

# A Low-rank Matching Attention based Cross-modal Feature Fusion Method for Conversational Emotion Recognition

Yuntao Shou, Huan Liu\*, Xiangyong Cao, Deyu Meng and Bo Dong

arXiv:2306.17799v2 [cs.CV] 15 Nov 2024

**Abstract**—Conversational emotion recognition (CER) is an important research topic in human-computer interactions. Although recent advancements in transformer-based cross-modal fusion methods have shown promise in CER tasks, they tend to overlook the crucial intra-modal and inter-modal emotional interaction or suffer from high computational complexity. To address this, we introduce a novel and lightweight cross-modal feature fusion method called Low-Rank Matching Attention Method (LMAM). LMAM effectively captures contextual emotional semantic information in conversations while mitigating the quadratic complexity issue caused by the self-attention mechanism. Specifically, by setting a matching weight and calculating inter-modal features attention scores row by row, LMAM requires only one-third of the parameters of self-attention methods. We also employ the low-rank decomposition method on the weights to further reduce the number of parameters in LMAM. As a result, LMAM offers a lightweight model while avoiding overfitting problems caused by a large number of parameters. Moreover, LMAM is able to fully exploit the intra-modal emotional contextual information within each modality and integrates complementary emotional semantic information across modalities by computing and fusing similarities of intra-modal and inter-modal features simultaneously. Experimental results verify the superiority of LMAM compared with other popular cross-modal fusion methods on the premise of being more lightweight. Also, LMAM can be embedded into any existing state-of-the-art CER methods in a plug-and-play manner, and can be applied to other multi-modal recognition tasks, e.g., session recommendation and humour detection, demonstrating its remarkable generalization ability.

**Index Terms**—Attention, Cross Modal Feature Fusion, Low Rank Decomposition, Multimodal Emotion Recognition

## I. INTRODUCTION

With the development of the multi-modal research field, Conversational Emotion Recognition (CER) that utilizes three

\* Corresponding Author: Huan Liu (huanliu@xjtu.edu.cn)

This work is supported by the National Nature Science Foundation of China (No. 62272374, No. 62192781, No. 62250009, No. 62137002, No. 62202367, No. 62272375), Natural Science Foundation of Shaanxi Province (No. 2024JC-JCQN-62), Project of China Knowledge Center for Engineering Science and Technology, Project of Chinese academy of engineering “The Online and Offline Mixed Educational Service System for ‘The Belt and Road’ Training in MOOC China”, and the K. C. Wong Education Foundation.

Y. Shou, X. Cao, and B. Dong are with School of Computer Science and Technology, the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, and the Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong, Xi'an, Shaanxi 710049, China. (shouyuntao@stu.xjtu.edu.cn, caoxiangyong@mail.xjtu.edu.cn, dong.bo@mail.xjtu.edu.cn).

H. Liu are with School of Computer Science and Technology, Xi'an Jiaotong, Xi'an, Shaanxi 710049, China, and the BigKE Joint Innovation Center, SHMEEA, Shanghai, China. (huanliu@xjtu.edu.cn).

D. Meng is with School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China. (dymeng@mail.xjtu.edu.cn).

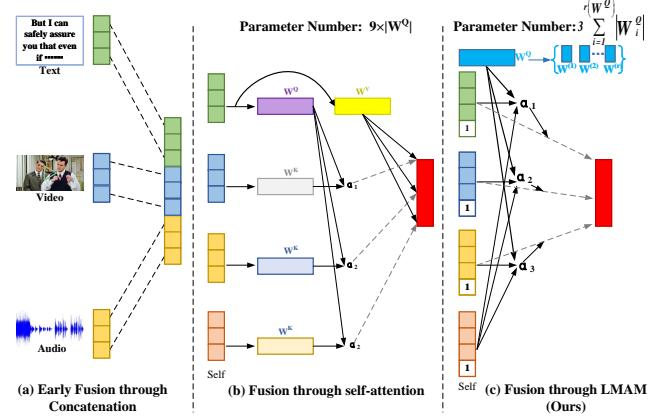


Fig. 1. (a) Early fusion through multimodal concatenation. (b) Multimodal feature fusion through self-attention.  $W^Q$ ,  $W^K$ , and  $W^V$  represent the learnable parameters of the query, key, and value vectors respectively. The parameter number of self-attention is 9 times of the parameter number of  $W^Q$ . (c) Multimodal feature fusion through LMAM. The parameter number of our LMAM fusion method is less than 3 times of the parameter number of  $W^Q$ .  $|W_i^Q|$  means the parameter number of the column vectors of  $W_Q$  after low rank decomposition.  $r(\cdot)$  indicates the rank of the matrix.

modal data (i.e., video, audio and text) to identify the speaker’s emotional changes during the conversation has become a hot research topic [6], [29]. Nowadays, CER has shown its promising performance in many practical social media scenarios. For example, in the field of intelligent recommendation, a recommendation system with emotional tendencies can recommend products that users are more interested in by identifying changes in consumers’ emotions. Therefore, it is of great importance to accurately identify the speaker’s emotional changes during the conversation.

Many existing studies [33], [38], [40] have proven that multi-modal feature fusion plays an important role in CER tasks. For example, as shown in Fig. 1(a), previous work [2] achieved multi-modal feature fusion by simply concatenating multi-modal features. Although this method has utilized multimodal data to a certain extent, it often adopts a strategy of treating each modality equally without fully considering the difference in the amount of emotional information between different modalities. The above method has two limitations: 1) the modality with less information may introduce noise during the fusion process, weakening the performance of the overall model; 2) the dominant role of the modality with more information in emotion recognition is not fully utilized. In addition, existing methods cannot capture contextual emotional

semantic information within and between modalities, which limits the performance of CER tasks.

To tackle the above problems, many Transformer-based methods (e.g., CTNet [24], and EmoCaps [23], etc) have been proposed and have become the current mainstream multi-modal feature fusion methods in CER task. Specifically, as shown in Fig. 1(b), these Transformer-based multi-modal feature fusion methods mainly use the self-attention mechanism to calculate the attention scores of different modal features, and achieve the fusion of intra-modal and inter-modal contextual emotional semantic information based on attention scores. Although the Transformer-based method can more accurately capture and fuse multimodal features in CER tasks, it also has two limitations. 1) This attention-based method has high computational complexity, especially when processing large-scale data, the demand for computing resources increases significantly. 2) As the model complexity increases, the risk of overfitting also increases accordingly, especially when the amount of training data is insufficient or the data diversity is limited. This may lead to a decrease in the generalization ability of the model and make it difficult to maintain stable performance in practical applications.

To solve these problems, existing research proposes low-rank decomposition methods [19] to reduce the number of parameters. For example, Zhu et al. [50] proposed a Low-rank Tensor Multi-modal Fusion method to achieve multi-modal data fusion through a low-rank tensor outer product and attention mechanism. Jin et al. [17] proposed a Fine-grained Temporal Low-rank Multi-modal Fusion (FT-LMF) method to fuse multi-modal information in time and space through low-rank tensor outer product. Sahay et al. [34] proposed Low-rank fusion based transformers to extract potential interactive information in multi-modal information through low-rank tensor outer product and cross-modal Transformer. Although the above method achieves the fusion of multimodal information through low-rank tensor outer product, this approach has significant limitations. Specifically, this fusion method lacks fine-grained processing capabilities when capturing the potential relationship information between discourses. Although the low-rank tensor outer product method can reduce the computational complexity to a certain extent, due to its inherent structural constraints, it cannot fully express the complex interactions between modalities, resulting in inaccurate capture of potential relationship information.

Different from the aforementioned CER methods, we thoroughly leverage the respective advantages of low-rank decomposition and Transformer techniques. Specifically, as shown in Fig. 1(c), we directly introduce a low-rank weight  $W$  into the self-attention, and then perform row-by-row attention calculation on multi-modal features to achieve fine-grained cross-modal information fusion. Our method not only retains the accuracy of Transformer in feature capture, but also effectively reduces the computational complexity through low-rank decomposition, avoiding the overfitting problem that may be caused by traditional methods.

Overall, we introduce a novel and lightweight cross-modal feature fusion method called Low-Rank Matching Attention Method (LMAM). LMAM effectively captures contextual

emotional semantic information in conversations while mitigating the quadratic complexity issue caused by the self-attention mechanism. Different from the existing Transformers with a self-attention mechanism, the number of parameters required by LMAM is less than one-third of the self-attention mechanism. Furthermore, LMAM is much easier to learn and may also reduce the risk of over-fitting. Fig. 1 illustrates a comparison of different cross-modal fusion methods.

Specifically, we first use the text, video and audio features pre-extracted by BERT, 3D-CNN and openSMILE respectively to obtain the corresponding query feature vector through a linear layer. In order to reduce the amount of parameters required by the network, we introduce low-rank weight decomposition to achieve compression of the learnable parameters  $W$  of the linear layer. Then we perform row-by-row attention calculations on the text query feature vector, video query feature vector, and audio query feature vector with the other two modal features (i.e., matching feature vectors). Finally, we fuse the obtained multi-modal features to obtain the final high-level discourse representation with emotional contextual semantic information.

Overall, the contributions of this work are summarized as:

- We introduce a novel and lightweight cross-modal feature fusion method called Low-Rank Matching Attention Method (LMAM). LMAM effectively captures contextual emotional semantic information in conversations while mitigating the quadratic complexity issue caused by the self-attention mechanism.
- By introducing only one learnable matching feature vectors and further utilizing the low-rank decomposition method, LMAM can significantly reduce the model complexity of existing Transformer-based cross-modal fusion methods. Furthermore, LMAM can also reduce the risk of overfitting.
- Extensive experiments also verify that the proposed LMAM method can be embedded into the existing DL-based CER methods to improve their recognition accuracy in a plug-and-play manner. In addition, LMAM is a general cross-modal feature fusion method and has potential application value in other multi-modal feature fusion tasks.

## II. RELATED WORK

### A. Conversational Emotion Recognition

Conversational emotion recognition (CER) involves cross-field knowledge such as cognitive science and brain science, and has received extensive attention from researchers [25], [44], [49]. Current CER research mainly includes three directions, i.e., sequential context-based emotion recognition, distinguishing speaker state-based emotion recognition, and speaker information-based emotion recognition [22].

For the sequential context-based CER approaches, Poria et al. [31] proposed a Bidirectional LSTM (bcLSTM), which utilizes recurrent neural units to extract the speaker's context information in the video, audio, and text features, and then uses the attention mechanism to fusion the information. Hazarika et al. [11] designed an interactive conversational

memory network (ICON) to extract the multi-modal features of different speakers following the idea of hierarchical modeling, and then input them into the global attention network for information fusion. Xing et al. [43] introduced an Adapted Dynamic Memory Network (A-DMN) to fine-grainedly model the dependencies between contextual utterances. Shahin et al. [36] proposed a dual-channel long short-term memory compressed-CapsNet to improve the hierarchical representation of contextual information.

For the different speaker states-based CER methods, Majumder et al. [28] proposed a DialogueRNN with three gating neural units (i.e., global GRU, party GRU and emotion GRU) to encode and update context and speaker information. Hu et al. [14] proposed a Contextual Reasoning Networks (CRN) to distinguish the speaker's emotional changes in the perceptual stage and the cognitive stage.

For the speaker information-based CER methods, Ghosal et al. [7] proposed a DialogueGCN to model the dialogue relationship between speakers by constructing a speaker relationship graph from the concatenated multi-modal feature vectors. Sheng et al. [37] designed a summarization and aggregation graph inference network (SumAggGIN) to consider global inferences related to dialogue topics and local inferences with adjacent utterances. Hu et al. [14] proposed a dialogue contextual reasoning network (DCRN) to extract contextual information from a cognitive perspective, and designed a multi-round reasoning module to fuse the emotional information. The DCDM proposed by Su et al. [39] introduced a causal directed acyclic graph (DAG) structure to establish complex correlations between hidden emotional information and other observed dialogue elements and construct a dynamic time disentanglement model to capture the hidden emotional change information in the dialogue.

### B. Multimodal Feature Fusion Approaches

In this subsection, we briefly review the multi-modal feature fusion methods for the CER task. Liu et al. [26] designed a low-rank multi-modal fusion method (LFM) to reduce the computational complexity caused by the change of tensor dimensions. Hu et al. [16] proposed a multi-modal fused graph convolutional network (MMGCN) to model dialogue relations between speakers and fuse the cross-modal features. Lian et al. [24] proposed a Conversational Transformer Network to fuse complementary semantic information from different modalities. Hu et al. [13] proposed Multimodal Dynamic Fusion Network (MM-DFN), which performs emotion recognition by eliminating contextual redundant information. Ma et al. [27] proposed a transformer-based self-distillation model to dynamically learn the intra-modal and inter-modal information interaction. Li et al. [21] proposed a Graph network based Multi-modal Fusion Technology (GraphMFT) for CER. GraphMFT achieves the fusion of intra-modal and inter-modal contextual semantic information by building multiple graph attention networks. The CFN-ESA proposed by Li et al. [20] effectively integrated the data distribution between modalities and promotes the interaction of multimodal information by constructing a unimodal encoder, a cross-modal encoder and

a sentiment conversion module. Although these multi-modal fusion approaches can obtain discriminative fused feature by exploiting the information of different modalities, they are either computationally expensive or do not fully consider the complementary information of different modals.

## III. PRELIMINARY INFORMATION

### A. Problem Definition

We assume the participants in the dialogue are  $P = \{p_1, p_2, \dots, p_N\}$ , where  $N$  represents the number of participants ( $N \geq 2$ ). We define sequential context  $T = \{t_1, t_2, \dots, t_M\}$ , where  $M$  represents the total number of sessions and  $t_i$  represents the  $i$ -th utterance. The task of CER is to identify the discrete emotions (e.g., happy, sad, disgust, neutral, excited, etc.) in each utterance.

### B. Multimodal Feature Extraction

In the CER task, three types of modality data are included, namely text, video and audio. The feature extraction method of each modal is different, and the semantic information they contain is also different [8]. Next, we will briefly introduce their data preprocessing methods.

**Word Embedding:** To obtain the feature vector representation of characters that computers can understand [35], we use the large-scale pre-training model BERT to encode text features. First, we use the Tokenizer method to segment the text to get each word and its index. We then feed them into the BERT model for feature encoding, and use the first 100-dimensional features in the BERT model as our text feature vectors.

**Visual Feature Extraction:** Following Hazarika et al. [12], we use 3D-CNN to extract the speaker's facial expression features and gesture change features in video frames, which have an important impact on the model's understanding of the speaker's emotional changes. Specifically, we utilize 3D-CNN and a linear layer with 512 neurons to obtain video feature vectors with rich semantic information.

**Audio Feature Extraction:** Following Hazarika et al. [12], we use openSMILE [5] to extract acoustic features in audio (e.g., loudness, Mel-spectra, MFCC). Specifically, we utilize the IS12\_ComParE1 extractor<sup>1</sup> in openSMILE and a linear layer with 100 neurons to obtain speaker audio features.

### C. Self-attention Mechanism

Self-Attention Mechanism is an algorithm used for sequence data processing to calculate the correlation between various positions of the input sequence. Owing to its powerful information capture capabilities, self-attention mechanisms have received considerable research attention in various fields, e.g., natural language understanding and text generation. A typical attention mechanism is to obtain the attention score by calculating the correlation between each element in the input sequence and other elements. For a given input sequence  $X$ , query  $Q$ , key  $K$  and value  $V$  are calculated as follows:

$$Q, K, V = W^Q X, W^K X, W^V X \quad (1)$$

<sup>1</sup><http://audeering.com/technology/opensmile>

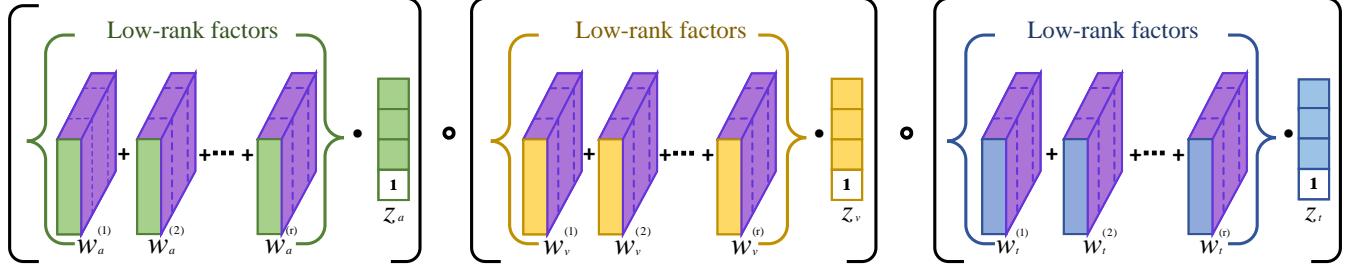


Fig. 2. LMAM achieves information fusion of multimodal features through parallel low-rank decomposition of weight (i.e.,  $\omega_a, \omega_v, \omega_t$ ) and modal features. We add an extra dimension to each modality feature and pad them with 1 to ensure that the intra-modal semantic information is preserved during inter-modal feature fusion.

where  $\mathbf{Q}$ ,  $\mathbf{K}$ , and  $\mathbf{V}$  are the learnable parameters.

Given the query  $\mathbf{Q}$ , key  $\mathbf{K}$  and value  $\mathbf{V}$ , the embeddings representation of the sequence are obtained through the dot product and Softmax function as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad (2)$$

In this work, we only use one query  $\mathbf{Q}$  as our matching feature vectors for self-attention calculation since the high complexity of the model easily leads to the risk of overfitting.

#### IV. PROPOSED LMAM CROSS-MODAL FUSION METHOD

In this section, we propose a novel cross-modal fusion method, namely Low-rank Matching Attention Mechanism (LMAM).

##### A. Matching Attention Layer

We concatenate the row vectors of all 1s with the extracted multi-modal features to obtain  $\xi$  as follows:

$$\xi = \{\xi_u, \xi_a, \xi_v\}, \quad (3)$$

where  $\xi_u$  represents context utterance features,  $\xi_a$  represents audio features, and  $\xi_v$  represents video features.

The existing CER approaches usually use feature splicing or feature summation to fuse the cross-modal feature [16], [26], [26], [45]. As introduced in the related work, these cross-modal fusion methods are either computationally expensive or do not fully consider the complementary information of different modals. Therefore, our goal is to construct an efficient and effective fusion method that captures the differences among multimodal features by computing the correlation among the three modalities of text, video and audio and realizes the fusion of complementary semantic information across modalities. Specifically, the computation process of our proposed LMAM fusion method is shown as follows.

For a given model input  $\mathbf{K}_i$ ,  $\mathbf{V}_i$  and  $\mathbf{M}_i$ , where  $\mathbf{K}_i = \xi_i$ ,  $\mathbf{V}_i = \xi_i$ , and  $\mathbf{M}_i = \xi_i$ . We first compute the query matrix  $\mathbf{Q}_i \in \mathbb{R}^{L_{Q_i} \times d_{Q_i}}$  by linear transformation from  $\mathbf{M}_i$  as follows:

$$\mathbf{Q}_i = \mathbf{M}_i \mathbf{W}^{Q_i} + \mathbf{b}^{Q_i}, \quad (4)$$

where  $\mathbf{K}_i$  and  $\mathbf{M}_i$  represent the  $i$ -th modal features.  $L_{Q_i}$  represents the sequence length of the modal features.  $d_{Q_i}$  represents the feature dimension after linear layer mapping,

---

#### Algorithm 1 Matching Attention Mechanism (LMAM)

**Input:** Text feature vectors  $\xi_t$ , video feature vectors  $\xi_v$  and audio feature vectors  $\xi_a$ ; the number of iterations  $\epsilon$ ; the size of dataset  $\phi$ .

**Output:** The enhanced multimodal fusion feature vectors  $\xi_f$ .

```

1: Initialize the model weights  $W$  and bias  $b$ .
2: Initialize the set of multimodal feature fusion  $\xi_{fusion}$ .
3: for EPOCH  $\leftarrow 1, 2, \dots, \epsilon$  do
4:   for  $i \leftarrow 1, 2, \dots, \text{len}(\phi)/32$  do
5:     Sample a batch  $\xi = \{\xi_t, \xi_v, \xi_a\}_{i=1}^{32}$ .
6:     for modal in  $\xi$  do
7:       for t in modal do
8:         att_emo, score=MatchingAttention(modal, t)
9:         att_emotions.append(att_emo.unsqueeze(0))
10:        scores.append(score[:, 0, :])
11:      end for
12:      att_emotions=torch.cat(att_emotions, dim=0)
13:       $\Psi = \text{att\_emotions} + \text{F.gelu}(\text{modal})$ 
14:    end for
15:  end for
16: end for
17:  $\xi_{fusion} = \{\Psi_1, \Psi_2, \dots, \Psi_{\text{len}(\phi)/32}\}$ .
18: return the enhanced multimodal feature vectors  $\xi_{fusion}$ .

```

---

and  $\mathbf{W}^{Q_i} \in \mathbb{R}^{d_m \times d_{Q_i}}$ ,  $d_m$  represents the feature embedding dimension of  $Q_i$ .

Next, we get the attention score using the following formula:

$$\alpha_i = \text{softmax}\left(\text{Tanh}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{d_k}}\right)\right) \quad (5)$$

where  $\top$  represents the matrix transpose operation.  $d_k$  represents the sequence length of the modal features,  $\alpha_i$  represents the attention score, and  $\alpha_i \in [0, 1]$ .

Subsequently, we perform matrix multiplication by the attention score and the modality feature  $I_i$  to obtain the attention output as follows:

$$\mathbf{A}_i = \alpha_i \mathbf{V}_i \quad (6)$$

where  $\mathbf{A}_i$  is the feature vector after attention calculation. As shown in Fig. 3, we present the fusion process of text features and audio features using a matched attention mechanism.

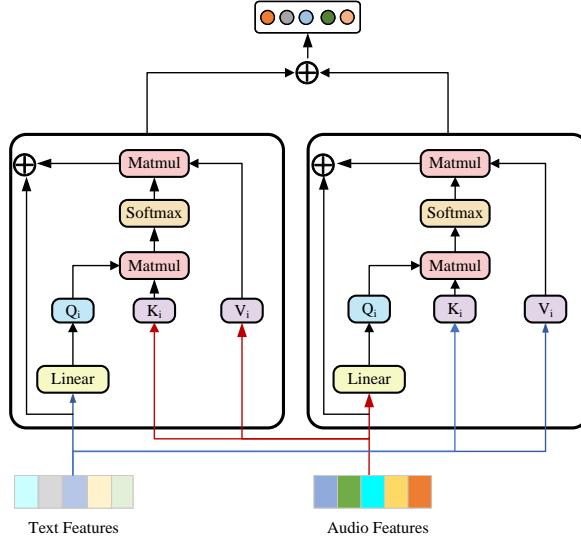


Fig. 3. An illustrative example of the proposed LMAM method fusing complementary semantic information between different modal features. We show the fusion process of text features and audio features. The fusion process for text and video, and video and audio follows the same paradigm.

Similarly, the fusion process for text and video, and video and audio follows the same paradigm.

To prevent the problem of gradient disappearance and information collapse in the model training, we also build a residual connection layer with normalization operation. Finally, we use a linear layer with ReLU activation function to get the final output of the LMAM. The formulas are as follows:

$$\begin{aligned} \mathbf{E}_i &= \text{Norm}(\mathbf{A}_i + \mathbf{I}_i), \\ \mathbf{O}_i &= \text{ReLU}(\text{Linear}(\mathbf{E}_i)). \end{aligned} \quad (7)$$

#### B. Low-rank Weight Decomposition

Low-rank weight decomposition technology achieves parameter reduction and model compression by decomposing the original weight matrix into the product of two or more low-rank matrices [19]. Since it can effectively reduce the complexity of the model, low-rank decomposition technology has received extensive research attention on many tasks, e.g., image processing and multi-modal fusion. Tucker decomposition is a commonly used technique for tensor decomposition, which can decompose a high-order tensor into a smaller core tensor and a set of factor matrices. Unlike other tensor decomposition methods, Tucker decomposition allows the core tensor to maintain complex relationships between multiple modes, thus being able to capture multimodal dependencies in the data.

As shown in Fig. 2, the idea behind low-rank decomposition in LMAM is to decompose the weight  $\mathbf{W}$  into specific factors that match the modal features. For any  $N$ -order weight  $\mathbf{W}_i$ , there is always a low-rank decomposition method. The formula is defined as follows:

$$\begin{aligned} \widetilde{\mathbf{W}}_i &= \text{Tucker}(\mathbf{W}_i) \\ &= \sum_{j=1}^R \omega_{n,i}^{(j)}, \omega_{n,i}^{(j)} \in \mathbb{R}_n^d \end{aligned} \quad (8)$$

where  $r$  represents the rank of the weight  $\mathbf{W}_i$ ,  $\left\{ \left\{ \omega_{n,i}^{(j)} \right\}_{n=1}^N \right\}_{j=1}^r$  is a collection of low-rank decomposition factors, and  $\eta$  is the size of the multimodal feature set  $\xi = \{\xi_u, \xi_a, \xi_v\}$ , i.e.,  $\eta = 3$ .  $\widetilde{\mathbf{W}}_i$  is computed during training.

Therefore, Eq. 4 can be calculated as:

$$\mathbf{Q}_i = \left( \sum_{j=1}^r \omega_{n,i}^{(j)} \right) \mathbf{M}_i + \mathbf{b}^{\mathbf{Q}_i}, \omega_{n,i}^{(j)} \in \mathbb{R}_n^d \quad (9)$$

The whole computational process of the LMAM method is shown in Figure 1(c) and the pseudocode of the LMAM method is summarized in Algorithm 1.

#### C. Comparison to Self-attention

Studies [4] have shown that the performance of the self-attention mechanism is lower than CNN, RNN and other methods when the amount of data is small, while its performance can gradually exceed CNN and RNN when the amount of data is very large. The difference in performance may be attributed to that the self-attention mechanism needs to learn the query vectors  $\mathbf{Q}$ , the key vectors  $\mathbf{K}$ , and the value vectors  $\mathbf{V}$  at the same time, which makes the optimization of the model more difficult. Unlike classic self-attention, LMAM only needs a very low rank weight to achieve better performance than self-attention. Specifically, we only set a learnable parameter  $\mathbf{W}^Q$  for cross-modal feature fusion and capture of complementary semantic information. Furthermore, we perform a parallel low-rank decomposition of  $\mathbf{W}^Q$  with modality-specific factors to further reduce the amount of parameters required for  $\mathbf{W}^Q$ . LMAM can reduce the difficulty of network optimization while maintaining performance.

TABLE I  
THE DIVISION OF TRAINING SET, VALIDATION SET AND TEST SET OF IEMOCAP AND MELD DATA SETS.

Datasets	Partition	Utterance Count	Dialogue Count
IEMOCAP	train+val	5810	120
	test	1623	31
MELD	train+val	11098	1153
	test	2610	280
CMU-MOSI	train+val	1513	62
	test	686	31
POM	train+val	700	N/A
	test	203	N/A

#### D. Network Architecture of LMAM Module

In this section, we design a network to implement the LMAM method. The overall network architecture of the LMAM module is illustrated in Figure 4(d). From Figure 4(d), we can observe that the LMAM module first receives three modal data as input, and then generates two types of vectors (i.e., query feature vector and matched feature vector) by a linear transformation layer. Subsequently, we compute the attention score based on these feature vectors. Finally, we generate the final fusion feature by conducting cross-modal feature fusion followed by a fine-tuning step.

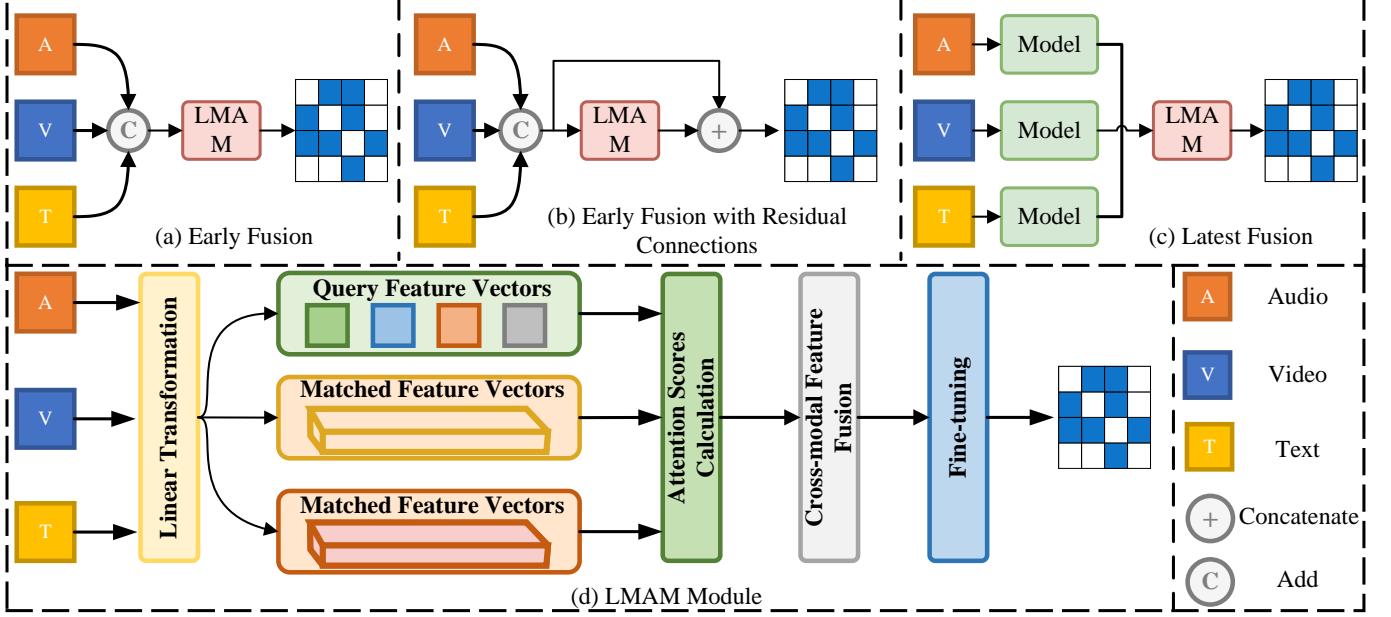


Fig. 4. Three embedding ways for cross-modal fusion using LMAM module. (a) Embedding the LMAM module before model's feature extraction. (b) Embedding the LMAM module and using the residual connections before model's feature extraction. (c) Embedding the LMAM module after model's feature extraction. (d) The overall flow of the LMAM module.  $Q_i$  is represented as the query feature vectors, and  $I_i$  is represented as the matching feature vectors.

TABLE II

EXPERIMENTAL RESULTS ON IEMOCAP DATASET. METHODS WITH \* REPRESENT THE METHOD EQUIPPED WITH OUR LMAM MODULE WITHOUT ANY FURTHER CHANGES. THE BEST RESULT IN EACH COLUMN IS IN BOLD.

Methods	IEMOCAP						
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Average(w)
TextCNN [18]	27.73 29.81	57.14 53.83	34.36 40.13	<b>61.12</b> 52.47	46.11 50.09	<b>62.94</b> 55.78	48.89 48.08
TextCNN*	<b>37.24</b> <b>37.37</b>	<b>81.41</b> <b>72.97</b>	<b>50.57</b> <b>53.95</b>	57.21 <b>62.96</b>	<b>70.28</b> <b>63.87</b>	59.01 <b>59.16</b>	<b>60.35</b> <b>59.34</b>
$\Delta$	<b>↑9.51</b> <b>↑7.56</b>	<b>↑24.27</b> <b>↑19.14</b>	<b>↑16.21</b> <b>↑13.82</b>	↓3.91 <b>↑10.49</b>	<b>↑24.17</b> <b>↑13.78</b>	↓3.87 <b>↑3.29</b>	<b>↑11.46</b> <b>↑11.26</b>
bc-LSTM [31]	29.16 34.49	57.14 60.81	54.19 51.80	57.03 56.75	51.17 57.98	<b>67.12</b> 58.97	55.19 54.96
bc-LSTM+Att [31]	30.56 <b>35.63</b>	56.73 62.09	57.55 53.00	59.41 <b>59.24</b>	52.84 58.85	65.88 <b>59.41</b>	56.31 56.08
bc-LSTM*	<b>86.67</b> 31.52	<b>66.43</b> <b>73.08</b>	<b>63.09</b> <b>55.54</b>	<b>81.51</b> 49.79	<b>68.08</b> <b>77.01</b>	50.39 59.31	<b>65.55</b> <b>60.29</b>
$\Delta$	<b>↑56.11</b> ↓4.11	<b>↑9.7</b> <b>↑10.99</b>	<b>↑5.54</b> <b>↑2.54</b>	<b>↑22.1</b> ↓9.45	<b>↑15.24</b> <b>↑18.16</b>	↓15.49 ↓0.1	<b>↑10.36</b> <b>↑5.33</b>
DialogueRNN [28]	25.63 33.11	<b>75.14</b> <b>78.85</b>	58.56 59.24	64.76 <b>65.23</b>	<b>80.27</b> 71.85	<b>61.16</b> 58.97	63.40 62.77
DialogueRNN*	<b>63.51</b> <b>66.43</b>	72.43 78.64	<b>62.94</b> <b>59.53</b>	<b>78.89</b> 55.04	77.00 <b>77.65</b>	55.07 <b>59.07</b>	<b>66.84</b> <b>65.79</b>
$\Delta$	<b>↑37.88</b> <b>↑33.32</b>	↓2.71 ↓0.21	↑4.38 <b>↑0.29</b>	<b>↑14.13</b> ↓10.19	↓3.27 ↑5.8	↓6.09 ↑0.1	<b>↑3.44</b> <b>↑3.02</b>
DialogueGCN [7]	40.63 42.71	<b>89.14</b> <b>84.45</b>	61.97 <b>63.54</b>	67.51 <b>64.14</b>	65.46 63.08	<b>64.13</b> <b>66.90</b>	65.91 65.62
DialogueGCN*	<b>62.60</b> <b>59.69</b>	76.86 78.98	<b>62.82</b> 60.06	<b>74.40</b> 63.48	<b>72.38</b> <b>77.93</b>	56.40 58.30	<b>66.39</b> <b>66.12</b>
$\Delta$	<b>↑21.97</b> <b>↑16.98</b>	↓12.28 ↓5.47	<b>↑0.85</b> ↓3.48	<b>↑6.89</b> ↓0.66	<b>↑6.92</b> <b>↑14.85</b>	↓7.73 ↓8.6	<b>↑0.48</b> <b>↑0.51</b>
MM-DFN [13]	40.17 42.22	<b>74.27</b> 78.98	<b>69.13</b> <b>66.42</b>	70.25 <b>69.97</b>	76.99 75.56	<b>68.58</b> <b>66.33</b>	68.77 68.20
MM-DFN*	<b>74.58</b> <b>69.57</b>	74.01 <b>81.03</b>	64.93 63.71	<b>73.91</b> 66.67	<b>81.40</b> <b>80.00</b>	59.84 61.71	<b>69.94</b> <b>69.69</b>
$\Delta$	<b>↑34.41</b> <b>↑27.35</b>	↓0.26 <b>↑2.05</b>	↓4.2 ↓2.71	<b>↑3.66</b> ↓3.3	<b>↑4.41</b> <b>↑4.44</b>	↓8.74 ↓4.62	<b>↑1.16</b> <b>↑1.48</b>
M2FNet [3]	65.92 60.00	<b>79.18</b> <b>82.11</b>	65.80 <b>65.88</b>	<b>75.37</b> <b>68.21</b>	74.84 72.60	<b>59.87</b> <b>68.31</b>	69.11 69.86
M2FNet*	<b>73.35</b> <b>69.53</b>	74.37 81.42	<b>67.26</b> 63.51	72.86 66.23	<b>81.57</b> <b>81.29</b>	59.64 62.50	<b>70.31</b> <b>70.07</b>
$\Delta$	<b>↑7.43</b> <b>↑9.53</b>	↓4.81 ↓0.69	<b>↑1.46</b> ↓2.37	↓2.51 ↓1.98	<b>↑6.73</b> <b>↑8.69</b>	↓0.23 ↓5.81	<b>↑1.20</b> <b>↑0.21</b>
EmoCaps [23]	<b>70.34</b> 72.86	77.39 82.45	64.27 65.10	71.79 <b>69.14</b>	84.50 73.90	<b>63.94</b> 63.41	71.22 70.06
EmoCaps*	69.93 <b>74.31</b>	<b>82.52</b> <b>85.47</b>	<b>68.41</b> <b>67.03</b>	<b>79.49</b> 65.26	<b>84.85</b> <b>80.14</b>	63.33 <b>68.38</b>	<b>73.67</b> <b>73.01</b>
$\Delta$	↓0.41 <b>↑1.45</b>	<b>↑5.13</b> <b>↑3.02</b>	<b>↑4.14</b> <b>↑1.93</b>	<b>↑7.7</b> ↓3.88	<b>↑0.35</b> <b>↑6.24</b>	↓0.61 <b>↑4.97</b>	<b>↑2.45</b> <b>↑2.95</b>

As shown in Figure 4, there are three ways to use the proposed LMAM module, i.e., early fusion, early fusion with residual connections, and late fusion. For early fusion, we concatenate the three modalities and then input them into the LMAM module for feature fusion. For early fusion with residual connections, the concatenated features vectors of our

three modalities are added to the feature vectors after feature fusion through the LMAM module. For late fusion, we extract the contextual semantic information from the model (e.g., EmoCaps) and then input it to the LMAM module for feature fusion. It should be noted that the selection of the LMAM fusion ways depends on the baseline model itself. In the

TABLE III

EXPERIMENTAL RESULTS ON MELD DATASET. METHODS WITH \* REPRESENT THE METHOD EQUIPPED WITH OUR LMAM MODULE WITHOUT ANY FURTHER CHANGES. THE BEST RESULT IN EACH COLUMN IS IN BOLD.

Methods	MELD							
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Average(w)
	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1	Acc. F1
TextCNN [18]	<b>71.23</b> 74.91	<b>43.35</b> <b>45.51</b>	<b>4.63</b> <b>3.71</b>	18.25 21.17	46.14 49.47	<b>8.91</b> <b>8.36</b>	35.33 34.51	52.48 55.09
TextCNN*	70.23 <b>75.79</b>	36.47 44.78	0.00 0.00	<b>24.19</b> <b>21.19</b>	<b>50.58</b> <b>52.46</b>	0.00 0.00	<b>43.43</b> <b>41.68</b>	<b>53.17</b> <b>56.55</b>
Δ	↓1.00 ↑0.88	↓6.88 ↓0.73	↓4.63 ↓3.71	↑5.94 ↑0.02	↑4.44 ↑2.99	↓8.91 ↓8.36	↑8.10 ↑7.17	↑0.69 ↑1.47
bc-LSTM [31]	<b>71.45</b> 73.84	46.82 <b>47.71</b>	<b>3.84</b> <b>5.46</b>	22.47 <b>25.19</b>	51.61 51.34	<b>4.31</b> <b>5.23</b>	36.71 38.44	54.18 55.89
bc-LSTM+Att	70.45 <b>75.55</b>	46.43 46.35	0.00 0.00	21.77 16.27	49.30 50.72	0.00 0.00	41.77 40.71	53.73 55.82
bc-LSTM*	70.78 75.46	<b>47.18</b> 46.47	0.00 0.00	<b>26.09</b> 24.58	<b>52.33</b> <b>53.11</b>	0.00 0.00	<b>43.23</b> <b>40.92</b>	<b>54.97</b> <b>56.85</b>
Δ	↑0.33 ↓0.09	↑0.75 ↑0.12	↑0.00 ↑0.00	↑4.32 ↑8.31	↑3.03 ↑2.39	↑0.00 ↑0.00	↑1.46 ↑0.21	↑1.24 ↑1.03
DialogueRNN [28]	<b>72.12</b> 73.54	<b>46.42</b> <b>49.47</b>	1.61 1.23	23.97 <b>23.83</b>	<b>52.01</b> 50.74	1.52 1.73	41.01 41.54	55.10 55.97
DialogueRNN*	71.74 <b>75.76</b>	45.83 48.23	<b>3.13</b> <b>2.77</b>	<b>31.71</b> 17.93	49.25 <b>53.04</b>	<b>2.01</b> <b>2.58</b>	<b>42.40</b> <b>42.21</b>	<b>55.27</b> <b>56.93</b>
Δ	↓0.38 ↑2.22	↓0.59 ↓1.24	↑1.52 ↑1.54	↑7.74 ↓5.90	↓2.76 ↑2.30	↑0.49 ↑0.85	↑1.39 ↑0.67	↑0.17 ↑0.96
DialogueGCN [7]	75.61 77.45	51.32 52.76	<b>5.14</b> <b>10.09</b>	30.91 32.56	<b>54.31</b> <b>56.08</b>	<b>11.62</b> <b>11.27</b>	42.51 44.65	58.74 60.55
DialogueGCN*	<b>78.19</b> <b>77.82</b>	<b>52.27</b> <b>54.11</b>	2.17 2.31	<b>35.79</b> <b>36.43</b>	54.15 55.07	4.05 2.12	<b>48.31</b> <b>47.22</b>	<b>60.96</b> <b>60.98</b>
Δ	↑2.58 ↑0.37	↑0.95 ↑1.35	↓2.97 ↓7.78	↑4.88 ↑3.87	↓0.16 ↓1.01	↓7.57 ↓9.15	↑5.80 ↑2.57	↑2.22 ↑0.43
MM-DFN [13]	<b>78.17</b> <b>77.76</b>	52.15 50.69	0.00 0.00	25.77 22.93	<b>56.19</b> <b>54.78</b>	0.00 0.00	<b>48.31</b> <b>47.82</b>	60.30 59.44
MM-DFN*	77.08 76.56	<b>53.79</b> <b>56.84</b>	<b>2.07</b> <b>4.11</b>	<b>38.10</b> <b>31.92</b>	53.63 50.53	<b>4.23</b> <b>7.10</b>	47.99 46.08	<b>60.65</b> <b>59.62</b>
Δ	↓1.09 ↓1.20	↑1.64 ↑6.15	↑2.07 ↑4.11	↑12.33 ↑8.99	↓2.56 ↓4.25	↑4.23 ↑7.10	↓0.32 ↓1.74	↑0.35 ↑0.18
M2FNet [3]	<b>68.88</b> <b>67.98</b>	72.76 58.66	5.57 3.45	50.09 <b>47.03</b>	68.49 65.50	<b>17.69</b> <b>25.24</b>	57.33 55.25	63.64 60.87
M2FNet*	68.40 67.27	<b>73.15</b> <b>60.37</b>	<b>9.13</b> <b>11.25</b>	<b>51.77</b> 46.68	<b>69.11</b> <b>65.92</b>	15.19 17.62	<b>60.76</b> <b>57.31</b>	<b>64.13</b> <b>60.97</b>
Δ	↓4.48 ↓0.71	↑0.39 ↑1.71	↑3.56 ↑7.80	↑1.68 ↓0.35	↑0.62 ↑0.42	↓2.50 ↓7.62	↑3.43 ↑2.06	↑0.50 ↑0.10
EmoCaps [23]	75.24 <b>75.12</b>	63.57 63.19	<b>3.45</b> <b>3.03</b>	<b>43.78</b> <b>42.52</b>	58.34 57.05	<b>7.01</b> <b>7.69</b>	58.79 57.54	63.52 62.97
EmoCaps*	<b>76.37</b> 74.28	<b>66.57</b> <b>64.74</b>	3.11 2.14	40.17 42.35	<b>63.33</b> <b>62.52</b>	6.21 7.05	<b>59.45</b> <b>60.26</b>	<b>64.93</b> <b>63.88</b>
Δ	↑1.13 ↓0.84	↑3.00 ↑1.55	↓0.34 ↓0.89	↓3.61 ↓0.17	↑4.99 ↑5.47	↓0.80 ↓0.84	↑0.66 ↑2.72	↑1.41 ↑0.97

following experiment, we mainly adopt the latter two ways of fusion, i.e., early fusion with residual connections framework and late fusion.

## V. EXPERIMENTS

In this section, we conduct several experiments to verify the effectiveness of our proposed LMAM cross-modal fusion method. Specifically, the overall experimental setting is shown as follows. Firstly, we choose seven state-of-the-art DL-based approaches, including TextCNN [18], bc-LSTM [31], DialogueRNN [28], DialogueGCN [7], MM-DFN [13], M2FNet [3], and EmoCaps [23], as backbones and embed the proposed LMAM fusion method into these approaches. In particular, TextCNN uses CNN, 3D-CNN and openSMILE to extract text, video and audio features respectively and input the obtained multi-modal data into 1D-CNN to complete emotion classification. Secondly, we compare our proposed LMAM method with four popular cross-modal fusion methods, including classical add operation and concatenate operation, and latest low-rank multi-modal fusion (LFM) [26], tensor fusion network (TFN) [45], Dual Low-Rank Multimodal Fusion (Dual-LMF) [17], Low-Rank Multimodal Fusion with Self-Attention (Att-LMF) [50]. and Low-Rank Multimodal Fusion with multi-modal Transformer (LMF-MuLT) [34]. Thirdly, we conduct an ablation study to verify the necessity of considering the multi-modal datasets. Finally, we apply the proposed LMAM method to other multi-modal recognition tasks. All the experiments are conducted on a PC with Intel Core i7-8700K CPU, and one GeForce RTX 3090 with 24GB memory.

### A. Datasets and Evaluation Metrics

The IEMOCAP [1], MELD [32] CMU-MOSI [47], and POM [30] datasets are widely used for conversational emotion recognition. Therefore, this paper selects these four benchmark datasets to verify the effectiveness of our LMAM fusion method. The IEMOCAP dataset contains three modal data to meet the needs of multimodal research, namely video, text, and audio. The IEMOCAP dataset contains 151 dialogues and 7433 utterances of 5 actors and 5 actresses. The emotional labels of the IEMOCAP dataset were annotated by at least three experts, and they divided the labels into six categories, namely “happy”, “neutral”, “sad”, “excited”, “angry” and “frustrated”. The MELD also includes video, text, and audio three modal data. The MELD dataset contains 13,708 utterances and 1,433 dialogues by multiple actors for 13.7 hours. The emotional labels of the MELD dataset were annotated by at least five experts, and they divided the labels into seven categories, namely “fear”, “neutral”, “angry”, “joy”, “sadness”, “disgust” and “surprise”. The CMU-MOSI data set is a multi-modal sentiment analysis data set, including video, audio, text and other data modalities. The CMU-MOSI dataset is annotated with an emotion label in the range [-3, 3]. The POM dataset contains 903 movie review videos, each of which is annotated with 16 speaker features. The speaker in each video is annotated with an emotion label including confidence, enthusiasm, and other characteristics.

The IEMOCAP dataset only contains the training set and the test set, so we divide the test set into a test set and a validation set at a ratio of 8:2. The MELD dataset includes a training set, a test set, and a validation set. Similarly, the

TABLE IV

EXPERIMENTAL RESULTS ON CMU-MOSI, AND POM DATASET. METHODS WITH \* REPRESENT THE METHOD EQUIPPED WITH OUR LMAM MODULE WITHOUT ANY FURTHER CHANGES. THE BEST RESULT IN EACH COLUMN IS IN BOLD.

Methods	CMU-MOSI					POM		
	Acc-2 ( $\uparrow$ )	F1 ( $\uparrow$ )	MAE ( $\downarrow$ )	Corr ( $\uparrow$ )	Acc-7 ( $\uparrow$ )	MAE ( $\downarrow$ )	Acc ( $\uparrow$ )	Corr ( $\uparrow$ )
TextCNN [18]	62.2	62.0	1.42	0.41	23.3	0.88	34.0	0.13
TextCNN*	<b>69.7</b>	<b>69.6</b>	<b>1.31</b>	<b>0.49</b>	<b>24.6</b>	<b>0.87</b>	<b>34.0</b>	<b>0.21</b>
$\Delta$	<b>↑7.5</b>	<b>↑7.6</b>	<b>↑0.11</b>	<b>↑0.08</b>	<b>↑1.3</b>	<b>↑0.01</b>	<b>↑0.0</b>	<b>↑0.08</b>
bc-LSTM [31]	73.9	73.9	1.08	0.61	28.7	0.84	34.8	0.28
bc-LSTM+Att	75.2	75.2	1.05	0.62	34.6	0.82	35.3	0.30
bc-LSTM*	<b>77.0</b>	<b>77.1</b>	<b>0.96</b>	<b>0.64</b>	<b>34.7</b>	<b>0.80</b>	<b>39.1</b>	<b>0.35</b>
$\Delta$	<b>↑1.8</b>	<b>↑1.9</b>	<b>↑0.09</b>	<b>↑0.02</b>	<b>↑0.1</b>	<b>↑0.02</b>	<b>↑3.8</b>	<b>↑0.05</b>
DialogueRNN [28]	75.7	75.8	1.02	0.62	34.6	0.81	36.2	0.32
DialogueRNN*	<b>77.4</b>	<b>77.0</b>	<b>1.00</b>	<b>0.67</b>	<b>34.9</b>	<b>0.77</b>	<b>39.3</b>	<b>0.37</b>
$\Delta$	<b>↑1.7</b>	<b>↑1.2</b>	<b>↑0.02</b>	<b>↑0.05</b>	<b>↑0.3</b>	<b>↑0.04</b>	<b>↑3.1</b>	<b>↑0.05</b>
DialogueGCN [7]	74.3	74.0	1.06	0.61	29.6	0.83	35.0	0.29
DialogueGCN*	<b>74.9</b>	<b>74.7</b>	<b>1.04</b>	<b>0.62</b>	<b>31.8</b>	<b>0.79</b>	<b>36.7</b>	<b>0.35</b>
$\Delta$	<b>↑0.6</b>	<b>↑0.7</b>	<b>↑0.02</b>	<b>↑0.01</b>	<b>↑2.2</b>	<b>↑0.04</b>	<b>↑1.7</b>	<b>↑0.06</b>
MM-DFN [13]	76.7	76.7	0.91	0.68	33.3	0.79	42.5	0.41
MM-DFN*	<b>77.1</b>	<b>77.3</b>	<b>0.89</b>	<b>0.71</b>	<b>33.9</b>	<b>0.76</b>	<b>43.1</b>	<b>0.41</b>
$\Delta$	<b>↑0.4</b>	<b>↑0.6</b>	<b>↑0.02</b>	<b>↑0.03</b>	<b>↑0.6</b>	<b>↑0.03</b>	<b>↑0.6</b>	<b>↑0.0</b>
M2FNet [3]	78.9	78.5	0.87	0.73	34.6	0.75	43.4	0.41
M2FNet*	<b>79.6</b>	<b>79.9</b>	<b>0.87</b>	<b>0.75</b>	<b>35.5</b>	<b>0.73</b>	<b>43.6</b>	<b>0.43</b>
$\Delta$	<b>↑0.7</b>	<b>↑1.4</b>	<b>↑0</b>	<b>↑0.02</b>	<b>↑0.9</b>	<b>↑0.02</b>	<b>↑0.2</b>	<b>↑0.02</b>
EmoCaps [23]	80.2	80.3	0.86	0.78	36.7	0.71	43.7	0.43
EmoCaps*	<b>82.7</b>	<b>82.9</b>	<b>0.81</b>	<b>0.89</b>	<b>38.2</b>	<b>0.63</b>	<b>44.9</b>	<b>0.47</b>
$\Delta$	<b>↑2.5</b>	<b>↑2.6</b>	<b>↑0.05</b>	<b>↑0.11</b>	<b>↑1.5</b>	<b>↑0.08</b>	<b>↑1.2</b>	<b>↑0.04</b>

CMU-MOSI and POM datasets also include a training set, a test set, and a validation set. Two popular metrics are chosen to evaluate the performance of each method on the IEMOCAP and MELD datasets, i.e., classification accuracy and  $F1$  score. Five metrics are chosen to evaluate the performance of each method on the CMU-MOSI dataset, i.e., binary classification accuracy (Acc-2), seven classification accuracy (Acc-7), mean absolute error (MAE), Pearson's correlation (Corr), and  $F1$  score. Three metrics are chosen to evaluate the performance of each method on the POM dataset, i.e., classification accuracy, MAE, and Corr.

#### B. Performance Verification Experiment

To verify the effectiveness of our designed LMAM module, we first test our method in a plug-and-play way by directly embedding the LMAM module into seven state-of-the-art DL-based CER methods. The experimental results are shown in Table II, Table III, Table IV. From Table II, Table III, and Table IV, it can be easily seen that all the seven backbones have a significant performance improvement on the four datasets after using our proposed LMAM module. The performance improvement may attribute to the full interaction and fusion of different modal information in our proposed LMAM method, while the seven backbone networks only make a simple fusion of cross-modal information and thus neglect some complementary semantic information between different modals. Besides, we also compare the emotion recognition results of bc-LSTM+Att and bc-LSTM\* (i.e., bc-LSTM+LMAM), and the performance of bc-LSTM\* is significantly better than that of the bc-LSTM+Att, which implies that the proposed LMAM module is better than the self attention module.

Since the above experiment embeds our proposed LMAM module into the backbones, thus it will increase the parameter number of the backbone network. To verify that the performance improvement doesn't come from the increase of model complexity but the reasonable design of our LMAM module, we increase the parameter number of four backbones (i.e., bc-LSTM, MM-DFN, M2FNet, and EmoCaps) to the same as after embedding the LMAM module. The experimental results are shown in Table V. It can be observed from Table V that the performance of both bc-LSTM, MM-DFN, M2FNet, and EmoCaps methods embedded with the LMAM module are

TABLE V

THE RESULTS OF THE EQUAL PARAMETERS EXPERIMENT ON IEMOCAP AND MELD DATASETS. THE PARAMETERS OF METHODS WITH  $\diamond$  ARE INCREMENTED TO BE THE SAME AS METHODS WITH \*. THE BEST RESULT IN EACH COLUMN IS IN BOLD.

Method	Params	IEMOCAP		MELD	
		Acc.	F1	Acc.	F1
bc-LSTM [31]	0.53M	55.19	54.96	54.18	55.89
bc-LSTM $\diamond$	1.15M	54.37	52.19	53.11	50.93
bc-LSTM*	1.15M	<b>65.55</b>	<b>60.29</b>	<b>54.97</b>	<b>56.85</b>
MM-DFN [13]	2.21M	68.77	68.20	60.30	59.44
MM-DFN $\diamond$	2.83M	67.33	67.02	60.01	57.65
MM-DFN*	2.83M	<b>69.94</b>	<b>69.69</b>	<b>60.65</b>	<b>59.62</b>
M2FNet [3]	8.47M	69.11	69.86	63.64	60.87
M2FNet $\diamond$	9.09M	69.46	68.79	62.49	60.91
M2FNet*	9.09M	<b>70.31</b>	<b>70.07</b>	<b>64.13</b>	<b>60.97</b>
EmoCaps [23]	12.74M	71.22	71.06	63.52	62.97
EmoCaps $\diamond$	13.36M	71.18	70.56	62.17	62.11
EmoCaps*	13.36M	<b>73.67</b>	<b>73.01</b>	<b>64.93</b>	<b>63.88</b>

TABLE VI

THE RESULTS OF THE PARAMETER SIZE AND RUNNING TIME EXPERIMENT ON IEMOCAP, MELD, CMU-MOSI, AND POM DATASETS. THE PARAMETERS OF METHODS WITH  $\diamond$  ARE INCREMENTED TO BE THE SAME AS METHODS WITH \*.

Methods	Params	IEMOCAP		MELD		CMU-MOSI		POM	
		Running Time	IEMOCAP	Running Time	MELD	Running Time	CMU-MOSI	Running Time	POM
TextCNN [18]	0.23M	0.68s		1.15s		0.25s		0.11s	
TextCNN*	0.85M	1.65s		2.34s		0.32s		0.14s	
bc-LSTM [31]	0.53M	1.21s		2.17s		0.27s		0.19s	
bc-LSTM+Att	4.7M	4.13s		6.63s		0.83s		0.44s	
bc-LSTM*	1.15M	2.18s		2.89s		0.46s		0.25s	
DialogueRNN [28]	14.47M	47.96s		76.92s		12.56s		7.73s	
DialogueRNN*	15.09M	48.54s		78.43s		13.17s		8.42s	
DialogueGCN [7]	12.92M	56.18s		82.34s		17.48s		9.32s	
DialogueGCN*	13.54M	57.77s		88.53s		18.23s		9.74s	
MM-DFN [13]	2.21M	3.15s		6.17s		1.41s		0.71s	
MM-DFN*	2.83M	3.89s		7.04s		1.83s		0.79s	
M2FNet [3]	8.47M	12.48s		25.16s		3.39s		1.15s	
M2FNet*	9.09M	13.11s		28.33s		3.77s		1.82s	
EmoCaps [23]	12.74M	16.73s		31.07s		4.82s		2.56s	
EmoCaps*	13.36M	17.54s		34.35s		5.14s		3.06s	

better than the bc-LSTM, MM-DFN, M2FNet, and EmoCaps models with the same parameter number, which proves that the performance improvement is not due to the increase of parameter number but is brought by our LMAM module.

### C. Parameter Size and Runtime Analysis

To verify that the proposed LMAM is a lightweight plug-and-play module, we analyze the changes in parameter size and running time after adding the LMAM method to the baselines on different emotion recognition datasets. In particular, we perform all experiments under the same experimental configuration to ensure the fairness of the experiments. As shown in Table VI, we observe that after adding our proposed LMAM module to the baselines, the running time and model parameter size only increase slightly, but the performance of emotion recognition can be greatly improved. The experimental results show that through appropriate model optimization or architecture improvement, the computational overhead can be effectively controlled while improving the model performance. The performance improvement is attributed to the fact that the proposed LMAM can more effectively realize feature fusion, so that the model can better capture and utilize the complementary information between different modalities, thereby improving the accuracy of emotion recognition. At the same time, the increase in running time and parameter size is small, indicating that the design of the model successfully avoids excessive growth in resource requirements while retaining complexity and depth.

### D. Statistical Significance

In this section, to further verify the significance level between the LMAM method proposed in this paper and the baseline method, we conducted a paired t-test to ensure the reliability of our experimental conclusions. The significance level results are shown in Tables VIII and VII. We run the baseline model and the baseline model with the LMAM module 5 times and obtained the corresponding F1 value as

an evaluation metric of significance statistics. All emotion categories on IEMOCAP and MELD datasets are statistically significant in F1 values under paired t-test ( $p < 0.05$ ).

### E. Comparison with Other Cross-modal Fusion Methods

In this section, to further verify the superiority of our proposed LMAM module, we also conduct an experiment to compare our LMAM module with other four typical cross-modal fusion approaches, i.e., classical add operation and concatenate operation, and latest low-rank multi-modal fusion (LFM) [26], tensor fusion network (TFN) [45], Dual Low-Rank Multimodal Fusion (Dual-LMF) [17], Low-Rank Multimodal Fusion with Self-Attention (Att-LMF) [50], Low-Rank Multimodal Fusion with multi-modal Transformer (LMF-MuLT) [34], AuxFormer [9], Cross-modality Context fusion and Semantic Refinement Network (CMCF-SRNet) [48], and Low-Rank Adaption (LoRA) [15]. The selected backbone network is EmoCaps [23] and the used datasets are also IEMOCAP and MELD.

TABLE IX

THE COMPARISON RESULTS OF OUR LMAM FUSION METHOD WITH OTHER CROSS-MODAL METHODS. THE BEST RESULT IS HIGHLIGHTED IN BOLD.  $\diamond$  REPRESENTS RANDOMLY INITIALIZED PRE-TRAINING WEIGHTS, AND \* USING THE PRE-TRAINED WEIGHTS OF EMOCAPS.

Methods	IEMOCAP		MELD	
	Acc.	F1	Acc.	F1
Add [24]	71.3	71.0	64.3	64.0
Concatenate [2]	68.9	68.1	62.6	61.1
TFN [45]	70.8	70.1	63.0	62.8
LFM [26]	72.3	71.7	64.1	62.3
Dual-LMF [17]	72.6	72.0	64.8	63.2
Att-LMF [50]	72.5	71.3	64.3	62.9
LMF-MuLT [34]	72.5	71.9	64.6	62.5
AuxFormer [9]	71.6	72.1	63.9	63.0
CMCF-SRNet [48]	72.3	71.4	62.3	63.3
LoRA $\diamond$ [15]	60.7	58.3	55.6	57.5
LoRA* [15]	73.3	73.0	64.4	63.5
<b>LMAM (Ours)</b>	<b>73.7</b>	<b>73.0</b>	<b>64.9</b>	<b>63.9</b>

TABLE VII

STATISTICALLY SIGNIFICANT RESULTS ON IEMOCAP DATASET. WE REPORT P-VALUES FOR EACH EMOTION CATEGORY UNDER PAIRED T-TESTS.

Methods	IEMOCAP					
	Happy	Sad	Neutral	Angry	Excited	Frustrated
TextCNN [18]	$9.8 \times 10^{-10}$	$6.0 \times 10^{-10}$	$6.1 \times 10^{-7}$	$1.1 \times 10^{-6}$	$4.8 \times 10^{-7}$	$4.2 \times 10^{-6}$
bc-LSTM [31]	$4.8 \times 10^{-3}$	$6.7 \times 10^{-4}$	$3.1 \times 10^{-4}$	$5.6 \times 10^{-5}$	$1.7 \times 10^{-3}$	$3.9 \times 10^{-3}$
bc-LSTM+Att	$4.3 \times 10^{-5}$	$8.9 \times 10^{-8}$	$5.5 \times 10^{-6}$	$1.2 \times 10^{-8}$	$5.9 \times 10^{-7}$	$3.2 \times 10^{-3}$
DialogueRNN [28]	$4.9 \times 10^{-8}$	$3.4 \times 10^{-2}$	$2.9 \times 10^{-5}$	$2.8 \times 10^{-3}$	$6.4 \times 10^{-8}$	$1.7 \times 10^{-9}$
DialogueGCN [7]	$3.9 \times 10^{-7}$	$5.2 \times 10^{-8}$	$5.1 \times 10^{-7}$	$1.9 \times 10^{-6}$	$7.3 \times 10^{-3}$	$9.4 \times 10^{-8}$
MM-DFN [13]	$1.0 \times 10^{-8}$	$4.5 \times 10^{-5}$	$5.8 \times 10^{-5}$	$1.1 \times 10^{-5}$	$6.8 \times 10^{-7}$	$4.5 \times 10^{-7}$
M2FNet [3]	$9.3 \times 10^{-7}$	$3.5 \times 10^{-3}$	$1.7 \times 10^{-4}$	$4.1 \times 10^{-4}$	$5.8 \times 10^{-7}$	$1.1 \times 10^{-8}$
Emocaps [23]	$5.1 \times 10^{-4}$	$4.3 \times 10^{-5}$	$2.1 \times 10^{-4}$	$7.2 \times 10^{-6}$	$3.5 \times 10^{-8}$	$2.6 \times 10^{-7}$

TABLE VIII

STATISTICALLY SIGNIFICANT RESULTS ON MELD DATASET. WE REPORT P-VALUES FOR EACH EMOTION CATEGORY UNDER PAIRED T-TESTS.

Methods	MELD						
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger
TextCNN [18]	$1.1 \times 10^{-8}$	$2.9 \times 10^{-3}$	$1.8 \times 10^{-6}$	$1.4 \times 10^{-2}$	$7.6 \times 10^{-5}$	$2.1 \times 10^{-9}$	$8.4 \times 10^{-9}$
bc-LSTM [31]	$2.3 \times 10^{-4}$	$4.3 \times 10^{-3}$	$2.5 \times 10^{-7}$	$4.6 \times 10^{-10}$	$7.8 \times 10^{-4}$	$1.1 \times 10^{-7}$	$5.6 \times 10^{-4}$
bc-LSTM+Att	$2.4 \times 10^{-4}$	$5.1 \times 10^{-3}$	$2.5 \times 10^{-7}$	$1.9 \times 10^{-3}$	$6.6 \times 10^{-4}$	$2.7 \times 10^{-5}$	$8.9 \times 10^{-4}$
DialogueRNN [28]	$3.7 \times 10^{-4}$	$4.9 \times 10^{-3}$	$2.2 \times 10^{-3}$	$8.7 \times 10^{-8}$	$5.4 \times 10^{-5}$	$2.8 \times 10^{-3}$	$1.5 \times 10^{-3}$
DialogueGCN [7]	$1.3 \times 10^{-2}$	$3.4 \times 10^{-4}$	$1.7 \times 10^{-3}$	$6.7 \times 10^{-6}$	$3.7 \times 10^{-3}$	$1.2 \times 10^{-10}$	$5.2 \times 10^{-5}$
MM-DFN [13]	$3.7 \times 10^{-3}$	$7.2 \times 10^{-8}$	$2.5 \times 10^{-6}$	$8.9 \times 10^{-10}$	$6.6 \times 10^{-6}$	$9.1 \times 10^{-9}$	$4.7 \times 10^{-3}$
M2FNet [3]	$1.4 \times 10^{-2}$	$6.9 \times 10^{-4}$	$7.6 \times 10^{-10}$	$3.0 \times 10^{-3}$	$2.9 \times 10^{-2}$	$9.2 \times 10^{-10}$	$6.1 \times 10^{-4}$
Emocaps [23]	$6.9 \times 10^{-5}$	$1.7 \times 10^{-3}$	$4.4 \times 10^{-3}$	$4.7 \times 10^{-2}$	$8.4 \times 10^{-7}$	$3.9 \times 10^{-2}$	$5.7 \times 10^{-5}$

The experimental results are recorded in Table IX. As shown in the Table IX, the LMAM method achieves the best experimental results on the IEMOCAP and MELD datasets, with Acc of 73.0% and 65.4%, respectively, and F1 values of 73.0% and 64.9%, respectively. Specifically, compared with the Add method, the Acc and F1 values of the LMAM method on the IEMOCAP dataset are increased by 1.7% and 2.0%, respectively, and the Acc and F1 values on the MELD dataset are increased by 1.1% and 0.9%, respectively. Compared with the Concatenate method, the Acc and F1 values of the LMAM method on the IEMOCAP dataset are increased by 4.1% and 4.9%, respectively, and the Acc and F1 values on the MELD dataset are increased by 2.8% and 3.8%, respectively. We think this is because the Add method and the Concatenate method do not model complementary semantic information within and between modalities. Additionally, compared with the TFN, LFM, Dual-LMF, Att-LMF, LMF-MuLT, AuxFormer and CMCF-SRNet methods, the LMAM method has also achieved better performance in the accuracy and F1 value of emotion recognition, which further illustrates the superiority of our designed LMAM fusion method. Moreover, to further illustrate the effectiveness of our proposed LMAM method, we also compare it with LoRA. Specifically, we divide LoRA into two versions, one with randomly initialized pre-trained weights and the other with pre-trained weights using EmoCaps. Experimental results show that LoRA using EmoCaps' pre-trained weights performs similarly to our proposed LMAM method, but the performance of LoRA is very poor when the pre-trained weights are randomly initialized. We believe this is because the pre-trained weights of LoRA are frozen, so it is very dependent on the prior information of the pre-trained weights. Although the LMAM method we proposed

randomly initializes the weights, since our method is end-to-end and the weights is learnable, it can learn weights with better generalization performance.

TABLE X  
EXPERIMENTAL RESULTS OF USING SINGLE-MODAL DATA AND MULTI-MODAL DATA ON IEMOCAP AND MELD DATASETS. T, A, AND V REPRESENT TEXT, AUDIO, AND VIDEO, RESPECTIVELY.

Modality	IEMOCAP		MELD	
	Acc.	F1	Acc.	F1
T	70.8	69.2	63.1	61.8
A	66.0	65.3	60.4	59.2
V	58.8	56.8	53.9	50.7
T+A	72.3	71.5	64.4	62.7
T+V	70.8	69.6	63.5	62.1
V+A	66.7	65.8	61.0	59.9
T+A+V	<b>73.7</b>	<b>73.0</b>	<b>64.9</b>	<b>63.9</b>

### F. Ablation study

1) *Necessity of multi-modal data:* To illustrate the necessity of multi-modal research, we used the EmoCaps method equipped with LMAM module as the backbone to conduct a comparative experiment of uni-modal, bi-modal and multi-modal on the IEMOCAP and MELD datasets. The experimental results are shown in Table X. We conducted a uni-modal experiment to utilize only one of the three modalities (i.e., text, video, and audio), a bi-modal experiment to use any two modalities and a multi-modal experiment to use all the three modalities. For the uni-modal experiments, we found that the features of the text modality performed best for emotion recognition on both datasets, followed by the

features of the audio modality, and the worst performance of the features of the video modality. In the bi-modal experiment, the emotion recognition effect of text + audio is the best, followed by the emotion recognition of text + video, and the emotion recognition of video + audio is the worst. In the multimodal experiment, we can find that the emotion recognition effect of the combination of the three modalities is the best. Experimental results demonstrate that it is necessary to consider the multi-modal study. Furthermore, designing multi-modal feature fusion methods to improve the effect of emotion recognition is also necessary.

2) *Comparison of different embedding ways:* To compare the performance of the early fusion and early fusion with residual connections embedding ways introduced in Section 4.2, we conduct comparative experiments using the bc-LSTM, MMDFN, M2FFNet, and EmoCaps algorithms on the IEMOCAP and MELD datasets. The experimental results are shown in Table XI. As can be seen, the LMAM module with residual connections can obtain better performance compared with the early fusion without residual connection.

TABLE XI

METHODS WITH \* REPRESENT THE METHOD EQUIPPED WITH OUR LMAM MODULE WITHOUT ANY FURTHER CHANGES. METHODS WITH \*(R) REPRESENT THE METHOD EQUIPPED WITH OUR LMAM MODULE WITH RESIDUAL CONNECTIONS. THE BEST RESULT IS HIGHLIGHTED IN BOLD.

Methods	IEMOCAP		MELD	
	Acc.	F1	Acc.	F1
bc-LSTM*	59.49	59.16	57.69	55.47
bc-LSTM*(R)	<b>61.77</b>	<b>60.49</b>	<b>59.56</b>	<b>57.49</b>
MM-DFN*	67.93	67.15	63.28	61.12
MM-DFN*(R)	<b>69.82</b>	<b>69.68</b>	<b>68.02</b>	<b>65.24</b>
M2FFNet*	68.35	57.96	67.47	66.59
M2FFNet*(R)	<b>70.27</b>	<b>70.07</b>	<b>68.34</b>	<b>67.25</b>
EmoCaps*	71.49	71.01	65.03	64.22
EmoCaps*(R)	<b>73.67</b>	<b>73.01</b>	<b>64.93</b>	<b>63.88</b>

### G. Complexity Analysis

We assume that the query vectors  $W^Q$ , key vector  $W^K$ , and value vector  $W^V$  in the self-attention mechanism have the same dimensions as the input multimodal feature  $d_n$ . Theoretically, the computational complexity of the LMAM method proposed in this paper is  $O\left(\sum_{n=1}^M \sum_{i=1}^{r(d_n)} (d_n^{(i)})^3\right)$  compared to  $O\left(\sum_{n=1}^M (d_n)^3\right)$  of self-attention model. Furthermore, we evaluate the computational complexity and computation time of LMAM and self-attention mechanism. As shown in the Table XII, the training time and parameter of the LMAM method are much smaller than the self-attention mechanism. Here, we set the rank size to 45.

### H. Rank Settings

We verified the impact of different rank parameter settings on the accuracy of emotion recognition on the IEMOCAP dataset. The experimental results are shown in Fig. 5. We observe that when  $rank = 45$ , the training effect of the model is the best, and the training effect of the model is stable

TABLE XII  
COMPARISON THE TRAINING SPEED AND PARAMETERS OF SELF-ATTENTION MECHANISM AND LMAM METHOD IN AN EPOCH. EMOCAPS IS SELECTED AS OUR ARCHITECTURE.

Methods	Training time (s)	Parameters (M)
LMAM	17.8	0.62
Self-attention [41]	58.7	3.42

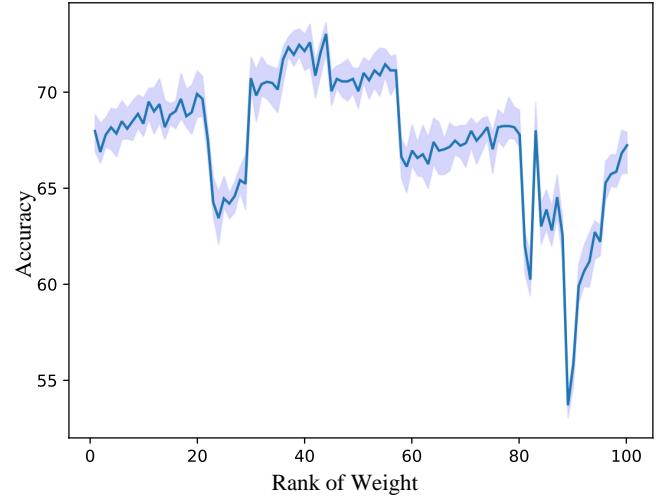


Fig. 5. The impact of different rank parameter settings on the experimental accuracy. When the rank exceeds 55, the training results of the model start to be unstable.

when the rank is between 30 and 55. When  $rank > 45$ , the training result of the model becomes unstable and the effect is poor. Therefore, better experimental results can be obtained by using lower ranks. Therefore, better experimental results can be obtained by using lower ranks.

### I. Potential Applications

In order to verify that our LMAM method has a potential application in other multi-modal recognition tasks, we further apply the proposed LMAM module to session recommendation and humor detection tasks. Specifically, we embed our LMAM method into dual channel hypergraph convolutional network (DHCN) [42] for session recommendation task and Contextual Memory Fusion Network (C-MFN) [46] for humor detection task, respectively. The session recommendation task is conducted on the Digitetica dataset<sup>2</sup>, and the humor detection task is carried out on the UR-FUNNY [10] dataset. The experimental results are illustrated in Table XIII and Table XIV. As shown in Table XIII and Table XIV, we can observe that our proposed LMAM module can improve the performance of the backbone networks in other multi-modal recognition tasks.

## VI. CONCLUSION

In this paper, we propose a novel cross-modal feature fusion method to enable better cross-modal feature fusion. To capture the complementary emotional contextual semantic information

<sup>2</sup><http://cikm2016.cs.iupui.edu/cikm-cup/>

TABLE XIII

EXPERIMENTAL RESULTS OF DHCN METHOD ON THE DIGETICA DATASET FOR THE SESSION RECOMMENDATION TASK. WE USE P@K (PRECISION) AND MRR@K (MEAN RECIPROCAL RANK) TO EVALUATE THE RECOMMENDATION RESULTS. \* MEANS THE METHOD EQUIPPED WITH THE LMAM MODULE. THE BEST RESULT IS HIGHLIGHTED IN BOLD.

Method	Digetica					
	P@5	P@10	P@20	MRR@5	MRR@10	MRR@20
DHCN [42]	27.24	39.87	53.18	15.28	17.53	18.44
DHCN*	<b>28.19</b>	<b>40.38</b>	<b>53.70</b>	<b>16.43</b>	<b>17.90</b>	<b>18.69</b>

TABLE XIV

EXPERIMENTAL RESULTS OF C-MFN METHOD ON THE UR-FUNNY DATASET FOR THE HUMOR DETECTION TASK. C-MFN (C) MEANS USING ONLY CONTEXTUAL INFORMATION WITHOUT PUNCHLINES. C-MFN (P) MEANS USING ONLY PUNCHLINES WITH NO CONTEXTUAL INFORMATION. \* MEANS THE METHOD EQUIPPED WITH THE LMAM MODULE. THE BEST RESULT IS HIGHLIGHTED IN BOLD.

UR-FUNNY					
Modality	T	A+V	T+A	T+V	T+A+V
C-MFN(P) [10]	62.85	53.30	63.28	63.22	64.47
C-MFN(C) [10]	57.96	50.23	57.78	57.99	58.45
C-MFN [10]	64.44	57.99	64.47	64.22	65.23
C-MFN(P)*	<b>65.43</b>	<b>59.36</b>	<b>66.04</b>	<b>66.59</b>	<b>66.87</b>
C-MFN(C)*	<b>59.46</b>	<b>53.69</b>	<b>58.39</b>	<b>58.68</b>	<b>59.23</b>
C-MFN*	<b>65.66</b>	<b>59.34</b>	<b>65.68</b>	<b>64.97</b>	<b>67.00</b>

in different modalities, we utilize a low-rank matching attention mechanism (LMAM) to realize the interaction between multimodal features and use low-rank weights to improve efficiency. LMAM is better than the existing fusion methods while has a lower complexity. Extensive experimental results verify that LMAM can be embedded into any existing DL-based CER methods to improve their performance in a plug-and-play manner. We also mathematically prove the effectiveness of our method. Further, LMAM is a general cross-modal feature fusion method and has potential application value in other multi-modal recognition tasks, e.g., session recommendation and humor detection.

Although the LMAM method can achieve better performance improvement and increase the running speed of the model, the introduction of low-rank decomposition technology may cause the model to lose contextual semantic information. Therefore, in future research work, we will study how to achieve efficient fusion of multi-modal features without losing contextual semantic information.

## REFERENCES

- [1] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [2] E. Cambria, D. Hazarika, S. Poria, A. Hussain, and R. Subramanyam, “Benchmarking multimodal sentiment analysis,” in *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLING 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18*. Springer, 2018, pp. 166–179.
- [3] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, “M2fnet: multi-modal fusion network for emotion recognition in conversation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4652–4661.
- [4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*.
- [5] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 2010, pp. 1459–1462.
- [6] A. Geetha, T. Mala, D. Priyanka, and E. Uma, “Multimodal emotion recognition with deep learning: advancements, challenges, and future directions,” *Information Fusion*, vol. 105, p. 102218, 2024.
- [7] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, “Dialoguegen: A graph convolutional neural network for emotion recognition in conversations,” 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)*. ACL, 2019, pp. 154–164.
- [8] S. Ghosh, G. V. Singh, A. Ekbal, and P. Bhattacharyya, “Comma-deer: Common-sense aware multimodal multitask approach for detection of emotion and emotional reasoning in conversations,” in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 6978–6990.
- [9] L. Goncalves and C. Busso, “Auxformer: Robust approach to audiovisual emotion recognition,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7357–7361.
- [10] M. K. Hasan, W. Rahman, A. B. Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency, and M. E. Hoque, “Ur-funny: A multimodal language dataset for understanding humor,” 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)*, 2019, pp. 2046–2056.
- [11] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, “Icon: Interactive conversational memory network for multimodal emotion detection,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. ACL, 2018, pp. 2594–2604.
- [12] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, “Conversational memory network for emotion recognition in dyadic dialogue videos,” in *Proceedings of the conference. Association for Computational Linguistics.*, vol. 2018, 2018, p. 2122.
- [13] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, “Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7037–7041.
- [14] D. Hu, L. Wei, and X. Huai, “Dialoguercn: Contextual reasoning networks for emotion recognition in conversations,” 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)*, 2021, pp. 7042–7052.
- [15] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, “Lora: Low-rank adaptation of large language models,” in *International Conference on Learning Representations*.
- [16] J. Hu, Y. Liu, J. Zhao, and Q. Jin, “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)*, 2021, pp. 5666–5675.
- [17] T. Jin, S. Huang, Y. Li, and Z. Zhang, “Dual low-rank multimodal fusion,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 377–387.
- [18] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2014, pp. 1746–1751.
- [19] T. G. Kolda and B. W. Bader, “Tensor decompositions and applications,” *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.
- [20] J. Li, X. Wang, Y. Liu, and Z. Zeng, “Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition,” *IEEE Transactions on Affective Computing*, 2024.
- [21] J. Li, X. Wang, G. Lv, and Z. Zeng, “Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation,” *Neurocomputing*, vol. 550, p. 126427, 2023.
- [22] S. Li, H. Yan, and X. Qiu, “Contrast and generation make bart a good dialogue emotion recognizer,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11002–11010.
- [23] Z. Li, F. Tang, M. Zhao, and Y. Zhu, “Emocaps: Emotion capsule based model for conversational emotion recognition,” in *Findings of*

- the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1610–1618.
- [24] Z. Lian, B. Liu, and J. Tao, “Ctnet: Conversational transformer network for emotion recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.
- [25] H. Liu, K. Li, J. Fan, C. Yan, T. Qin, and Q. Zheng, “Social image–text sentiment classification with cross-modal consistency and knowledge distillation,” *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3332–3344, 2022.
- [26] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, “Efficient low-rank multimodal fusion with modality-specific factors,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, 2018, pp. 2247–2256.
- [27] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, “A transformer-based model with self-distillation for multimodal emotion recognition in conversations,” *IEEE Transactions on Multimedia*, pp. 1–13, 2023.
- [28] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, “Dialoguernn: An attentive rnn for emotion detection in conversations,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01. AAAI, 2019, pp. 6818–6825.
- [29] S. M. Mohammad, “Ethics sheet for automatic emotion recognition and sentiment analysis,” *Computational Linguistics*, vol. 48, no. 2, pp. 239–278, 2022.
- [30] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, “Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach,” in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 50–57.
- [31] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, “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)*. ACL, 2017, pp. 873–883.
- [32] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “Meld: A multimodal multi-party dataset for emotion recognition in conversations,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL, 2019, pp. 527–536.
- [33] M. Ren, X. Huang, W. Li, D. Song, and W. Nie, “Lr-gcn: Latent relation-aware graph convolutional network for conversational emotion recognition,” *IEEE Transactions on Multimedia*, pp. 1–1, 2021.
- [34] S. Sahay, E. Okur, S. H. Kumar, and L. Nachman, “Low rank fusion based transformers for multimodal sequences,” in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 2020, pp. 29–34.
- [35] C. Schuster and S. Hegelich, “From bert’s point of view: Revealing the prevailing contextual differences,” in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1120–1138.
- [36] I. Shahin, N. Hindawi, A. B. Nassif, A. Alhudhaif, and K. Polat, “Novel dual-channel long short-term memory compressed capsule networks for emotion recognition,” *Expert Systems with Applications*, vol. 188, p. 116080, 2022.
- [37] D. Sheng, D. Wang, Y. Shen, H. Zheng, and H. Liu, “Summarize before aggregate: a global-to-local heterogeneous graph inference network for conversational emotion recognition,” in *Proceedings of the 28th International Conference on Computational Linguistics*. ICCL, 2020, pp. 4153–4163.
- [38] Y. Shou, T. Meng, W. Ai, S. Yang, and K. Li, “Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis,” *Neurocomputing*, vol. 501, pp. 629–639, 2022.
- [39] Y. Su, Y. Wei, W. Nie, S. Zhao, and A. Liu, “Dynamic causal disentanglement model for dialogue emotion detection,” *IEEE Transactions on Affective Computing*, 2024.
- [40] G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, and D. Jiang, “Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models,” *Knowledge-Based Systems*, vol. 235, p. 107598, 2022.
- [41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [42] X. Xia, H. Yin, J. Yu, Q. Wang, L. Cui, and X. Zhang, “Self-supervised hypergraph convolutional networks for session-based recommendation,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 5, 2021, pp. 4503–4511.
- [43] S. Xing, S. Mai, and H. Hu, “Adapted dynamic memory network for emotion recognition in conversation,” *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1426–1439, 2022.
- [44] L. Yang, Y. Shen, Y. Mao, and L. Cai, “Hybrid curriculum learning for emotion recognition in conversation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11595–11603.
- [45] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.
- [46] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, “Memory fusion network for multi-view sequential learning,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [47] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos,” *arXiv preprint arXiv:1606.06259*, 2016.
- [48] X. Zhang and Y. Li, “A cross-modality context fusion and semantic refinement network for emotion recognition in conversation,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 13099–13110.
- [49] Y. Zhang, H. Liu, D. Wang, D. Zhang, T. Lou, Q. Zheng, and C. Quek, “Cross-modal credibility modelling for eeg-based multimodal emotion recognition,” *Journal of Neural Engineering*, vol. 21, no. 2, p. 026040, 2024.
- [50] H. Zhu, Z. Wang, Y. Shi, Y. Hua, G. Xu, and L. Deng, “Multimodal fusion method based on self-attention mechanism,” *Wireless Communications and Mobile Computing*, vol. 2020, pp. 1–8, 2020.



**Yuntao Shou** is currently pursuing the M.S. degree in the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an. His research interests include graph learning and emotion recognition. (Email: shouyuntao@stu.xjtu.edu.cn)



**Huan Liu** received the B.Sc. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2013 and 2020, respectively. He is currently an Associate Professor with the School of Computer Science and Technology, Xi'an Jiaotong University. His research interests include affective computing, machine learning, and deep learning. (Email: huanliu@xjtu.edu.cn)



**Xiangyong Cao** received the B.Sc. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2012 and 2018, respectively. From 2016 to 2017, he was a Visiting Scholar with Columbia University, New York, NY, USA. He is currently an Associate Professor with the School of Computer Science and Technology, Xi'an Jiaotong University. His research interests include statistical modeling and image processing. (Email: caoxiangyong@mail.xjtu.edu.cn)



**Deyu Meng** received the B.Sc., M.Sc., and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2001, 2004, and 2008, respectively. From 2012 to 2014, he took his two-year sabbatical leave with Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Professor with the School of Mathematics and Statistics, Xi'an Jiaotong University, and an Adjunct Professor with the Faculty of Information Technology, The Macau University of Science and Technology, Macau, China. His research interests include model-based deep learning, variational networks, and meta-learning (Email: dymeng@mail.xjtu.edu.cn)



**Bo Dong** received the Ph.D. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2014. From 2014 to 2017, he was Postdoc Researcher of control science and engineering, Xi'an Jiaotong University, where he is currently a Professor with the School of Distance Education. His main research interests are data mining, and pattern recognition. (Email: dong.bo@mail.xjtu.edu.cn)