Journal of Intelligent Information Systems
https://doi.org/10.1007/s10844-025-00975-z

RESEARCH

Towards robust multimodal emotion recognition in conversation with multi-modal transformer and variational distillation fusion

Xiaofei Zhu[1] · **Shuming Jiang**[1]

Received: 25 May 2025 / Revised: 23 July 2025 / Accepted: 28 July 2025

Abstract
Multimodal Emotion Recognition in Conversation (MERC) utilizes multimodal information such as language, visual, and audio to enhance the understanding of human emotions. Current multimodal interaction frameworks inadequately resolve inherent information conflicts and redundancy due to their assumption of equivalent quality across heterogeneous modalities. In addition, inappropriate evaluation of the importance of modalities can also cause this problem. To address this issue, we introduce a Language-Focused Augmented Transformer with Variational Distillation Fusion network called LFVD. In contrast to previous work, we suggest focusing on language modality through the Language-Focused Augmented Transformer, which extracts task-relevant signals from visual and audio modalities to help us understand language. Concurrently, this architecture derives conversational emotional atmosphere representation to refine multimodal integration, thereby mitigating the influence of redundant and conflicting information. Furthermore, Variational Distillation Fusion has been proposed in which multimodal representations are probabilistically encoded as variational distributions over Gaussian manifolds rather than deterministic embeddings. Subsequently, the importance of each modality is estimated automatically based on distribution differences. Experiments on the IEMOCAP and MELD datasets show that our proposed model outperforms previous state-of-the-art baseline models.

Keywords Emotion recognition in conversation · Distillation learning · Multimodal information fusion

✉ Xiaofei Zhu
zxf@cqut.edu.cn

Shuming Jiang
jsm@stu.cqut.edu.cn

[1] College of Computer Science and Engineering, Chongqing University of Technology, Hongguang Avenue, Chongqing 400054, Chongqing, China

Published online: 04 September 2025

1 Introduction

Emotion Recognition in Conversation (ERC) aims to correctly recognize the emotions expressed by each speaker in a conversation. Recently, ERC has attracted a great deal of interest because of its valuable applications in recommendation systems (Zheng et al., 2022), healthcare services (Pujol et al., 2019), and affective computing systems (Czerwinski et al., 2016). Traditional ERC identifies the emotion of an utterance by analyzing the language in a conversation. However, language information alone cannot fully capture the emotional nuances. Emotions can be conveyed not only through words but also through the speaker's facial expressions, and voice intonation. Therefore, MERC which integrates audio and video information in addition to unimodal ERC has attracted more research interest.

Since conversations contain rich multimodal contextual information, how to effectively utilize the information of each modality is an important issue. Recent approaches devoted to the modeling of multimodal information can be divided into two categories: graph-based methods (Hu et al., 2022; Tu et al., 2024b; Wu et al., 2025; Gan et al., 2025; Wang et al., 2025; Ai et al., 2025) and transformer-based methods (Zhong et al., 2019; Zou et al., 2023; Liu et al., 2025). For graph-based approaches, they treat the utterances in the conversation as nodes and construct connecting edges from the relationships that exist between different utterances (whether it is the same speaker or the same emotion), and subsequently, the current node collects information about surrounding utterances. For transformer-based methods, they mainly utilize cross-modality transformers to capture intra- and inter-modal interactions. And they often use language modalities to augment other modalities. We argue that simply using the language modality to augment the other two modalities does not do a good job of improving the quality of these two modalities or even undermining the cues that are otherwise present in them that are favorable for emotion recognition.

Existing methods have shown their advantages in interacting and fusing multimodal information, but most of them treat different modalities as having the same quality, especially these graph-based methods, such as MMGCN (Hu et al., 2021) and GS-MCC (Ai et al., 2025). They form an undirected graph of all modalities, equally connecting each modality so as to equally fuse the information of each modality in subsequent operations such as convolution, and we consider such a connection as having the same quality for each modality. Such a symmetric fusion approach usually suffers from the influence of redundant and conflicting information. Whereas, the fact is that existing work and our ablation study (see Table 4) show that language, visual, and audio modalities contribute differently to the overall prediction performance (Pham et al., 2019; Lei et al., 2023). Therefore, we propose that the language modality serves as the dominant modality, which is of high quality and contains more information relevant to the MERC task, while the auxiliary modality, namely audio and visual, is of relatively low quality and inevitably contains information irrelevant to the task. In multimodal information fusion, we hope that the auxiliary modality can be leveraged to complement the dominant modality in the task and we want to avoid introducing information irrelevant to the MERC task into the final fusion representation.

To this end, we propose a novel Language-Focused AugmentedTransformer withVariational Distillation Fusion for solving the above problem, named LFVD. Our model consists of three main components: Feature Representation, Language-Focused Augmented Transformer (LFA), and Cross-Modalityity Variational Distillation Fusion (CMVD). The Feature Representation is used to extract the representations of each modality. LFA differs

from previous symmetric multimodal information interaction strategies in that it focuses on language. Specifically, inspired by Wang et al. (2024) we design LFA as a scheme to further enhance the language modality, and follow the approach of transferring auxiliary information to the dominant modality in a targeted manner, interacting the information of each modality. Then, the emotional atmosphere representations of the conversations therein are extracted and dual contrast learning is used to enhance the features of each modality. In CMVD, we introduce variational distillation learning and cross-modality fusion. Specifically, the variational module obtains a more robust modal representation capability by transforming feature representations into distributional representations. The variational distillation module uses the dominant modality as a teacher and the two auxiliary modalities as students. Subsequently, when performing cross-modality fusion, evaluate the distribution differences to represent the importance of the information of each modality. In summary, the main contributions of this paper are as follows:

1. We propose a novel LFA that enhances each modality by focusing on the language modality and extracting the conversational atmosphere representation, thereby reducing the influence of redundant and conflicting information during the multimodal interaction process.
2. We develop CMVD to learn a robust representation by encoding multimodal representations as variational distributions, which are subsequently used for distillation learning and cross-modality fusion, thereby improving the quality of multimodal representations.
3. Extensive experiments on two public benchmark multimodal datasets, including IEMOCAP and MELD, show that our proposed LFVD outperforms all state-of-the-art baseline models.

2 Related work

Previous work (Majumder et al., 2019; Ghosal et al., 2019) on ERC has focused on unimodal, i.e., language modality, and they have focused primarily on language modality for emotion recognition. DialogueRNN (Majumder et al., 2019) employs three GRUs, the global GRU, the party GRU, and the emotion GRU, to model speaker, contextual, and emotion information. Based on DialogueRNN, DialogueGCN (Ghosal et al., 2019) also introduces Graph Convolutional Networks (GCN) to model intra- and inter-speaker dependencies with the advantage of graph structure to enhance the propagation of contextual information.

However, when dealing with more complex emotional scenarios, there are limitations of unimodal approaches compared to multimodal. Unimodality cannot provide enough information for emotion recognition, so recently there has been a keen interest in investigating information such as facial expressions, audio information, to obtain more effective multimodal representations. How to capture the rich interaction information between each modality is crucial to improve the accuracy of emotion recognition. And due to the recent deepening of multimodal learning research (Ma et al., 2024; Song et al., 2024; Wang et al., 2023), it provides us with more perspectives to think about. Ma et al. (2024) proposed a Transformer-based self-distillation model SDT for multimodal conversation sentiment recognition that effectively models interactions within and outside the modality and improves modal representation. Zou et al. (2022) proposed the Master Modal Transformer (MMTr)

method, which effectively improves the fusion effect of multimodal conversation emotion recognition by introducing the concept of master modality and enhancing the inter-modal interactions using multi-head attention. Tsai et al. (2019) introduced the multimodal transformer, which combines the basic modules of the transformer fusion method with a multi-head attention mechanism to achieve Cross-Modality information fusion by using different modalities as query, value, and key. Li et al. (2023) proposes a multimodal emotion recognition method that fuses global contextual features with unimodal features, and solves the problem of over-smoothing of existing graph neural networks by jointly optimizing modal fusion and graph comparison learning. Hu et al. (2021) proposed a multimodal fused graph convolutional network called MMGCN which uses three modalities to construct a multimodal graph. It establishes connections at the internal nodes of each modality and establishes connections between the modalities. Nguyen et al. first proposed the use of directed acyclic graphs (DAG) to integrate multimodal features and introduced course learning to deal with the sentiment category class imbalance problem. Nguyen et al. (2024a) proposed a unified framework using directed acyclic graphs (DAG) to integrate language, audio, and visual features, and introduced course learning to deal with emotional shifts and data imbalance. Tu et al. (2024a) introduced a network called Multi-Knowledge Enhanced Interaction Graph Network (MKE-IGN), which facilitates the modeling of the relationship between utterances and different types of CSK by integrating a variety of knowledge such as language and visual CSK into edge representations.

The main difference between our proposed approach and the above approaches is that existing work treats all modalities equally when performing multimodal information interaction, i.e., all modalities are assumed to have the same quality. However, it has been shown that the quality of the different modalities varies, so we believe that multimodal interactions should be performed in a non-reciprocal manner. Also previous approaches to non-reciprocal fusion have taken the approach of using the dominant modality to augment the weaker auxiliary modality, whereas our approach is just the opposite we further augment the dominant modality by the auxiliary modality and extract the conversation atmosphere representation from it to augment the multimodal representation. Our hypothesis is that directly using a dominant modality (like language) to augment weaker ones can sometimes overwhelm or even corrupt their inherent, subtle emotional cues. Instead of this direct "strong-to-weak" augmentation, our "opposite" approach is more nuanced. We first use language as a lens to extract relevant signals from the other modalities. These signals are then integrated to form a unified, global conversational atmosphere representation. This global context is then used to augment all modalities, including language itself. This two-step process ensures that the augmentation is based on a holistic understanding of the conversation's emotional tone, rather than just the raw semantic power of the text, leading to a more robust and conceptually sound fusion.

To this end, we propose a novel Language-Focused Augmented Transformer, which takes the language modality as the dominant modality and the other modalities as the auxiliary modalities, and further enhances the language modality with the help of the other modalities. Unlike the approach that inspired ours, Wang et al. (2024) merely use the cross-attention mechanism with language as the dominant to extract the enhanced multimodal representation of individual samples is a sample-level enhancement, whereas our approach utilizes the language modality to obtain the global atmosphere representation of the whole conversation, which is an enhancement at the conversation level. And in order to prevent the situation

where the discourse of the same conversation is augmented by the same global atmosphere representation resulting in the inability to distinguish similar emotions, we introduce a dual contrastive learning mechanism. Second, in fusing multimodal features unlike the existing work, we propose a Cross-Modality Variational Distillation Fusion method, which calculates the weights of each modality based on multimodal Gaussian distributions for fusion and improves the quality of the auxiliary modalities through distillation learning.

3 Methods

In this section, we provide a detailed description of each component of our model which consists of three main parts: Feature Representation, Language-Focused Augmented Transformer, and Cross-Modality Variational Distillation Fusion. The architecture of the proposed model is illustrated in Fig. 1.

3.1 Task definition

Let $U = \{\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_n\}$ be a conversation produced by $m \geq 2$ speakers, consisting of n utterances. Each utterance is represented by a triplet $\mathrm{u}_i = \{\mathrm{u}_i^a, \mathrm{u}_i^v, \mathrm{u}_i^l\}$, where u_i^a, u_i^v, and u_i^l denote the audio, visual, and language features of $\mathbf{u_i}$, respectively. MERC aims to predict the emotion label y_i of each utterance u_i based on its previous utterances.

3.2 Feature representation

The Feature Representation module consists of two submodules, i.e., Utterance Feature Extraction and Utterance-Level Augmentation.

3.2.1 Utterance feature extraction

In this paper, we follow up-to-date previous works (Yang et al., 2025; Guo et al., 2024) and employ pre-trained models for feature extraction. We use RoBERTa (Liu et al., 2019), OpenSMILE (Eyben et al., 2010), and DenseNet (Iandola et al., 2014) to extract language, audio, and visual features respectively. Formally, we have:

$$\begin{aligned} \mathrm{H}_i^l &= \mathrm{RoBERTa}(\boldsymbol{u}_i^l) \in \mathbb{R}^{d_l}, \\ \mathrm{H}_i^a &= \mathrm{DenseNet}(\boldsymbol{u}_i^a) \in \mathbb{R}^{d_a}, \\ \mathrm{H}_i^v &= \mathrm{OpenSMlLE}(\boldsymbol{u}_i^v) \in \mathbb{R}^{d_v}, \end{aligned} \tag{1}$$

where d_l, d_a, and d_v represent the dimensions of the language, audio, and visual features, respectively.

3.2.2 Utterance-level augmentation

The features (denoted as $\mathrm{H}^\xi \in \mathbb{R}^{n \times d_\xi}$, where $\xi \in \{a, v, l\}$) extracted by the pre-trained encoder lack of contextual information in the conversation. To further capture and extract

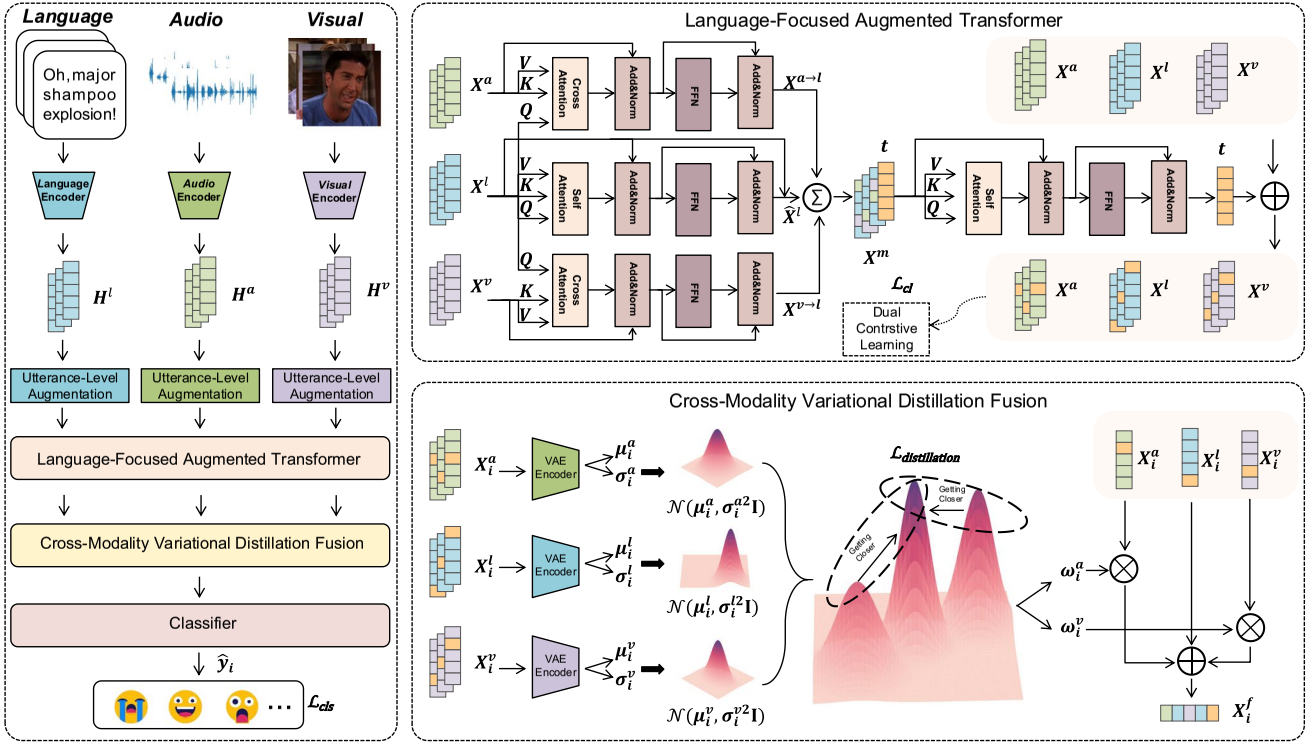

Fig. 1 Overall of LFVD, which consists of three main components: Feature Representation, Language-Focused Augmented Transformer module, and Cross-Modality Variational Distillation Fusion module

the internal contextual information of each modality and project the features of all modalities into the same semantic space. For this purpose, we use Bi-GRU (Cho et al., 2014) and DAG (Shen et al., 2021) to extract the contextual information in the conversation. Considering the inherently sequential nature of conversations, we take advantage of Bi-GRU to extract contextual information in the time dimension:

$$\hat{\mathrm{H}}^{\xi} = [\hat{\mathrm{H}}_1^{\xi}, \hat{\mathrm{H}}_2^{\xi}, \ldots, \hat{\mathrm{H}}_n^{\xi}] = \text{Bi-GRU}(\mathrm{H}_1^{\xi}, \mathrm{H}_2^{\xi}, \ldots, \mathrm{H}_n^{\xi}) \in \mathbb{R}^{n \times d_r}, \tag{2}$$

where d_r is the output dimension of the Bi-GRU.

The following steps are proposed to refine the contextual information, as well as model speaker identity and location relationships. Following the previous work (Shen et al., 2021), we introduce the Directed Acyclic Graph Network (DAG-ERC), which uses a DAG to model the information flow in conversations. In each layer l of DAG-ERC, the hidden state of utterances is updated recurrently, capturing the temporal flow from the first to the last utterance:

$$\mathrm{X}^{\xi} = \mathrm{DAG}(\hat{\mathrm{H}}^{\xi}). \tag{3}$$

The results of the enhancement of each modality according to the above method are: X^{a}, X^{v} and $\mathrm{X}^{l} \in \mathbb{R}^{n \times d_x}$.

3.3 Language-focused augmented transformer

To solve the problem of redundant and conflicting information introduced during multimodal information interaction. Our framework primarily focuses on the language modality, leveraging the other two modalities as auxiliary components to enrich language features. Then, these enriched language features are used to derive conversational atmosphere representation, which subsequently optimizes the representations of each modality. This approach entails introducing global information for each utterance, thereby minimizing the possibility of redundant and conflicting information. To further enhance each modality, a contrast learning strategy is introduced. At the core of our architecture lies a hybrid attention mechanism. The mechanism combines dual self-attention layers with dual cross-modal attention layers, where the language modality consistently serves as the query. Building upon this foundation, we developed a novel Multi-head Language-guided Attention (MLA) mechanism that extends the conventional multi-head self-attention framework.

Specifically, given the input feature matrix X^{ξ}, we first define a set of weight matrices $\mathrm{W}_{Q_\xi}, \mathrm{W}_{K_\xi}, \mathrm{W}_{V_\xi}$ to obtain query, key and value as:

$$\mathrm{Q}_{\xi} = \mathrm{X}^{\xi}\mathrm{W}_{Q_\xi}, \mathrm{K}_{\xi} = \mathrm{X}^{\xi}\mathrm{W}_{K_\xi}^{T}, \mathrm{V}_{\xi} = \mathrm{X}^{\xi}\mathrm{W}_{V_\xi}. \tag{4}$$

Like the original transformer, we extend the language-guided attention to Multi-head Language-guided Attention (MLA). This is denoted as:

$$\begin{aligned} \mathrm{MLA}(\mathrm{Q},\mathrm{K},\mathrm{V}) &= (head_1 \parallel head_2 \parallel \cdots \parallel head_h)\mathrm{W}_o, \\ head_i &= \text{attention }(\mathrm{Q}_i,\mathrm{K}_i,\mathrm{V}_i), i \in \{1,2,\cdots,h\}, \\ attention(\mathrm{Q}_i,\mathrm{K}_i,\mathrm{V}_i) &= softmax\left(\frac{\mathrm{Q}_i\mathrm{K}_i^T}{\sqrt{d_h}}\right)\mathrm{V}_i, \end{aligned} \tag{5}$$

where $(\cdot \parallel \cdot)$ denotes vector concatenation operation, $\mathrm{W}_o \in \mathbb{R}^{h\cdot d_h \times d_x}$ represents the learnable parameter matrix, h denotes the number of headers used in the attention mechanism and d_h represents the dimension of the head.

Firstly, MLA is used to extract complementary information about the language modality from other modal features:

$$\begin{aligned} \tilde{\mathrm{X}}^{a\to l} &= \mathrm{MLA}(\mathrm{Q}_l,\mathrm{K}_a,\mathrm{V}_a), \\ \bar{\mathrm{X}}^{a\to l} &= \mathrm{LN}(\tilde{\mathrm{X}}^{a\to l} + \mathrm{X}^a), \\ \mathrm{X}^{a\to l} &= \mathrm{LN}(\mathrm{FFN}(\bar{\mathrm{X}}^{a\to l}) + \bar{\mathrm{X}}^{a\to l}), \end{aligned} \tag{6}$$

where $\mathrm{LN}(\cdot)$ and $\mathrm{FFN}(\cdot)$ indicate , normalization and feedforward network, respectively. In the same way, we get $\mathrm{X}^{v\to l}$.

To obtain a more advanced representation of the language modality and contextual information, we use MLA to further process the language modality:

$$\begin{aligned} \tilde{\mathrm{X}}^{l\to l} &= \mathrm{MLA}(\mathrm{Q}_l,\mathrm{K}_l,\mathrm{V}_l), \\ \bar{\mathrm{X}}^{l\to l} &= \mathrm{LN}(\tilde{\mathrm{X}}^{l\to l} + \mathrm{X}^l), \\ \hat{\mathrm{X}}^{l} &= \mathrm{LN}(\mathrm{FFN}(\bar{\mathrm{X}}^{l\to l}) + \bar{\mathrm{X}}^{l\to l}), \end{aligned} \tag{7}$$

Subsequently, the obtained features $\mathrm{X}^{a\to l}$, $\mathrm{X}^{v\to l}$, and $\hat{\mathrm{X}}^{l}$ are aggregated together to obtain a high-level representation of the language modality. To preserve low-level language representations, we employ the notion of residual connectivity, incorporating initial language features X^l. Finally, the resulting feature representation is denoted as $\mathrm{X}^m \in \mathbb{R}^{n\times d_x}$:

$$\mathrm{X}^m = \mathrm{X}^{a\to l} + \mathrm{X}^{v\to l} + \hat{\mathrm{X}}^{l} + \mathrm{X}^l. \tag{8}$$

The resulting X^m contains multimodal information about the conversation, and to enhance the representation of each modality, we extract the global information in X^m using the idea of adding a learnable vector t. This vector t is concatenated to $\mathrm{X}^m = [\mathrm{t}, \mathrm{x}_1^{\xi}, \ldots, \mathrm{x}_n^{\xi}] \in \mathbb{R}^{(n+1)\times d_x}$. The global dependencies are to be simulated through the mechanism of attention, thereby enabling t to express global information in conversation. Finally, the conversation's emotional atmosphere representation t is extracted from $\hat{\mathrm{X}}^m$ and augmented with a multimodal representation of each utterance in the conversation via a broadcast mechanism.

$$\begin{aligned} \tilde{\mathrm{X}}^m &= \mathrm{MLA}(\mathrm{Q}_m,\mathrm{K}_m,\mathrm{V}_m, \\ \bar{\mathrm{X}}^m &= \mathrm{LN}(\tilde{\mathrm{X}}^m + \mathrm{X}^m), \\ \hat{\mathrm{X}}^m = [\mathrm{t}, \mathrm{x}_1^{\xi}, \ldots, \mathrm{x}_n^{\xi}] &= \mathrm{LN}(\mathrm{FFN}(\bar{\mathrm{X}}^m) + \bar{\mathrm{X}}^m), \\ \mathrm{X}^{\xi} &= \mathrm{X}^{\xi} + \mathrm{t}, \end{aligned} \tag{9}$$

where $\xi \in \{a, v, l\}$, the enhanced multimodal representation $\mathrm{X}^a, \mathrm{X}^v, \mathrm{X}^l \in \mathbb{R}^{n \times d_x}$.

To foster multimodal feature alignment during training, we introduce Dual Contrastive Learning, which consists of two parts: 1) fully supervised contrastive learning based on the emotional label $\mathcal{L}_{\text{full}}$, and 2) self-supervised inter-modal contrastive learning based on the multimodal space semantics $\mathcal{L}_{\text{self}}$.

For different modalities ξ and ξ' of the same sample, we minimize the distance between utterances from different modalities within the same sample, while maximizing the distance between utterances from different samples:

$$\mathcal{L}_{self} = -\frac{1}{N} \sum_{\xi,\xi' \in \{a,v,l\}}^{\xi \neq \xi'} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathrm{x}_i^{\xi}, \mathrm{x}_i^{\xi'})/\tau)}{\sum_{j=1}^{N} \mathbb{I}_{j \neq i} \exp(\text{sim}(\mathrm{x}_i^{\xi}, \mathrm{x}_j^{\xi'})/\tau)}, \tag{10}$$

where $\text{sim}(\cdot, \cdot)$ denotes the dot product of L2-normalized feature vectors, τ is the similarity scaling factor, and $\mathbb{I}_{j \neq i}$ is an indicator function that equals 1 when $j \neq i$, and 0 otherwise.

For each modality ξ, we minimize the distance between utterance features with the same label in the semantic space, while maximizing the distance between utterances with different labels:

$$\mathcal{L}_{full} = -\frac{1}{N} \sum_{\xi \in \{a,v,l\}} \sum_{i=1}^{N} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\text{sim}(\mathrm{x}_i^{\xi}, \mathrm{x}_p^{\xi})/\tau)}{\sum_{k=1}^{N} \mathbb{I}_{k \notin \rangle} \exp(\text{sim}(\mathrm{x}_i^{\xi}, \mathrm{x}_k^{\xi})/\tau)}, \tag{11}$$

where $\mathcal{P}(i)$ denotes the set of utterances with the same label as the i-th utterance.

3.4 Cross-modality variational distillation fusion

The Cross-Modality Variational Distillation Fusion consists of two submodules, i.e., Variational Distillation Learning and Cross-Modality Fusion.

3.4.1 Variational distillation learning

We have proposed a distribution-based knowledge distillation method to improve the quality of the auxiliary modality and filter out information irrelevant to the task at the same time.

The unimodal features are fixed for each input sample, making it difficult to directly estimate their distributions. To address this, we adopt a generative approach to model the unimodal features, where the unimodal features are drawn from a latent space with isotropic Gaussian priors:

$$q(\mathrm{z}_i^{\xi} | \mathrm{x}_i^{\xi}) = \mathcal{N}(\mathrm{z}_i^{\xi} \mid \mu(\mathrm{x}_i^{\xi}), \sigma(\mathrm{x}_i^{\xi})). \tag{12}$$

In this way, the representation of each multimodal sample is not restricted to deterministic point embeddings but is a consistent fuzzy representation over multiple multivariate Gaussian distributions. Subsequently, we align the audio and visual modalities with the language modality by minimizing the Kullback-Leibler (KL) divergence. In this process, the lan-

guage modality acts as the teacher, and the other two modalities act as the students, thereby improving the representation of the auxiliary modality:

$$\begin{aligned} L_{distillation} = & \sum_{m \neq l}^{m \in \mathcal{M}} \mathrm{D}_{KL}\left(q(\mathrm{z}_i^m|\mathrm{x}_i^m) \| q(\mathrm{z}_i^l|\mathrm{x}_i^l)\right) \\ & + \sum_{m \in \mathcal{M}} \mathrm{D}_{KL}\left(q(\mathrm{z}_i^m|\mathrm{x}_i^m) \| \mathcal{N}(0, \mathrm{I})\right), \end{aligned} \tag{13}$$

where $D_{KL}(\cdot||\cdot)$ stands for the KL divergence, $\mathcal{M} \in \{a, v, l\}$, the first part of the formula aims to align the distributions of the audio and visual modalities with the distribution of the language modality. The second part of the formula pushes each distribution towards a normal distribution, to introduce a strong prior constraint in the latent space, thereby ensuring the continuity and controllability of the distributions.

3.4.2 Cross-modality fusion

Evaluating the importance of different modalities during Cross-Modality Fusion can reduce the introduction of redundant and conflicting information, which is beneficial for improving multimodal representations. Furthermore, the distributional divergence between unimodal features is interpreted as the information gap between different modalities. Specifically, we do this by evaluating the KL divergence between unimodal distributions.

Thus, the weight of different modalities in data sample x_i can be measured by the averaged KL divergence between unimodal distributions as follows:

$$w_i^{l \to v} = \mathrm{D}_{KL}\left(q\left(\mathrm{z}_i^l \,||\, \mathrm{x}_i^l\right) \,||\, q\left(\mathrm{z}_i^v \,||\, \mathrm{x}_i^v\right)\right), \tag{14}$$

$$w_i^{v \to l} = \mathrm{D}_{KL}\left(q\left(\mathrm{z}_i^v \,||\, \mathrm{x}_i^v\right) \,||\, q\left(\mathrm{z}_i^l \,||\, \mathrm{x}_i^l\right)\right), \tag{15}$$

$$w_i^v = \mathrm{sigmoid}\left(\frac{1}{2}\left(w_i^{l \to v} + w_i^{v \to l}\right)\right), \tag{16}$$

where the sigmoid function is used to scale the average KL divergence distance to the range of 0 to 1, for subsequent use. Likewise, we can compute w_i^a. Then, we perform the weighted fusion using the obtained weights:

$$\mathrm{x}_i^f = w_i^a \mathrm{x}_i^a + w_i^v \mathrm{x}_i^v + \mathrm{x}_i^l, \tag{17}$$

where w_i^a and w_i^v represent the weights of the audio and visual modes, respectively. x_i^f represents the final representation of the utterance x_i, which is then passed into the classifier to obtain the predicted label.

3.5 Emotion classifier

In the previous section, we obtained the final representation of the utterance u_i denoted as x_i^f. Then, we pass it through a fully connected network to predict the corresponding label:

$$\hat{y}_i = \mathrm{Classifier}(\mathrm{x}_i^f), \tag{18}$$

where $\mathrm{Classifier}(\cdot)$ is composed of a fully connected layer followed by a function $\mathrm{softmax}(\cdot)$. We utilize the cross-entropy loss for training the classifier. In particular, the cross-entropy loss between the predicted label distribution $\hat{y}_i$ and the true label distribution y_i is defined as follows: We use the cross-entropy loss function as:

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

where N represents the number of utterances in the batch, and C denotes the number of emotion categories. The final loss function is defined as:

$$\mathcal{L}_{\mathrm{cl}} = \gamma_1 \mathcal{L}_{\mathrm{self}} + \gamma_2 \mathcal{L}_{\mathrm{full}}, \tag{20}$$

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{cl} + \gamma_3 \mathcal{L}_{distillation}, \tag{21}$$

where γ_1, γ_2, and γ_3 are hyperparameters.

4 Experiments

4.1 Datasets and evaluations

We use the MELD (Poria et al., 2019) and IEMOCAP (Busso et al., 2008) datasets to evaluate our proposed model. Detailed information on these two datasets can be found in Table 1.

MELD inspired by *Friends*, a multi-speaker conversation dataset was compiled, comprising 1,433 dialogues and 13,708 utterances. Each utterance is classified into one of seven emotional categories: *Neutral, Surprise, Fear, Sadness, Joy, Disgust, Angry.*

IEMOCAP this dataset encompasses dyadic interactions involving ten speakers, consisting of 153 dialogues and 7,433 utterances. Each utterance is classified into one of six emotions: *Happy, Sad, Neutral, Angry, Excited, Frustrated.*

Table 1 Data distribution of IEMOCAP and MELD

Dataset	Conversation			Utterance			Utterance per Conversation	Classes
	train	valid	test	train	valid	test	Utterance/Conversation	
MELD	1039	114	280	9989	1109	2610	9	7
IEMOCAP	120		31	5810		1623	52	6

Evaluation metrics to balance the model's overall performance and category-level fairness, we adopt overall accuracy and the weighted average F1-score as evaluation metrics. Additionally, to provide a detailed view of the model's performance across individual categories, we also report the F1-score for each class.

4.2 Baselines

We used state-of-the-art models as the baseline for comparison.

DialogueRNN (Majumder et al., 2019): introduce an RNN-based neural network architecture that integrates each input utterance within the unique characteristics of the speaker, providing a more enriched context.

DialogueGCN (Ghosal et al., 2019): introduce a GCN-based ERC approach that models conversational context for emotion recognition by leveraging self-dependency and inter-speaker dependency.

MMGCN (Hu et al., 2021): effectively leverages multimodal dependencies while incorporating speaker information to model inter-speaker relationships.

CTNet (Lian et al., 2021): introduce a transformer-based structure to model intra- and Cross-Modality interactions, with word-level and segment-level features as input to capture temporal information in utterances.

MM-DFN (Hu et al., 2022): introduce a graph-based dynamic fusion module that integrates multimodal contextual features to reduce redundancy and enhance each modality.

SCMM (Yang et al., 2023): propose the Self-adaptive Context and Modal-interaction Modeling (SCMM) framework. Contains context representation module which consists of three submodules to model multiple contextual representations, modal-interaction module for multimodal information interaction and self-adaptive path selection module integrating multimodal data.

CMCF-SRNet (Zhang & Li, 2023): propose a Cross-Modality locality-constrained transformer to explore the multimodal interaction and investigate a graph-based semantic refinement transformer, which solves the limitation of insufficient semantic relationship information between utterances.

MultiDAG (Nguyen et al., 2024b): introduce a multimodal ERC approach combining Directed Acyclic Graphs (DAG) for integrating language, audio, and visual features with Curriculum Learning (CL) to address emotional shifts and data imbalance, enhancing model performance.

AdaIGN (Tu et al., 2024b): introduce an adaptive graph network that balances intra- and inter-speaker dependencies, employs the Gumbel Softmax trick to adaptively select nodes

and edges, and enhances multimodal ERC with adaptive selection policies and a task-specific loss.

4.3 Detailed settings

Our model is implemented using PyTorch on a single NVIDIA RTX 4090 GPU. We adopt the Adam optimizer, with the initial learning rate set to 6.2e-5 for IEMOCAP and 2.5e-4 for MELD. Additionally, the L2 regularization factor is set to 7e-3, with a batch size of 20 for IEMOCAP and 32 for MELD. The maximum number of training epochs is set to 100 for IEMOCAP and 80 for MELD. For γ_1, γ_2, and γ_3 are set 3.5e-2, 5e-3, and 3e-3 for IEMOCAP, 1.35e-3, 5e-3 and 3e-3 for MELD. To ensure fairness, all reported results are the average of five random runs in the test set.

4.4 Overall performances

We evaluated the overall performance of our model against the baseline method on the IEMOCAP and MELD datasets, as shown in Table 2 and Table 3, respectively. Our model consistently outperforms the state-of-the-art methods, showing superior accuracy (ACC) and weighted F1 (W-F1) scores on both datasets. On the IEMOCAP dataset, our model outperforms all baseline models in overall performance. Specifically, the overall accuracy improved by 2.49% and the weighted W-F1 score improved by 2.01% compared to the best baseline model (AdaIGN). In addition, our model achieved the highest performance in all four sentiment categories of Happy, Sad, Neutral, and Excited. Similarly, on the MELD dataset, our model outperformed all baseline models. Compared to the best baseline model (AdaIGN), the overall accuracy improved by 0.70% and the weighted W-F1 score improved by 0.24%. In addition, our model also achieved the best results in the Neutral, Fear, Sadness, Disgust, and Angry five emotion categories for this dataset. Although our model demonstrates the best results in overall performance (ACC and W-F1), it does not achieve the best results across all emotion categories. For instance, in the IEMOCAP dataset, the best models for the Angry and Frustrated categories are MultiDAG and AdaIGN, respectively. Upon

Table 2 Experimental results on IEMOCAP dataset

Model	IEMOCAP							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	ACC	W-F1
DialogueRNN♯	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89
DialogueGCN♯	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89
MMGCN♯	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26
CTNet♭	51.30	79.90	65.80	67.20	78.70	58.80	68.00	67.50
MM-DFN♭	42.22	78.98	66.42	69.77	75.56	66.33	68.21	68.18
SCMM♭	45.37	78.76	63.54	66.05	76.70	66.18	–	67.53
CMCF-SRNet♭	52.20	80.90	68.80	70.30	76.70	61.60	70.50	69.60
MultiDAG♭	49.65	81.40	69.53	**70.33**	71.61	66.94	69.11	69.08
AdaIGN♭	53.04	81.47	71.26	65.87	76.34	**67.79**	70.49	70.74
ours	**54.69**	**85.43**	**74.60**	67.65	**79.32**	66.66	**72.98**	**72.75**

Best results in bold, second best underlined. ♯ and ♭ results come from Hu et al. (2022) and original papers, respectively

Table 3 Experimental results on MELD dataset

Model	MELD								
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Angry	ACC	W-F1
DialogueRNN♯	76.97	47.69	–	20.41	50.92	–	45.52	60.31	57.66
DialogueGCN♯	75.97	46.05	–	19.60	51.20	–	40.83	58.62	56.36
MMGCN♯	76.33	48.15	–	26.74	53.02	–	46.09	60.42	58.31
CTNet♭	77.40	52.70	10.00	32.50	56.00	11.20	44.60	62.00	60.50
MM-DFN♭	77.76	50.69	–	22.93	54.78	–	47.82	62.49	59.46
SCMM♭	–	–	–	–	–	–	–	–	59.44
CMCF-SRNet♭	–	–	–	–	–	–	–	62.80	62.30
MultiDAG♭	–	–	–	–	–	–	–	64.41	64.00
AdaIGN♭	79.75	**60.53**	–	43.70	**64.54**	–	56.15	67.62	66.79
ours	**80.02**	58.61	**19.63**	**43.73**	64.45	**27.71**	**58.24**	**68.32**	**67.03**

Best results in bold, second best underlined. ♯ and ♭ results come from Hu et al. (2022) and original papers, respectively

careful analysis of these baseline models and our own model, we observed that MultiDAG and AdaIGN all construct multimodal interaction graphs to capture fine-grained information from each modality. These baseline models capture complex emotional cues in dialogues through the construction of sophisticated interaction graphs, enabling the effective recognition of different emotions based on the extracted cues. In contrast, our model focuses on enhancing and integrating information from different modalities with a language-centric approach, without the need to construct intricate interaction graphs to capture emotional cues. This is the reason why our model does not achieve optimal results across all emotion categories.

In summary, the LFVD proposed in this paper can effectively mitigate the occurrence of information conflict and redundancy when interacting with multimodal information, and also performs well when fusing multimodal information. This is further supported by the statistical analysis, where the p-value is $\ll 0.05$ compared to the AdaIGN.

4.5 Ablation study

4.5.1 Effect of each modalities

Table 4 shows the performance of our model. We have drawn the following conclusions: (1) Multimodal data input yields superior model performance compared to single-modal approaches, with the language modality demonstrating significantly better results than the other two modalities. This provides strong evidence for us to use the language modality as the dominant modality. (2) By comparing L and L+V+A, it can be found that the model performance is better when adding the infor-mation of the other two modalities compared to using only the text mode. This is because by integrating information from multiple modalities, a more comprehensive and detailed representation can be obtained, enabling the model to capture complex emotional cues in conversations.

Table 4 Ablation experiments for exploring the effects of each modality as well as each combinations of modalities

	IEMOCAP		MELD	
	ACC	W-F1	ACC	W-F1
T	68.87	67.92	67.23	66.45
V	40.12	37.89	45.78	34.67
A	56.03	54.21	48.56	42.34
T+V	68.74	68.89	67.82	66.53
T+A	70.98	70.23	67.91	66.78
A+V	62.45	61.87	49.32	43.89
T+A+V	72.98	72.75	68.32	67.03

Table 5 Ablation experiments for exploring the importance of each components

	IEMOCAP		MELD	
	ACC	W-F1	ACC	W-F1
w/o LFA-Transformer	67.34	67.52	67.55	66.22
w/o CEAR	68.76	68.94	68.28	66.64
w/o DL	72.70	71.91	68.05	66.87
w/o CMVD	71.29	70.79	67.78	66.42
w/o CMVD-Fusion	72.03	71.76	67.51	66.56
w/o CMVD-Distillation	71.60	71.04	67.74	66.79
w/ours(A)	70.73	70.32	68.20	66.76
w/ours(V)	70.61	69.91	67.70	66.37
ours	72.98	72.75	68.32	67.03

4.5.2 Effect of each components

To assess the contributions of each module in our model, we perform ablation experiments on two datasets. Furthermore, for the modules LFA and CMVD, we tested the effect of replacing the dominant modality with either the audio or visual modality.

- **w/o LFA-Transformer** without the Language-Focused Augmented Transformer.
- **w/o CEAR** without the conversational emotional atmosphere representation.
- **w/o DL** without the Dual Contrstive Learning.
- **w/o CMVD** without the Cross-Modality Variational Distillation Fusion.
- **w/o CMVD-Fusion** without the Cross-Modality Fusion.
- **w/o CMVD-Distillation** without the Variational Distillation.
- **w/ ours(A)** replace the main modality in the model with the audio modality.
- **w/ ours(V)** replace the main modality in the model with the visual modality.

The results of the ablation experiments are shown in the Table 5 (1) The removal of the Language Focused Augmentation Transformer module (LFA-Transformer) resulted in a degradation of the model's performance due to the loss of the model's ability to interact with multimedia messages, which was more pronounced for the IEMOCAP dataset than for MELD. This is because the average conversation length of the IEMOCAP dataset is significantly longer than that of the MELD dataset. As a result, the module is able to learn multimodal global information from longer conversations, thus effectively enhancing individual utterances in the conversation. In addition, we tested the removal of the cross-modal variant distillation fusion module (CMVD) and found that the removal of this module also

affects the model performance. This suggests that the module's approach of distilling and weighting the remaining two modalities with language as the teacher effectively improves the final fusion representation and improves model performance. This highlights the fact that learning by refining and weighting the fusion of multimodal information is crucial for multimodal tasks. Moreover, we report the performance loss of removing one of the sub-modules, which shows that each module has its own contribution. (2) We find that using either the audio modality (w/ our(A)) or the visual modality (w/ our(V)) as the dominant modality leads to performance degradation compared to using the language modality. This phenomenon can be attributed to the higher quality of the language modality, which provides the best quality multimodal cues for fusion and extraction. These experiments fully validate the importance and justification of using language modality as the dominant modality in model construction.

4.5.3 Effect of different loss items

Table 6 shows the effect of each of the Loss Items proposed in this paper on the performance of the model. The performance of the two datasets shows a decreasing trend as the Loss Items are gradually removed, and the performance of the model is impaired regardless of the removal of that Loss Items, which illustrates the necessity of the individual Loss Items in the paper.

4.5.4 Effect of different fusion methods

To validate the efficacy of the Cross-Modality Variational Distillation Fusion (CMVD-Fusion) approach introduced in our framework, we conducted a comprehensive comparison against five widely used multimodal fusion strategies: Add, Cat, and Gate.

As shown in Fig. 2, the performance of our proposed CMVD-Fusion method outperforms other fusion methods, demonstrating its effectiveness. The main reason lies in the fact that our method does not directly merge or concatenate the modality representations. Instead, it evaluates the distributional differences between modalities and applies a Gaussian distribution-based weighted fusion approach. This method alleviates the issue of modality heterogeneity, making it more effective than typical fusion strategies.Previous gate-based fusion methods only consider the direct relationships between different modalities. In contrast, our Gaussian distribution-based weighted fusion approach also takes into account the uncertainties between modalities, thereby achieving a better fusion of the diverse modality representations.

Table 6 Ablation study on different loss items

	IEMOCAP		MELD	
	ACC	W-F1	ACC	W-F1
w/o $\mathcal{L}_{\text{self}}$	72.77	72.14	68.05	66.93
w/o $\mathcal{L}_{\text{full}}$	72.09	71.45	67.70	66.74
w/o $\mathcal{L}_{\text{distillation}}$	71.60	71.04	67.74	66.79
w/o $\mathcal{L}_{\text{all}}$	70.73	70.77	68.39	66.50
ours	72.98	72.75	68.32	67.03

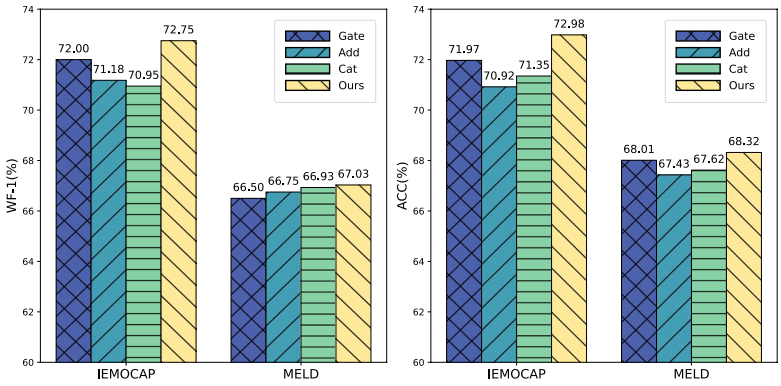

Fig. 2 Performance of different fusion methods on two datasets: the left panel shows the WF-1 score, and the right panel displays the Accuracy score

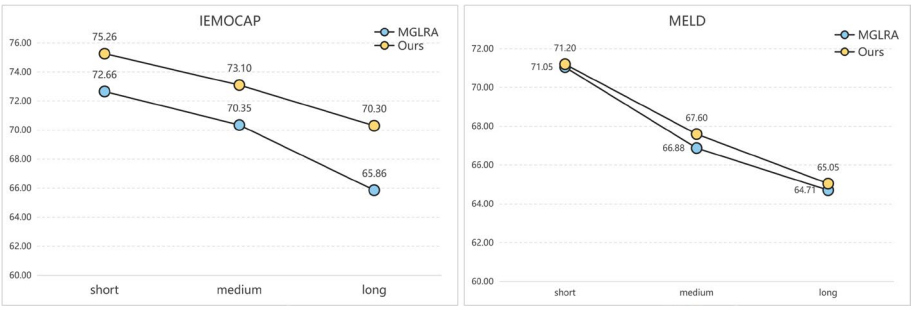

Fig. 3 Comparison of the performance of MGLRA and LFVD under different conversation lengths

4.6 Analysis of conversation length

We report the performance of the model (WF-1) for different conversation lengths, as shown in Fig. 3. Based on the conversation length distribution characteristics of the IEMOCAP and MELD datasets, we adopt differentiated classification criteria for multi-scale analysis. Specifically, for the IEMOCAP dataset, short conversations are defined as length ≤ 43 (corresponding to the first quartile $Q_1 = 43$), medium conversations as $43 < \text{length} \leq 59$ (corresponding to the third quartile $Q_3 = 59$), and long conversations as length > 59; whereas, for the MELD dataset, due to presenting significant short conversation characteristics, the classification thresholds are set as ≤ 5 ($Q_1 = 5$), $5 < \text{length} \leq 13$ ($Q_3 = 13$), and > 13.

The results show that our approach consistently outperforms MGLRA (Meng et al., 2024) on both datasets across all lengths of conversations.Furthermore, the performance of the model progressively decreases as the length of the conversation increases, a phenomenon that can be attributed to the fact that long conversations are usually accompanied by enhanced cross-cutting semantic dependencies, resulting in key emotion cues propagating in the temporal sequence information decay during the process, a phenomenon that is particularly significant in the Transformer architecture.

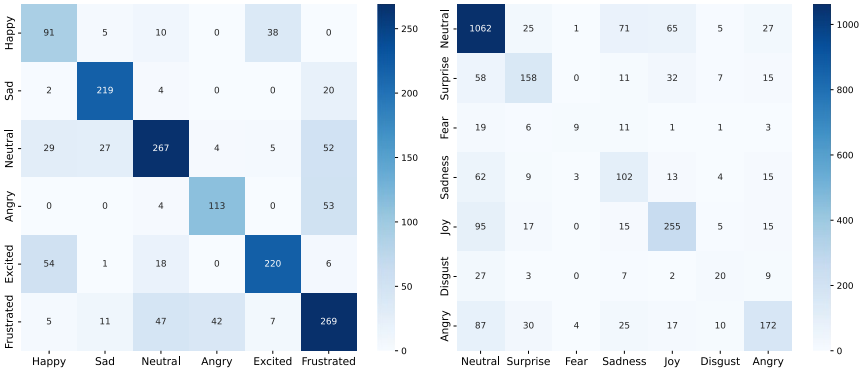

Fig. 4 Confusion matrix for IEMOCAP(left) and MELD(right)

4.7 Error analysis

In Fig. 4, we visualize the model's prediction results through a confusion matrix. From the figure, we observe that in the IEMOCAP dataset, the model often misclassifies 'happy' samples as 'excited' or 'neutral'. A similar trend is seen with 'excited' samples. This is largely due to the similarity between 'excited' and 'happy', as well as the fact that 'neutral' samples constitute a significant portion of the dataset, causing the model to focus more on them. In the MELD dataset, 'neutral' is the dominant class, accounting for as much as 48%. As a result, the model tends to focus on classifying 'neutral' samples during training, which leads to the underperformance in classifying other emotions. Particularly, the 'fear' and 'disgust' categories, which are less represented in the dataset, are often misclassified as 'neutral'. Moreover, misclassifications between similar emotional categories, such as'surprise' and 'joy', are also observed in this dataset. Although the model somewhat alleviates these issues, further improvements in recognizing similar emotions and addressing the underrepresentation of certain emotion categories could significantly enhance its performance.

5 Conclusion

We propose the LFVD, designed to effectively reduce the redundant and conflicting information that arises during multimodal information interaction. Specifically, we introduce a novel Language-Focused Augmented Transformer, which enhances multimodal information in a manner driven by the language modality. Additionally, we present Cross-Modality Variational Distillation Fusion, which aims to generate a robust modal representation by encoding multimodal representations as variational distributions. We conduct comparative experiments on two widely used datasets. The experimental results demonstrate that our method outperforms state-of-the-art techniques on both datasets, thereby verifying the effectiveness of our model and the hypothesis that focusing on the language modality can achieve effective multimodal interaction. As future work, to futher enhance the performance of our approach on all tasks, especially the two tasks with lower metrics, we will propose to develop novel contrastive learning strategy specialized for similar emotions.

Author Contributions S.J. Implemented the algorithms and conducted experiments, S.J. and X.Z. wrote the main manuscript text. All authors reviewed the manuscript.

Funding This work was supported by the Science and Technology Innovation Key R&D Program of Chongqing (CSTB2024TIAD-STX0027), the National Natural Science Foundation of China (62472059), the Chongqing Talent Plan Project, China (CSTC2024YCJH-BGZXM0022), the Open Research Fund of Key Laboratory of Cyberspace Big Data Intelligent Security (Chongqing University of Posts and Telecommunications), Ministry of Education (CBDIS202403).

Data Availability The IEMOCAP dataset (https://sail.usc.edu/iemocap/iemocaprelease.htm) and the MELD dataset (https://affective-meld.github.io/).

Declarations

Ethical Approval The authors state that this research complies with ethical standards. This research does not involve either human participants or animals.

Competing interests The authors declare no competing interests.

Availability of supporting data The datasets used during the current study are available. Additionally, the datasets generated, model settings, and training processes are available from the corresponding author upon reasonable request.

References

Ai, W., Zhang, F., Shou, Y., et al. (2025). Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(11), 11418–11426. https://doi.org/10.1609/aaai.v39i11.33242

Busso, C., Bulut, M., Lee, C., et al. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Lang Resour Evaluation, 42*(4), 335–359. https://doi.org/10.1007/S10579-008-9076-6

Cho, K., van Merriënboer ,B., Gulcehre, C., et al. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1724–1734, https://doi.org/10.3115/v1/D14-1179

Czerwinski, M., Gilad-Bachrach, R., Iqbal, S., et al. (2016). Challenges for designing notifications for affective computing systems. In: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: adjunct, pp. 1554–1559. https://doi.org/10.1145/2968219.2968548

Eyben, F., Wöllmer, M., Schuller, B.W. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th international conference on multimedia, pp. 1459–1462, https://doi.org/10.1145/1873951.1874246

Gan, X., Huang, X., Zou, S. (2025) Intentional tendency-based dynamic heterogeneous graph network for emotion recognition in conversations. Journal of Intelligent Information System pp. 1–22. https://doi.org/10.1007/s10844-025-00925-9

Ghosal, D., Majumder, N., Poria, S., et al. (2019). 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. https://doi.org/10.18653/v1/D19-1015

Guo, Z., Jin, T., Zhao, Z. (2024). Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In: Ku LW, Martins A, Srikumar V (eds) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp 1726–1736, https://doi.org/10.18653/v1/2024.acl-long.94

Hu, D., Hou, X., Wei, L., et al. (2022). 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), pp. 7037–7041. https://doi.org/10.1109/ICASSP43922.2022.9747397

Hu, J., Liu, Y., Zhao, J., et al. (2021). 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, pp. 5666–5675. https://doi.org/10.18653/v1/2021.acl-long.440

Iandola, F., Moskewicz, M., Karayev, S., et al. (2014). Densenet: Implementing efficient convnet descriptor pyramids. https://doi.org/10.48550/arXiv.1404.1869

Lei, Y., Yang, D., Li, M., et al. (2023). Text-oriented modality reinforcement network for multimodal sentiment analysis from unaligned multimodal sequences. In: CICAI (2), pp. 189–200. https://doi.org/10.1007/978-981-99-9119-8_18

Li, D., Wang, Y., Funakoshi, K., et al. (2023). Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition. CoRR. https://doi.org/10.48550/ARXIV.2311.11009

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

Liu, Y., Ott, M., Goyal, N., et al. (2019). Roberta: A robustly optimized BERT pretraining approach. https://doi.org/10.48550/arXiv.1907.11692

Liu, Y. K., Cai, J., Lu, B. L., et al. (2025). Multi-to-single: Reducing multimodal dependency in emotion recognition through contrastive learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(2), 1438–1446. https://doi.org/10.1609/aaai.v39i2.32134

Ma, H., Wang, J., Lin, H., et al. (2024). A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia, 26*, 776–788. https://doi.org/10.1109/TMM.2023.3271019

Majumder, N., Poria, S., Hazarika, D., et al. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. In: Proceedings of the AAAI conference on artificial intelligence, pp 6818–6825.https://doi.org/10.1609/AAAI.V33I01.33016818

Meng, T., Zhang, F., Shou, Y., et al. (2024). Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 4298–4312. https://doi.org/10.1109/TASLP.2024.3434495

Nguyen, C.V.T., Nguyen, C.B., Le, D.T., et al .(2024a). Curriculum learning meets directed acyclic graph for multimodal emotion recognition. In: Calzolari N, Kan MY, Hoste V, et al (eds) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 4259–4265

Nguyen, C.V.T., Nguyen, C.B., Le, D.T., et al .(2024b). Curriculum learning meets directed acyclic graph for multimodal emotion recognition. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024), pp 4259–4265

Pham, H., Liang, P.P., Manzini, T., et al. (2019). Found in translation: Learning robust joint representations by cyclic translations between modalities. In: AAAI, pp. 6892–6899.https://doi.org/10.1609/AAAI.V33I01.33016892

Poria, S., Hazarika, D., Majumder, N., et al. (2019). 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, https://doi.org/10.18653/v1/P19-1050

Pujol, F.A., Mora, H., Martínez, A. (2019). Emotion recognition to improve e-healthcare systems in smart cities. In: Research & Innovation Forum 2019 - Technology, Innovation, Education, and their Social Impact, RIIFORUM 2019, Rome, Italy, April 24-26, 2019, pp. 245–254. https://doi.org/10.1007/978-3-030-30809-4_23

Shen, W., Wu, S., Yang, Y., et al. (2021). Directed acyclic graph network for conversational emotion recognition. 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. 1551–1560. https://doi.org/10.18653/v1/2021.acl-long.123

Song, Z., Hu, Z., Zhou, Y., et al. (2024). Embedded heterogeneous attention transformer for cross-lingual image captioning. *IEEE Transactions on Multimedia, 26*, 9008–9020. https://doi.org/10.1109/TMM.2024.3384678

Tsai, Y.H.H., Bai, S., Liang, P.P., et al. (2019). Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the conference. Association for computational linguistics. Meeting, p. 6558, https://doi.org/10.18653/V1/P19-1656

Tu, G., Wang, J., Li, Z., et al. (2024a). Multiple knowledge-enhanced interactive graph network for multimodal conversational emotion recognition. In: Findings of the association for computational linguistics: EMNLP 2024, pp 3861–3874, https://doi.org/10.18653/v1/2024.findings-emnlp.222

Tu, G., Xie, T., Liang, B., et al. (2024). Adaptive graph learning for multimodal conversational emotion detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(17), 19089–19097. https://doi.org/10.1609/aaai.v38i17.29876

Wang, P., Zhou, Q., Wu, Y., et al. (2024). DLF: disentangled-language-focused multimodal sentiment analysis. CoRR. https://doi.org/10.48550/ARXIV.2412.12225

Wang, Y., Cui, Z., Li, Y. (2023). Distribution-consistent modal recovering for incomplete multimodal learning. In: 2023 IEEE/CVF international conference on computer vision (ICCV), pp. 21968–21977.https://doi.org/10.1109/ICCV51070.2023.02013

Wang, Y., Fang, X., Yin, H., et al. (2025). Big-fusion: Brain-inspired global-local context fusion framework for multimodal emotion recognition in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(2), 1574–1582. https://doi.org/10.1609/aaai.v39i2.32149

Wu, J., Wu, J., Zheng, Y., et al. (2025). MLGAT: multi-layer graph attention networks for multimodal emotion recognition in conversations. *Journal of Intelligent Information System, 63*(2), 375–394. https://doi.org/10.1007/S10844-024-00879-4

Yang, H., Gao, X., Wu, J., et al. (2023). Self-adaptive context and modal-interaction modeling for multimodal emotion recognition. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Findings of the Association for Computational Linguistics: ACL 2023, pp. 6267–6281. https://doi.org/10.18653/v1/2023.findings-acl.390

Yang, Y., Dong, X., & Qiang, Y. (2025). Mse-adapter: A lightweight plugin endowing llms with the capability to perform multimodal sentiment analysis and emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(24), 25642–25650. https://doi.org/10.1609/aaai.v39i24.34755

Zhang, X., Li, Y. (2023). A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 13099–13110. https://doi.org/10.18653/v1/2023.acl-long.732

Zheng, X., Zhao, G., Zhu, L., et al. (2022). Perd: Personalized emoji recommendation with dynamic user preference. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp 1922–1926. https://doi.org/10.1145/3477495.3531779

Zhong, P., Wang, D., Miao, C. (2019). Knowledge-enriched transformer for emotion detection in textual 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), pp. 165–176. https://doi.org/10.18653/v1/D19-1016

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

Zou, S., Huang, X., Shen, X. (2023). Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5994–6003, https://doi.org/10.1145/3581783.3611805