion [32]. However, the feature fusion process in previous works is relatively coarse-grained and cannot actually determine the contribution

of each modal feature to downstream tasks. We argue that label information plays an important role in guiding multi-modal information fusion. Therefore, how to properly fuse multi-modality and determine the contribution of multi-modal features to downstream tasks in a fine-grained manner remains a challenge.

To tackle the above problems, we propose an effective probability-guided fusion mechanism to achieve multi-modal contextual feature fusion, which utilizes the predicted label probability of each modal feature as the weight vectors of the modal features. Compared with other feature fusion models for emotion recognition tasks, the proposed fusion method can utilize the predicted label probability information in a fine-grained manner to actually determine the contribution of different modal features to the emotion prediction task.

To evaluate the effectiveness and efficiency of our proposed method, we conduct extensive experiments on two widely used benchmark datasets, IEMOCAP and MELD. In fact, the proposed method achieves state-of-the-art performance with low computational consumption, and experimental results demonstrate its effectiveness and efficiency.

Overall, our main contributions can be summarized as follows:

- We propose a Broad Mamba, which combines a broad learning system for searching abstract emotional features in a broad space and a SSM for data-dependent global emotional context information extraction. Different from previous SSMs, we design a bidirectional SSM convolution to extract global context information.
- We propose an effective probability-guided fusion mechanism to achieve multi-modal contextual feature fusion, which utilizes the predicted label probability of each modal feature as the weight vectors of the modal features.
- We conduct extensive experiments on the IEMOCAP and MELD datasets. Experimental results show that our proposed method achieves superior performance compared with the well-established Transformer or GNN architectures.

2 Related work

2.1 Multi-modal Emotion Recognition in Conversation

In the early eras, GRU [5] and LSTM [13] are the de-facto standard network designs for Natural Language Processing (NLP). Many recurrent neural network architectures [10], [27], [20], [11] have been proposed for various Multi-modal Emotion Recognition in Conversation (MERC). The pioneering work, Transformer changed the landscape by enabling efficient parallel computing under the premise of long sequence modeling. Transformer treats text as a series of 1D sequence data and applies an attention architecture to achieve sequence modeling. Transformer's surprising results on long sequence modeling and its scalability have encouraged considerable follow-up work for MERC [4], [33], [22]. One line of works focus on achieving intra-modal and inter-modal information fusion. For example, CTNet [23] proposes a single Transformer and cross Transformer. CKETF [7] constructs a Context and Knowledge Enriched Transformer. TL-ERC applies the Transformer with the transfer learning. Another pioneering work, Graph Neural Network (GNN) further improved the performance of ERC. The core idea of GNN is to learn the representation of nodes or graphs through the feature information

of nodes and the connection relationships in the graph structure [46]. For instance, DialogueGCN [6] proposes to use context information to build dialogue graphs. DER-GCN [1] fuses event relationships into speaker relationship graphs.

2.2 State Space Models

The State Space Models (SSMS) is used to describe the dynamic change process consisting of observed values and unknown internal state variables. Gu et al. [9] proposes a Structured State Space Sequence (S4) model, an alternative to the Transformer architecture that models long-range dependencies without using attention. The property of linear complexity of state space sequence lengths has received considerable research attention. Smith et al. [39] improves S4 by introducing MIMO SSM and efficient parallel scanning into the S4 layer to achieve parallel initialization and state reset of the hidden layer. He et al. [12] proposes introducing dense connection layers into SSM to improve the feature representation ability of shallow hidden layer states. Mehta et al. [28] improves the memory ability of the hidden layer by introducing gated units on S4. Recently, Gu et al. [8] proposes the general language model Mamba, which has better sequence modeling capabilities than Transformers and is linearly complex. Zhu et al. [47] introduces bidirectional SSM based on Mamba to improve the context information representation of the hidden layer.

3 Preliminary Information

3.1 Multi-modal Feature Extraction

Following previous work [26], we use RoBerta in this paper to obtain context-embedded representations of text. For video and audio features, following previous work [19], we utilize DenseNet and openSMILE for feature extraction and obtain video embedding features $\boldsymbol{\xi}_v$ and audio embedding features $\boldsymbol{\xi}_a$, respectively.

3.2 State Space Model

The State Space Model (SSMs) is an efficient sequence modeling model that can capture the dynamic changes of data over time. A typical SSM consists of a state equation and an observation equation, where the state equation describes the dynamic changes within the system, and the observation equation describes the connection between the system state and observations. Given an input $x(t) \in \mathbb{R}$ and a hidden state $h(t) \in \mathbb{R}$, $y(t)$ is obtained mathematically through a linear ordinary differential equations (ODE) as follows:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), y(t) = \mathbf{C}h(t) \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the evolution parameter and $\mathbf{B} \in \mathbb{R}^{N\times 1}, \mathbf{C} \in \mathbb{R}^{1\times N}$ are the projection parameters, and N is the latent state size.

Inspired by SSM, Mamba discretizes ODEs to achieve computational efficiency. Mamba discretizes the evolution parameter $\mathbf{A}$ and the projection parameter $\mathbf{B}$ by introducing a timescale parameter $\boldsymbol{\Delta}$ to obtain $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$. The formula is defined as follows:

$$\overline{\mathbf{A}} = \exp(\boldsymbol{\Delta}\mathbf{A}), \overline{\mathbf{B}} = (\boldsymbol{\Delta}\mathbf{A})^{-1}(\exp(\boldsymbol{\Delta}\mathbf{A}) - \mathbf{I}) \cdot \boldsymbol{\Delta}\mathbf{B} \tag{2}$$

In practice, we use a first-order Taylor series to obtain an approximation of $\overline{\mathbf{B}}$ as follows:

$$\overline{\mathbf{B}} = (e^{\boldsymbol{\Delta}\mathbf{A}} - \mathbf{I})\mathbf{A}^{-1}\mathbf{B} \approx (\boldsymbol{\Delta}\mathbf{A})(\boldsymbol{\Delta}\mathbf{A})^{-1}\boldsymbol{\Delta}\mathbf{B} = \boldsymbol{\Delta}\mathbf{B} \tag{3}$$

After obtaining the discretized $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, we rewrite Eq. 1 as follows:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, y_t = \mathbf{C}h_t + \mathbf{D}x_t \tag{4}$$

and then the output is computed via global convolutiona as follows:

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{\mathrm{M}}\overline{\mathbf{B}}), \mathbf{y} = \mathbf{x} * \overline{\mathbf{K}} + \mathbf{x} * \mathbf{D} \tag{5}$$

We adopted Mamba as a sequence modeling method in this work since Mamba can efficiently process sequence data without significant performance degradation.

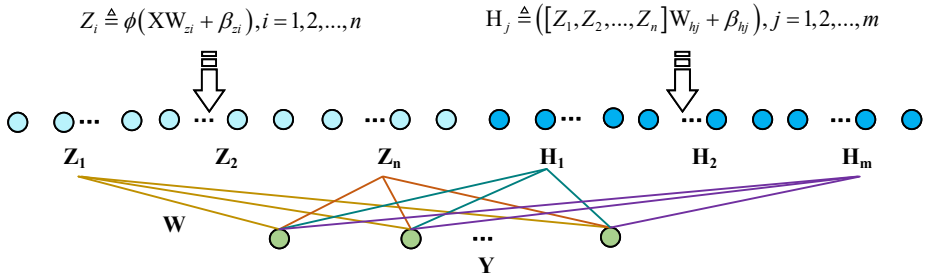

Fig. 2: The overall architecture of Broad Learning System (BLS). $\mathbf{Z}_i$ represents the feature nodes, $\mathbf{H}_i$ represents the enhancement nodes, and $\mathbf{Y}$ represents the predicted labels.

3.3 Broad Learning System

Broad Learning System (BLS) is different from traditional deep learning methods that it mainly focuses on discovering the relationship between features in the input data, rather than extracting features through multi-level nonlinear transformations. The core idea of BLS is to jointly solve the optimization problem by integrating the semantic information of feature nodes and enhancement nodes. The overall process of the BLS algorithm is shown in the Fig. 2.

Specifically, for a given input data $\mathbf{X} \in \mathbb{R}^{N \times M}$, where N represents the number of samples and M represents the dimension of the feature. The generated feature nodes are defined as follows:

$$\mathbf{Z}_i \triangleq \phi(\mathbf{X}\mathbf{W}_{z_i} + \boldsymbol{\beta}_{z_i}), i = 1, 2, \ldots, n \tag{6}$$

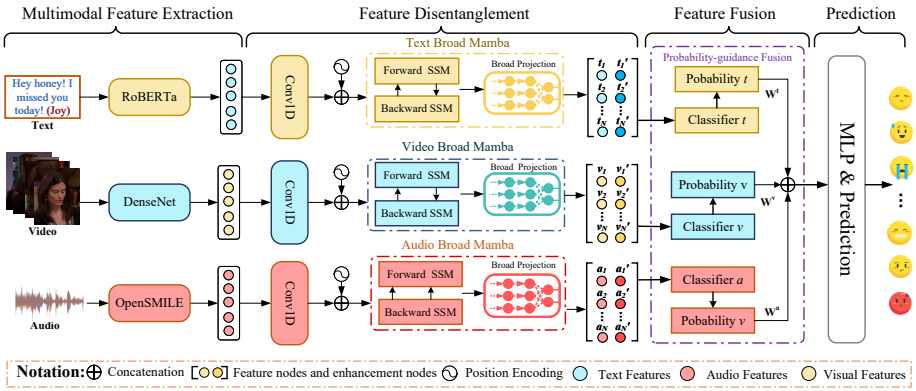

Fig. 3: The overall framework of the proposed model. Specifically, we first input the extracted multi-modal features into a 1-D convolutional layer for multi-scale feature extraction and introduce position encoding information to consider the position information of the series in the context. Then we input the obtained multi-modal features with multi-scale information into Broad Mamba to extract contextual semantic information and explore the potential data distribution in the broad space. Finally, we use a probability-guidance fusion model to complete the fusion of multi-modal features and achieve emotion prediction.

where $\mathbf{W}_{z_i} \in \mathbb{R}^{M \times d_z}$ and $\beta_z \in \mathbb{R}^{1 \times d_z}$ are the learnable parameters. d_z is the embedding dimensions of generated features and ϕ is the activation function. The set of generated feature nodes is represented as $\mathbf{Z}^n \triangleq [\mathbf{Z}_1, \ldots, \mathbf{Z}_n]$, n is the size of the set of generated feature nodes. Similarly, enhancement node features are defined as follows:

$$\mathbf{H}_j \triangleq \phi(\mathbf{Z}\mathbf{W}_{h_j} + \boldsymbol{\beta}_{h_j}), j = 1, 2, \ldots, m \tag{7}$$

where $\mathbf{W}_{h_i} \in \mathbb{R}^{d_z \times d_h}$ and $\beta_z \in \mathbb{R}^{1 \times d_h}$ are the learnable parameters. d_h is the embedding dimensions of enhancement features. The set of enhancement feature nodes is represented as $\mathbf{H}^m \triangleq [\mathbf{H}_1, \ldots, \mathbf{H}_m]$.

The final model output by concatenating feature nodes and enhancement nodes is as follows:

$$\mathbf{Y} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_n | \mathbf{H}_1, \ldots, \mathbf{H}_m]\mathbf{W} = [\mathbf{Z}^n | \mathbf{H}^m]\mathbf{W} \tag{8}$$

where $\mathbf{W}$ are the learnable parameters.

4 The proposed method

4.1 Feature Disentanglement

1D-Conv. To capture features of different scales and abstraction levels in multi-modal features (e.g., information such as the relationship between words and the importance of utterence), we input text features $\boldsymbol{\xi}_t$, video features $\boldsymbol{\xi}_v$ and audio features $\boldsymbol{\xi}_a$ into a 1D convolutional network (Conv1D) as follows:

$$\hat{\boldsymbol{\xi}}_t/\hat{\boldsymbol{\xi}}_a/\hat{\boldsymbol{\xi}}_v = Conv1D_{t/a/v}\left(\boldsymbol{\xi}_t, \boldsymbol{\xi}_a, \boldsymbol{\xi}_v\right) \tag{9}$$

where $\hat{\boldsymbol{\xi}}_t \in \mathbb{R}^{T_t \times d_m}$, $\hat{\boldsymbol{\xi}}_a \in \mathbb{R}^{T_a \times d_m}$, and $\hat{\boldsymbol{\xi}}_v \in \mathbb{R}^{T_v \times d_m}$, T_t, T_a, T_v represent the feature dimensions of text, audio, and video respectively, d_m represents the output feature dimensions.

Furthermore, to facilitate the model to capture the dependencies between long-distance positions in the sequence, we introduce sine and cosine position encoding embedding as follows:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right), PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{10}$$

where pos represents the position in the sequence. i represents the dimension index of position encoding, $i = 0, 1, ..., D-1$. D represents the embedded dimension. We input $\hat{\boldsymbol{\xi}}_t, \hat{\boldsymbol{\xi}}_a, and \hat{\boldsymbol{\xi}}_v$ $(\hat{\boldsymbol{\xi}}_t, \hat{\boldsymbol{\xi}}_a, \hat{\boldsymbol{\xi}}_v = Conv1D_{t/a/v}\left(\boldsymbol{\xi}_t, \boldsymbol{\xi}_a, \boldsymbol{\xi}_v\right) + PE)$ that encodes position information at each time step into Broad Mamba.

Broad Mamba. The overall architecture of the proposed Broad Mamba is shown in Fig. 4. In order to aggregate the contextual semantic information from the forward and backward directions, we build a bidirectional SSM convolution module. Specifically, the first kernel $\overleftarrow{\boldsymbol{\kappa}}$ performs a 1D convolution operator to obtain forward context information. The second kernel $\overrightarrow{\boldsymbol{\kappa}}$ performs a 1D convolution operator to obtain the mutual information associated with emotional information, and we add the two convolved results. The overall operating process is formally defined as follows:

$$\bar{\boldsymbol{\xi}}_j^{t/a/v} = \sum_{l \le j} \overleftarrow{\boldsymbol{\kappa}}_{j-l}^{t/a/v} \odot \hat{\boldsymbol{\xi}}_l^{t/a/v} + \sum_{l \ge j} \overrightarrow{\boldsymbol{\kappa}}_{l-j}^{t/a/v} \odot \hat{\boldsymbol{\xi}}_l^{t/a/v} + \boldsymbol{d}^{t/a/v} \odot \hat{\boldsymbol{\xi}}_j^{t/a/v} = \text{BiSSM}(\hat{\boldsymbol{\xi}}_j^{t/a/v}) \tag{11}$$

where $\overleftarrow{\boldsymbol{\kappa}}$, and $\overrightarrow{\boldsymbol{\kappa}}$ are obtained via Eq. 5.

To explore the potential data distribution of multi-modal data in the broad space and improve the performance of Mamba, we use Broad Learning Sytems (BLS) to enhance the emotional representation ability of features. Specifically, we map the features output by BiSSM to a random broad space and obtain feature nodes and enhancement nodes, and concatenate the feature nodes and enhancement nodes as the input of the feature fusion layer. Specifically, feature nodes can be formally defined as follows:

$$\mathbf{Z}_i^{t/a/v} \triangleq \text{BiSSM}(\hat{\boldsymbol{\xi}}_j^{t/a/v})\mathbf{W}_{z_i}^{t/a/v} + \boldsymbol{\beta}_{z_i}^{t/a/v},\ i = 1, 2, \ldots, n \tag{12}$$

and the enhancement nodes can be computed as:

$$\mathbf{H}_j^{t/a/v} \triangleq \text{ReLU}(\mathbf{Z}_{t/a/v}^n \mathbf{W}_{h_j}^{t/a/v} + \boldsymbol{\beta}_{h_j}^{t/a/v}), j = 1, 2, \ldots, m \tag{13}$$

Furthermore, we introduce l2 regularization into the loss function to avoid the overfitting phenomenon of BLS, which is formally defined as follows:

$$\mathcal{L}_{norm} = \|[\mathbf{Z}_{t/a/v}^n | \mathbf{H}_{t/a/v}^m]\mathbf{W}_b^{t/a/v} - \mathbf{Y}^{t/a/v}\|_2^2 + \lambda\|\mathbf{W}_b^{t/a/v}\|_2^2 \tag{14}$$

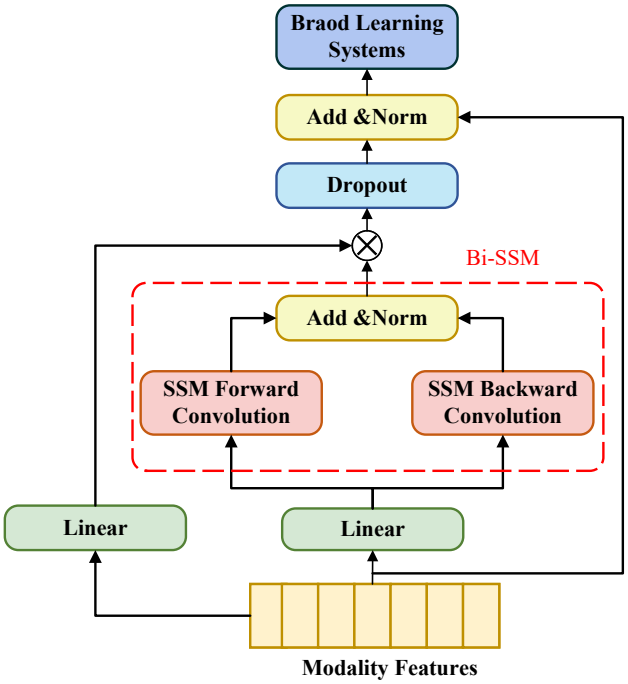

Fig. 4: The overall architecture of Broad Mamba. We use a bidirectional SSM to encode forward and reverse contextual semantic information.

where λ is the weight decay coefficient, $\mathbf{W}_b^{t/a/v}$ is the learnable parameters, $\mathbf{Y}^{t/a/v} = [\mathbf{Z}_1^{t/a/v}, \ldots, \mathbf{Z}_n^{t/a/v}, \ldots, \mathbf{H}_1^{t/a/v}, \ldots, \mathbf{H}_n^{t/a/v}]$.

By deriving and solving the Eq. 14, the solution to $\mathbf{W}_b^{t/a/v}$ can be calculated as follows:

$$\mathbf{W}_b^{t/a/v} = \left([\mathbf{Z}_{t/a/v}^n|\mathbf{H}_{t/a/v}^m]^\top [\mathbf{Z}_{t/a/v}^n|\mathbf{H}_{t/a/v}^m] + \lambda \mathbf{I}\right)^{-1} [\mathbf{Z}_{t/a/v}^n|\mathbf{H}_{t/a/v}^m]^\top \mathbf{Y}^{t/a/v} \tag{15}$$

Computation-Efficiency. SSM and the self-attention mechanism in Transformer both plays an important role in modeling global contextual semantic information. However, the self-attention mechanism is quadratic in complexity and is very time-consuming in training and inference. On the contrary, the computational complexity of SSM is $O(LlogL)$, so it can accelerate model inference in modeling long sequences.

4.2 Feature Fusion

Probability-guided Fusion Model. Many studies have proven that different modalities have different contributions to the prediction of emotional labels, so modal features with higher contributions need to be given greater weight in the multi-modal feature

fusion process. Different from previous works that fuse modal features at a coarse-grained level without using label information for guidance, we design a probability-guided fusion model (PFM) that dynamically assigns weights to each modality by using the predicted emotion label probabilities of the modalities. Specifically, we build an emotion classifier for the feature representation of each modality to obtain the predicted probability of the label as the weight of the modal features in the fusion process. The fusion process is formally defined as follows:

$$\omega^{t/a/v} = \text{Sigmoid}\left(\text{MLP}^{t/a/v}\left(\text{Y}^{t/a/v}\right)\right) \tag{16}$$

and then we can obtain the fused multi-modal feature representations as follows:

$$h^f = \omega^t \text{Y}^t + \omega^a \text{Y}^a + \omega^v \text{Y}^v \tag{17}$$

4.3 Training Loss

During the optimization phase of the model, the overall training loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{norm} + \mathcal{L}_{emo} \tag{18}$$

5 EXPERIMENTS

In comparative experiments, our experimental results are the average of 10 runs with different weight initializations. The results of our experiments are statistically significant (all $p < 0.05$) under paired t-tests.

5.1 Implementation Details

In the experiments, the number of feature nodes n and the number of enhancement node features m are set to 10 and 30 respectively. Following previous work, we use the same split ratio of training, test, and validation sets for model training and inference.

5.2 Datasets and Evaluation Metrics

We conduct experiments using two popular MERC datasets, IEMOCAP [2] and MELD [30], which include three modal data: text, audio, and video. IEMOCAP contains 12 hours of conversations, each containing six emotion labels. The MELD dataset contains conversation clips from the TV show Friends and contains seven different emotion labels. In addition, in the experiments we report the F1 of the proposed method and other baseline methods on each emotion category and weighted average F1 (W-F1).

5.3 Overall Results

Tables 1 show the experimental results on the IEMOCAP and MELD data sets. Experimental results show that our method significantly improves the performance of emotion recognition. The performance improvement may be attributed to the effective extraction of contextual semantic information and efficient integration of underlying data distribution.

Furthermore, our method is optimal compared with other multi-modal fusion methods in experimental results. The results demonstrate the effectiveness of our model in achieving multi-modal semantic information fusion. We also give W-F1 for each emotion. Specifically, on the IEMOCAP data set, our model's W-F1 is optimal on happy, neutral, and frustrated. On the MELD data set, our model's W-F1 is optimal on happy, neutral, and frustrated.

Table 1: Comparison with other baselines on the IEMOCAP and MELD dataset. The best result in each column is in bold.

Methods	IEMOCAP								MELD							
	Params.	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-F1	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	W-F1
bc-LSTM [29]	1.28M	34.4	60.8	51.8	56.7	57.9	58.9	54.9	73.8	47.7	5.4	25.1	51.3	5.2	38.4	55.8
A-DMN [43]	-	50.6	76.8	62.9	56.5	77.9	55.7	64.3	78.9	55.3	8.6	24.9	57.4	3.4	40.9	60.4
DialogueGCN [6]	12.92M	42.7	**84.5**	63.5	64.1	63.1	66.9	65.6	72.1	41.7	2.8	21.8	44.2	6.7	36.5	52.8
RGAT [15],	15.28M	51.6	77.3	65.4	63.0	68.0	61.2	65.2	78.1	41.5	2.4	30.7	58.6	2.2	44.6	59.5
CoMPM [18]	-	60.7	82.2	63.0	59.9	78.2	59.5	67.3	82.0	49.2	2.9	32.3	61.5	2.8	45.8	63.0
EmoBERTa [16]	499M	56.4	83.0	61.5	69.6	78.0	68.7	69.9	**82.5**	50.2	1.9	31.2	61.7	2.5	46.4	63.3
CTNet [23]	8.49M	51.3	79.9	65.8	67.2	**78.7**	58.8	67.5	77.4	50.3	10.0	32.5	56.0	11.2	44.6	60.2
LR-GCN [31]	15.77M	55.5	79.1	63.8	69.0	74.0	68.9	69.0	80.8	57.1	0	36.9	**65.8**	11.0	54.7	65.6
AdaGIN [41]	6.3M	53.0	81.5	71.3	65.9	76.3	67.8	70.7	79.8	60.5	15.2	43.7	64.5	**29.3**	56.2	66.8
DER-GCN [1]	78.59M	58.8	79.8	61.5	**72.1**	73.3	67.8	68.8	80.6	51.0	10.4	41.5	64.3	10.3	**57.4**	65.5
Our Model	1.73M	**65.5**	81.6	**73.5**	70.1	76.3	**69.8**	**73.3**	79.7	**65.6**	**16.9**	**48.9**	63.0	27.0	57.1	**67.6**

We also report the model parameter quantities of the proposed method and the baseline method. The results show that the parameter amount of our model is 1.73M, which is far lower than other methods. The model complexity of other baseline methods is relatively high, but the emotion recognition effect is relatively poor. Experimental results demonstrate that the proposed method is an effective and efficient MERC model.

5.4 Running Time

In this section, we report the inference time of different baselines and our proposed method on the IEMOCAP and MELD datasets. As shown in Table 2, the inference time of our method is below 10s, which is much lower than some GCN-based methods and RNN-based methods. Experimental results demonstrate the high efficiency of SSMs. In addition, we also counted the Flops of each method, and the results showed that our method was only slightly higher than bc-LSTM.

5.5 Ablation Studies

Ablation studies for PE, BLS, PFM. As shown in Table 3, we found that the performance of the model will decrease after removing PE, which indicates that positional

Table 2: We tested the running time of the proposed method and other comparative methods.

Methods	FLOPs(G)	Running time (s)	
		IEMOCAP	MELD
bc-LSTM	0.46	8.3	10.4
DialogueRNN	5.03	61.7	138.2
RGAT	6.87	68.5	146.3
DialogueGCN	4.81	58.1	127.5
LR-GCN	6.98	87.7	142.3
DER-GCN	20.83	125.5	189.7
Ours	0.71	3.5	6.6

encoding information is quite important for understanding contextual semantic information. Furthermore, without BLS, the performance of the model also degrades. The performance degradation is attributed to the underlying contextual data distribution which is also crucial for emotion prediction. Finally, when the PFM module is removed, the performance of the model drops sharply. Experimental results demonstrate the necessity of each proposed module.

Table 3: Ablation studies for PE, BLS, PFM.

Methods	IEMOCAP		MELD	
	W-Acc.	W-F1	W-Acc.	W-F1
Ours	**73.1**	**73.3**	**68.0**	**67.6**
w/o PE	$72.4_{(\downarrow 0.7)}$	$72.0_{(\downarrow 1.3)}$	$66.7_{(\downarrow 1.3)}$	$66.3_{(\downarrow 1.3)}$
w/o BLS	$71.5_{(\downarrow 1.6)}$	$72.1_{(\downarrow 1.2)}$	$65.5_{(\downarrow 2.5)}$	$64.9_{(\downarrow 2.7)}$
w/o PFM	$70.3_{(\downarrow 2.8)}$	$70.7_{(\downarrow 2.6)}$	$65.8_{(\downarrow 2.2)}$	$65.3_{(\downarrow 2.3)}$

Ablation studies for multi-modal features. To show the impact of different modal features on experimental results, we conducted ablation experiments to verify the combination of different modal features. From the experimental results in Table IV, it is found that: (1) In the single-modal experimental results of the model, the accuracy of emotion recognition in text mode is far better than the other two modes, indicating that text features play a dominant role in emotion recognition effect. (2) The emotion recognition effect using bimodal features is better than its own single-modality result. (3) The emotion recognition effect using three modal features is optimal. Experimental results prove the necessity of fusion of multi-modal features for emotion recognition.

Effect of Different Fusion Strategies. To study the effectiveness of the probability-guided fusion method proposed in this paper, we compare it with some previous multi-modal fusion strategies.

As shown in Fig. 5, compared with other fusion methods, the probability-guided fusion strategy we proposed has better emotion recognition effects on the two data sets. The results show that the emotion recognition effect of directly adding or concatenating multi-modal features to achieve multi-modal information fusion is relatively poor.

Table 4: The effect of our method using unimodal features, bimodal and multi-modal features, respectively.

Modality	IEMOCAP		MELD	
	W-Acc.	W-F1	W-Acc	W-F1
T+A+V	**73.1**	**73.3**	**68.0**	**67.6**
T	$65.5_{(\downarrow 7.6)}$	$65.7_{(\downarrow 7.6)}$	$64.6_{(\downarrow 3.4)}$	$63.9_{(\downarrow 3.7)}$
A	$58.6_{(\downarrow 14.5)}$	$58.8_{(\downarrow 14.5)}$	$52.7_{(\downarrow 15.3)}$	$52.0_{(\downarrow 15.6)}$
V	$49.4_{(\downarrow 23.7)}$	$49.7_{(\downarrow 23.6)}$	$40.1_{(\downarrow 27.9)}$	$41.4_{(\downarrow 26.2)}$
T+A	$71.3_{(\downarrow 1.8)}$	$70.2_{(\downarrow 3.1)}$	$65.2_{(\downarrow 2.8)}$	$65.6_{(\downarrow 2.0)}$
T+V	$68.7_{(\downarrow 4.4)}$	$67.4_{(\downarrow 5.9)}$	$65.0_{(\downarrow 3.0)}$	$63.7_{(\downarrow 3.9)}$
V+A	$62.1_{(\downarrow 11.0)}$	$62.2_{(\downarrow 11.1)}$	$51.3_{(\downarrow 16.7)}$	$51.9_{(\downarrow 15.7)}$

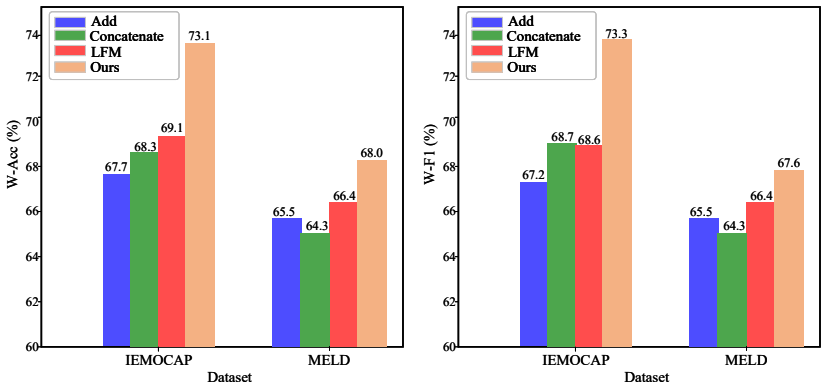

Fig. 5: Emotion recognition effects of different fusion methods. The experimental results are statistically significant (t-test with $p < 0.05$).

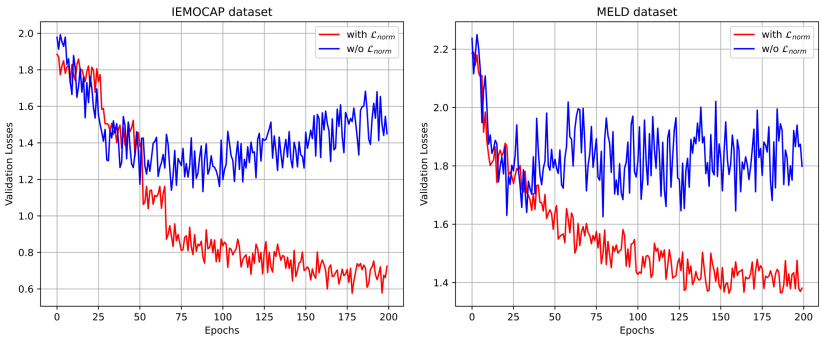

Fig. 6: Loss trends on IEMOCAP and MELD datasets.

The multi-modal information fusion effect of LMF is better than the adding method and the concatenating method. The probabilistic fusion strategy we propose introduces label information to guide the fusion of multi-modal information and further achieves parameter optimization of the model.

Effect of $\mathcal{L}_{norm}$. To illustrate the impact of $\mathcal{L}_{norm}$ on the experimental results, we conducted experiments on the IEMOCAP and MELD datasets to prove that $\mathcal{L}_{norm}$ can alleviate the problem of model overfitting. The experimental results are shown in Fig. 6. The loss curves on two datasets show that when $\mathcal{L}_{norm}$ is not introduced as a constraint, the model will overfit.

(a) IEMOCAP (b) MELD

Fig. 7: Visualizing feature embeddings for the multi-modal emotion on the IEMOCAP (**Left**) and MELD (**Right**) datasets. Each dot represents an utterance, and its color represents an emotion.

5.6 Multi-modal Representation Visualization

To intuitively demonstrate the classification results of our proposed method on the two data sets, we use t-SNE to project the high-dimensional multi-modal feature representation into a two-dimensional space, as shown in Fig. 7. The results show that the proposed method is able to effectively separate different emotion categories from each other.

5.7 Error Analysis

As shown in Fig. 8, we test the emotion classification results of DialogueRNN, DialogueGCN and the proposed method on the MELD dataset. In the disgust emotion category, the classification results of DialogueRNN and DialogueGCN are very poor, and they are all misclassified as neutral emotions. When the proposed method only uses text features, the emotion classification effect on the disgust category is unstable, but when multi-modal features are used, it can better classify disgust category emotions.

Turn	Speaker	Visual	Audio	Text	Dialogue RNN	Dialogue GCN	Ours (only text)	Ours	Ground Truth
1	Joey			Oh my god, you're back!	surprise	surprise	surprise	surprise	surprise
2	Phoebe			Ohh, let me see it! Let me see your hand!	surprise	surprise	surprise	surprise	surprise
3	Monica			Why do you want to see my hand?	*surprise*	*surprise*	neutral	neutral	neutral
4	Phoebe			I wanna see what's in your hand. I wanna see the trash.	*neutral*	*neutral*	*neutral*	disgust	disgust
5	Phoebe			Eww! Oh, it's all dirty. You should throw this out.	*neutral*	*neutral*	disgust	disgust	disgust

Fig. 8: An illustrative example of multi-modal emotion recognition in the MELD dataset.

6 Conclusions

In this work, we introduce a novel MERC method that comprehensively considers both feature disentanglement and multi-modal feature fusion. Specifically, during the feature disentanglement, we designed the broad Mamba, which incorporates the SSMs for data-dependent global emotional context modeling, and a broad learning system to explore the potential data distribution in the broad space. During the multi-modal feature fusion, we propose an effective probability-guided fusion mechanism to achieve multi-modal contextual feature fusion, which utilizes the predicted label probability of each modal feature as the weight vectors of the modal features. Extensive experiments conducted on two widely used benchmark datasets, IEMOCAP and MELD demonstrate the effectiveness and efficiency of our proposed method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62372478), the Research Foundation of Education Bureau of Hunan Province of China (Grant No. 22B0275), and the Hunan Provincial Natural Science Foundation Youth Project (Grant No. 2025JJ60420).

References

1. Ai, W., Shou, Y., Meng, T., Li, K.: Der-gcn: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. IEEE Transactions on Neural Networks and Learning Systems (2024)
2. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**, 335–359 (2008)

3. Chen, F., Sun, Z., Ouyang, D., Liu, X., Shao, J.: Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1064–1073 (2021)
4. Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., Onoe, N.: M2fnet: Multimodal fusion network for emotion recognition in conversation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4652–4661 (2022)
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on Deep Learning, December 2014 (2014)
6. Ghosal, D., Majumder, N., Poria, S., Chhaya, N., Gelbukh, A.: Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In: 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 (2019)
7. Ghosh, S., Varshney, D., Ekbal, A., Bhattacharyya, P.: Context and knowledge enriched transformer framework for emotion recognition in conversations. In: 2021 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2021)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
9. Gu, A., Goel, K., Re, C.: Efficiently modeling long sequences with structured state spaces. In: International Conference on Learning Representations (2021)
10. Hazarika, D., Poria, S., Mihalcea, R., Cambria, E., Zimmermann, R.: Icon: Interactive conversational memory network for multimodal emotion detection. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2594–2604 (2018)
11. Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.P., Zimmermann, R.: Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting. vol. 2018, p. 2122. NIH Public Access (2018)
12. He, W., Han, K., Tang, Y., Wang, C., Yang, Y., Guo, T., Wang, Y.: Densemamba: State space models with dense hidden connection for efficient large language models. arXiv preprint arXiv:2403.00818 (2024)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
14. Hu, J., Liu, Y., Zhao, J., Jin, Q.: Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 5666–5675 (2021)
15. Ishiwatari, T., Yasuda, Y., Miyazaki, T., Goto, J.: Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). pp. 7360–7370 (2020)
16. Kim, T., Vossen, P.: Emoberta: Speaker-aware emotion recognition in conversation with roberta. arXiv preprint arXiv:2108.12009 (2021)
17. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics (2014)
18. Lee, J., Lee, W.: Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5669–5679 (2022)

19. Li, B., Fei, H., Liao, L., Zhao, Y., Teng, C., Chua, T.S., Ji, D., Li, F.: Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5923–5934 (2023)
20. Li, W., Shao, W., Ji, S., Cambria, E.: Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis. Neurocomputing **467**, 73–82 (2022)
21. Li, W., Zhu, L., Mao, R., Cambria, E.: Skier: A symbolic knowledge integrated model for conversational emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 13121–13129 (2023)
22. Li, Z., Tang, F., Zhao, M., Zhu, Y.: Emocaps: Emotion capsule based model for conversational emotion recognition. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 1610–1618 (2022)
23. Lian, Z., Liu, B., Tao, J.: Ctnet: Conversational transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 985–1000 (2021)
24. Liu, S., Gao, P., Li, Y., Fu, W., Ding, W.: Multi-modal fusion network with complementarity and importance for emotion recognition. Information Sciences **619**, 679–694 (2023)
25. Lu, X., Zhao, Y., Wu, Y., Tian, Y., Chen, H., Qin, B.: An iterative emotion interaction network for emotion recognition in conversations. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 4078–4088 (2020)
26. Ma, H., Wang, J., Lin, H., Zhang, B., Zhang, Y., Xu, B.: A transformer-based model with self-distillation for multimodal emotion recognition in conversations. IEEE Transactions on Multimedia (2023)
27. Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., Cambria, E.: Dialoguernn: An attentive rnn for emotion detection in conversations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6818–6825 (2019)
28. Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B.: Long range language modeling via gated state spaces. In: International Conference on Learning Representations (2023)
29. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P.: Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 873–883 (2017)
30. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 527–536 (2019)
31. Ren, M., Huang, X., Li, W., Song, D., Nie, W.: Lr-gcn: Latent relation-aware graph convolutional network for conversational emotion recognition. IEEE Transactions on Multimedia **24**, 4422–4432 (2021)
32. Ren, M., Huang, X., Shi, X., Nie, W.: Interactive multimodal attention network for emotion recognition in conversation. IEEE Signal Processing Letters **28**, 1046–1050 (2021)
33. Shen, W., Chen, J., Quan, X., Xie, Z.: Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 13789–13797 (2021)
34. Shou, Y., Lan, H., Cao, X.: Contrastive graph representation learning with adversarial cross-view reconstruction and information bottleneck. Neural Networks **184**, 107094 (2025)
35. Shou, Y., Liu, H., Cao, X., Meng, D., Dong, B.: A low-rank matching attention based cross-modal feature fusion method for conversational emotion recognition. IEEE Transactions on Affective Computing (2024)
36. Shou, Y., Meng, T., Ai, W., Li, K.: Adversarial representation with intra-modal and inter-modal graph contrastive learning for multimodal emotion recognition. arXiv preprint arXiv:2312.16778 (2023)

37. Shou, Y., Meng, T., Ai, W., Yang, S., Li, K.: Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis. Neurocomputing **501**, 629–639 (2022)
38. Shou, Y., Meng, T., Ai, W., Yin, N., Li, K.: A comprehensive survey on multi-modal conversational emotion recognition with deep learning. arXiv preprint arXiv:2312.05735 (2023)
39. Smith, J.T., Warrington, A., Linderman, S.: Simplified state space layers for sequence modeling. In: The Eleventh International Conference on Learning Representations (2022)
40. Sun, J., Han, S., Ruan, Y.P., Zhang, X., Zheng, S.K., Liu, Y., Huang, Y., Li, T.: Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 658–670 (2023)
41. Tu, G., Xie, T., Liang, B., Wang, H., Xu, R.: Adaptive graph learning for multimodal conversational emotion detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19089–19097 (2024)
42. Wang, Y., Li, D., Shen, J.: Inter-modality and intra-sample alignment for multi-modal emotion recognition. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8301–8305. IEEE (2024)
43. Xing, S., Mai, S., Hu, H.: Adapted dynamic memory network for emotion recognition in conversation. IEEE Transactions on Affective Computing **13**(3), 1426–1439 (2020)
44. Yan, C., Liu, J., Liu, W., Liu, X.: Research on public opinion sentiment classification based on attention parallel dual-channel deep learning hybrid model. Engineering Applications of Artificial Intelligence **116**, 105448 (2022)
45. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1103–1114 (2017)
46. Zhang, D., Chen, F., Chen, X.: Dualgats: Dual graph attention networks for emotion recognition in conversations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7395–7408 (2023)
47. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)