Multi-modal Anchor Gated Transformer with Knowledge Distillation for Emotion Recognition in Conversation

Jie Li[1,2], **Shifei Ding**[1,2*], **Lili Guo**[1,2] and **Xuan Li**[1,2]
[1]School of Computer Science and Technology, China University of Mining and Technology
[2]Mine Digitization Engineering Research Center of Ministry of Education, China University of Mining and Technology
{jie_li, dingsf, liliguo, lixuan23}@cumt.edu.cn,

Abstract

Emotion Recognition in Conversation (ERC) aims to detect the emotions of individual utterances within a conversation. Generating efficient and modality-specific representations for each utterance remains a significant challenge. Previous studies have proposed various models to integrate features extracted using different modality-specific encoders. However, they neglect the varying contributions of modalities to this task and introduce high complexity by aligning modalities at the frame level. To address these challenges, we propose the Multi-modal Anchor Gated Transformer with Knowledge Distillation (MAGTKD) for the ERC task. Specifically, prompt learning is employed to enhance textual modality representations, while knowledge distillation is utilized to strengthen representations of weaker modalities. Furthermore, we introduce a multi-modal anchor gated transformer to effectively integrate utterance-level representations across modalities. Extensive experiments on the IEMOCAP and MELD datasets demonstrate the effectiveness of knowledge distillation in enhancing modality representations and achieve state-of-the-art performance in emotion recognition. Our code is available at: https://github.com/JieLi-dd/MAGTKD.

1 Introduction

Emotion plays a pivotal role in human communication, influencing not only the content but also the tone and context of interactions. Emotion Recognition in Conversation (ERC) aims to identify the emotional states expressed in each utterance within a dialogue. This task is essential for applications in areas such as conversational agents, healthcare systems, and recommendation engines. While emotions are traditionally expressed through text, they are also richly conveyed in the audio and visual modalities [Poria *et al.*, 2017; Wu *et al.*, 2025]. Figure 1 provides an illustrative example of multi-modal ERC, showcasing how information from various modalities can be integrated to improve emotion recognition.

Figure 1: Multi-modal conversation example from the MELD dataset.

Recent research in ERC has primarily focused on individual modalities. Text-based models often leverage context modeling [Majumder *et al.*, 2019; Song *et al.*, 2022; Hu *et al.*, 2023; Yang *et al.*, 2024] or incorporate external knowledge [Zhong *et al.*, 2019; Zhu *et al.*, 2021; Lee and Lee, 2022; Wang *et al.*, 2025]. Audio-based models make use of multi-task learning, attention mechanisms, and data augmentation strategies [Latif *et al.*, 2023; He *et al.*, 2025; Guo *et al.*, 2025], while video-based models extract key frames to enhance emotion recognition [Wei *et al.*, 2021; Poria *et al.*, 2017]. However, relying solely on a single modality for emotion recognition can overlook crucial emotional cues embedded in other modalities, leading to suboptimal performance. This limitation has spurred increasing interest in multi-modal ERC. Existing multi-modal models typically extract frame-level features for each modality, align these features across modalities, and fuse them for emotion classification [Tsai *et al.*, 2019; Guo *et al.*, 2022; Zheng *et al.*, 2023]. While effective, these approaches often treat all modalities equally, disregarding the varying significance of each modality in emotion recognition. Furthermore, the complex alignment process increases the computational burden, making these models less suitable for deployment in resource-constrained environments. To address these challenges, a more efficient and adaptive approach to modality representation and integration is needed.

Prompt-based learning has recently gained attention for its

*Corresponding Author.

success in both natural language processing (NLP) and multi-modal tasks. In ERC tasks focusing on textual data, well-designed prompts can effectively guide models to extract relevant contextual information, thereby improving the quality of utterance-level features [Song *et al.*, 2022; Son *et al.*, 2022; Yun *et al.*, 2024]. Additionally, knowledge distillation techniques have been widely explored to enhance the performance of student models by enabling them to mimic better the representations learned by teacher models [Lin *et al.*, 2022; Li *et al.*, 2023; Yun *et al.*, 2024; Ma *et al.*, 2024]. To overcome the challenges in multi-modal ERC, we introduce the Multi-modal Anchor Gated Transformer with Knowledge Distillation (MAGTKD), a framework designed to improve the integration of multi-modal information for emotion recognition tasks. Specifically, MAGTKD leverages context-aware prompts to extract high-quality utterance-level textual representations. These robust textual features are then used within a knowledge distillation framework to enhance the representation capacity of weaker modalities (e.g., audio and video), ultimately improving the fusion of multi-modal features for emotion recognition. In contrast to existing methods, which rely on frame-level feature interactions before fusion—thereby increasing computational complexity [Tsai *et al.*, 2019; Guo *et al.*, 2022; Zheng *et al.*, 2023]—MAGTKD directly fuses utterance-level features post-interaction, addressing the computational burden while maintaining or even improving performance.

We evaluate MAGTKD on two widely-used benchmark datasets, IEMOCAP and MELD. Experimental results demonstrate that MAGTKD achieves state-of-the-art performance on both datasets, surpassing existing methods in both accuracy and efficiency.

The key contributions of this work are as follows:

- We propose MAGTKD, a novel framework for ERC that effectively integrates multi-modal features, taking into account the varying contributions of different modalities to emotion classification.
- MAGTKD significantly reduces model complexity compared to traditional frame-level feature fusion methods.
- MAGTKD sets new benchmarks for ERC, achieving superior performance on the IEMOCAP and MELD datasets.

2 Related Works

2.1 Prompt Learning

Prompt Learning has emerged as an effective approach for leveraging pre-trained models by designing task-specific prompts to fine-tune and integrate them for downstream tasks, enabling improved modality representations. It has been widely adopted across various NLP tasks [Gao *et al.*, 2021; Heinzerling and Inui, 2021; Xu *et al.*, 2023]. Recently, researchers have begun exploring the application of prompt learning in multi-modal settings [Tsimpoukelli *et al.*, 2021; Khattak *et al.*, 2023; Zhu *et al.*, 2023]. [Tsimpoukelli *et al.*, 2021] presents a simple, yet effective, approach for transferring this few-shot learning ability to a multi-modal setting (vision and language). [Khattak *et al.*, 2023] proposes Multi-modal Prompt Learning (MaPLe) for both vision and language branches to improve alignment between the vision and language representations. [Zhu *et al.*, 2023] develop Visual Prompt multi-modal Tracking (ViPT), which learns the modal-relevant prompts to adapt the frozen pre-trained foundation model to various downstream multi-modal tracking tasks. With success in diverse NLP and multi-modal learning applications, we extend prompt learning to the emotion recognition task, aiming to harness its potential for enhancing emotional feature extraction and representation.

2.2 Konwledge Distillation

Knowledge Distillation (KD) aims to transfer knowledge from a large teacher network to a smaller student network. This knowledge transfer typically occurs at three levels: soft labels of the final layer [Hinton *et al.*, 2015], intermediate-layer features [Romero *et al.*, 2015], and the relationships between features across layers [Yim *et al.*, 2017]. Based on the learning strategy, KD can be categorized into offline [Passalis and Tefas, 2018; Li *et al.*, 2020] and online [Zhang *et al.*, 2018; Chung *et al.*, 2020] distillation. In offline distillation, the teacher model is pre-trained to guide the student model's learning. In contrast, online distillation involves simultaneous training of the teacher and student models with joint parameter updates. KD has demonstrated its effectiveness in transferring knowledge across modalities in multi-modal research [Albanie *et al.*, 2018]. Motivated by this, we adapt KD techniques to the multi-modal ERC task, enabling efficient knowledge transfer between modalities to enhance emotion recognition performance.

2.3 Modal Fusion

In the domain of modality fusion, existing works predominantly focus on extracting frame-level features and performing feature interactions at this granularity. [Tsai *et al.*, 2019] introduces the Multi-modal Transformer (MulT) to generically address the above issues in an end-to-end manner without explicitly aligning the data. [Zheng *et al.*, 2023] extracts three modal frame-level features and uses an attention mechanism to perform alignment operations on the three modal features. However, frame-level feature alignment often introduces significant computational complexity. Unlike these approaches, our work adopts utterance-level feature extraction and designs a novel model for multi-modal feature fusion, effectively reducing complexity while maintaining strong performance in emotion recognition tasks.

3 Methods

To enhance the representation of each modality and achieve effective multi-modal fusion, we propose the MAGTKD model for the ERC task. Figure 2 illustrates the overall architecture of the proposed framework, with detailed descriptions provided in the following subsections.

3.1 Task Definition

Given a set of speakers S, utterances U, and emotion labels Y, a conversation consisting of k utterances is represented as $[s_i, u_1, y_m, s_j, u_2, y_n, ..., s_i, u_k, y_m]$, where $s_i, s_j \in S$ are

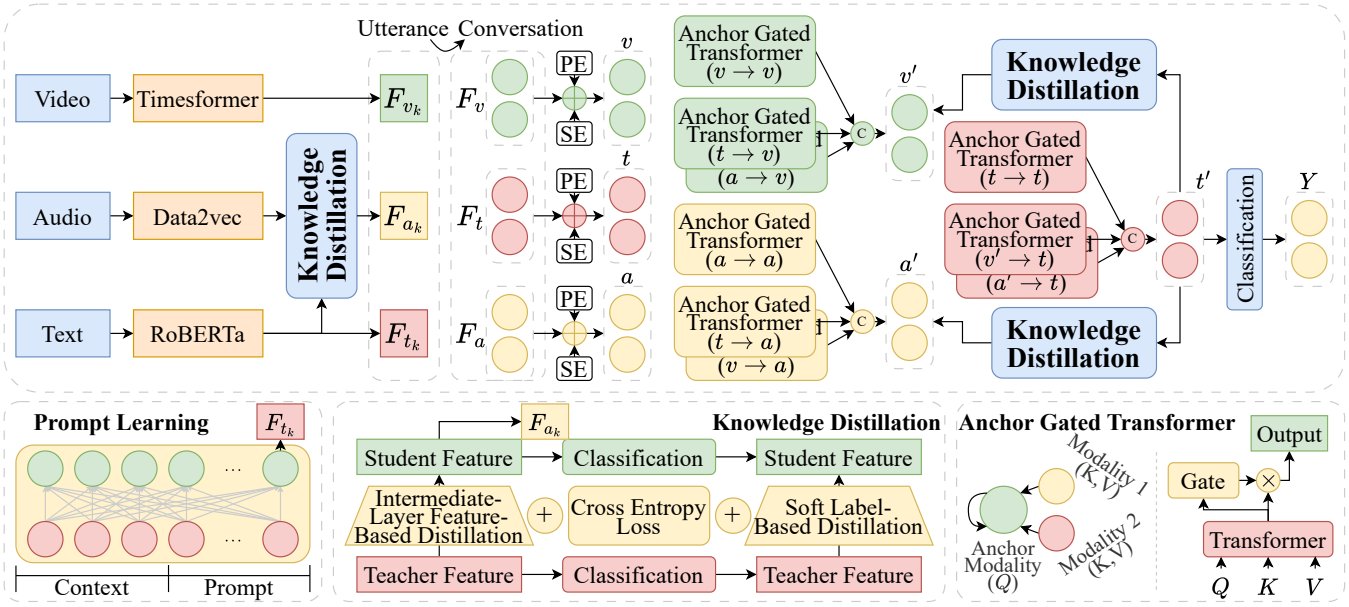

Figure 2: Illustration of the architecture of MAGTKD.

speakers and $y_m, y_n \in Y$ are one of the predefined emotion categories. If $i = j$, s_i and s_j represent the same speaker. Moreover, $u_k \in U$ represents the k-th utterance. Each utterance $u_k = \{t_k, a_k, v_k\}$ contains three modalities, where t, a, v represent text, audio and video, respectively. The goal of ERC is to predict to which emotion label y_m the utterance u_k belongs.

3.2 Feature Extraction

Figure 2 illustrates the extraction of modality-specific features using dedicated encoders for each input modality. In this section, we explain the feature generation process in detail:

Text: Following prior work [Lee and Lee, 2022; Song *et al.*, 2022], we utilize prompt learning to model both the context and speaker information. For the text encoder, we adopt RoBERTa [Liu *et al.*, 2019]. We construct the contextual representation C_k and the prompt P_k as follows.

$$C_k = Concat(s_i : t_1, s_j : t_2, \ldots, s_i : t_k) \quad (1)$$

$$P_k = For\ s_i : t_k\ Now\ s_i\ feels\ <mask> \quad (2)$$

$$F_{t_k} = Roberta(C_k\ </s>\ P_k) \quad (3)$$

where $<mask>$ represents a special token, $F_{t_k} \in R^{1\times d}$ is embedding of $<mask>$, representing the aggregated emotion feature, and d is the hidden dimension of the $<mask>$ token.

Audio: Self-supervised learning has achieved remarkable success not only in natural language processing but also in audio and video domains [Baevski *et al.*, 2020; Baevski *et al.*, 2022]. For the audio encoder, we employ Date2vec [Baevski *et al.*, 2022], with the audio segment a_k of the k-th utterance as input. The process of extracting audio features is formalized as follows.

$$F_{a_k} = Data2vec(a_k) \quad (4)$$

where $F_{a_k} \in R^{1\times d}$ is the embedding of a_k, and d is the hidden dimension of audio features.

Video: Similar to the audio feature extraction process, we utilize Timesformer [Bertasius *et al.*, 2021] as the video encoder. The process of extracting video features is formalized as follows.

$$F_{v_k} = Timesformer(v_k) \quad (5)$$

where v_k is the video input, and $F_{v_k} \in R^{1\times d}$ is the embedding of v_k, and d is the hidden dimension of video features.

3.3 Knowledge Distillation

Unlike traditional knowledge distillation methods that utilize KL divergence, the multi-modal ERC task involves cross-modal knowledge distillation. We adopt a collaborative distillation strategy based on soft labels and intermediate-layer features, using Pearson correlation coefficients as our cross-modal measurement approach.

$$d(u, v) = 1 - p(u, v) \quad (6)$$

where $p(u, v)$ is the Pearson correlation coefficient between two logit vectors u and v.

Soft Label-Based Distillation leverages the soft label outputs from the last layer of each modality encoder to compute the knowledge divergence across modalities at both the sample and feature levels. Pearson correlation coefficients are employed to measure the degree of knowledge disparity between different modalities. By reducing this disparity, the textual features transfer knowledge to the audio features. The process is formalized as:

$$Y^t_{i,:} = softmax(P^t_{i,:}/\tau) \quad (7)$$

$$Y^a_{i,:} = softmax(P^a_{i,:}/\tau) \quad (8)$$

$$L_s = \frac{\tau^2}{B}\sum_{i=1}^{B} d(Y^a_{i,:}, Y^t_{i,:}) + \frac{\tau^2}{C}\sum_{j=1}^{C} d(Y^a_{:,j}, Y^t_{:,j}) \quad (9)$$

where B is a training batch, C is the emotion categories, $P^t, P^a \in R^{B\times C}$ are the prediction matrix of text and audio modality, respectively. τ is a temperature parameter to control the softness of logits.

Intermediate-Layer Feature-Based Distillation computes similarity matrices within a batch for the textual modality as the target matrix (via dot product between text modality features and their transpose). Similarly, a source matrix is computed for the audio and text modalities. Using the softmax function, we derive the target and source distributions. The KL divergence between these distributions is minimized to enable the transfer of knowledge from textual features to audio features. The process is defined as:

$$T_i = \frac{exp(F_{i,j}/\tau)}{\sum_{s=1}^{B} exp(F_{i,s})}, \forall i,j \in B \tag{10}$$

$$S_i = \frac{exp(F^{'}_{i,j}/\tau)}{\sum_{s=1}^{B} exp(F^{'}_{i,s})}, \forall i,j \in B \tag{11}$$

$$L_f = \frac{1}{B}\sum_{i=1}^{B} KL(T_i||S_i) \tag{12}$$

where $F_{i,j}, F^{'}_{i,j} \in R^{B\times B}$ are the text-modal similarity matrix and the text-audio modal similarity matrix, respectively. T_i, S_i are target and source distributions.

The overall loss function includes the above two losses and the cross-entropy loss:

$$L_{CE} = -\frac{1}{B}\sum_{i=1}^{B} y_i \cdot \log p_i \tag{13}$$

$$L_{all} = L_{CE} + L_s + L_f \tag{14}$$

where y_i is true labels and p_i is predict labels.

3.4 Multi-modal Anchor Gated Transformer

In the first stage, we utilize prompt learning and knowledge distillation to extract utterance-level features for each modality. However, directly concatenating these features for emotion recognition, in figure 3, results in degraded model performance. To address this issue, we propose a second stage that employs a Multi-modal Anchor Gated Transformer (MAGT) to effectively integrate features across the three modalities. Specifically, each modality serves as an anchor to aggregate complementary information from other modalities. Specifically, we first use the audio and video features as anchors to aggregate information from the other modalities. Given the strong performance of the text modality, the raw text features are used as an anchor to aggregate the audio and video features enriched by other modalities.

We construct the dataset at the conversation level, where utterance-level features from different modalities are arranged sequentially based on temporal order. For each utterance, speaker and positional embeddings are added. Speaker embedding uniquely maps each speaker in the dataset to a sequence ID, which is then embedded using an embedding layer.

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d_{model}}}) \tag{15}$$

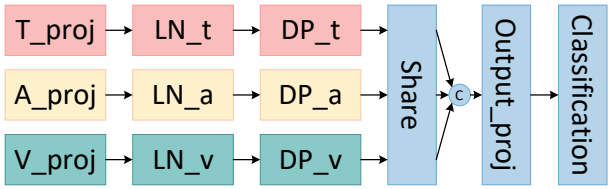

Figure 3: Architecture of the Concat model. Each modality passes through a linear layer, followed by Layer Normalization (LN) and Dropout (DP). The outputs are then processed by a shared weights module before classification through a final linear layer.

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d_{model}}}) \tag{16}$$

$$SE = Embedding[s_{u_1}, s_{u_2}, \ldots, s_{u_k}] \tag{17}$$

$$H_m = F_m + PE + SE \tag{18}$$

where PE is positional embedding and SE is speaker embedding. s_{u_k} represents speaker of utterance u_k. $F_m = (F_{m_1}, F_{m_2}, \ldots, F_{m_k})$ and $m \in \{t,a,v\}$. H_m is the positional- and speaker-aware utterance sequence representation for m modality.

Given the remarkable success of Transformers in NLP, the effectiveness of gating mechanisms for preserving salient features, and the superior performance of the text modality compared to audio and video in ERC tasks, we introduce an Anchor-Gated Transformer. This structure leverages a Transformer encoder with three inputs (Query $Q \in R^{l_q\times d_k}$, Key $K \in R^{l_k\times d_k}$, and Value $V \in R^{l_k\times d_k}$) to integrate features across modalities.

$$H_{m\to m} = Transformer(H_m, H_m, H_m) \tag{19}$$

$$H_{n\to m} = Transformer(H_m, H_n, H_n) \tag{20}$$

where $m \in \{t,a,v\}$ and $n \in \{t,a,v\} - \{m\}$. $H_{m\to m}$ represents anchor modality aggregating its own information. $H_{n\to m}$ represents an anchor modality aggregating information from other modalities.

To enhance the emotional representation of each modality, we incorporate a gating mechanism to filter out irrelevant information and retain the most effective emotional features.

$$\alpha_{n\to m} = \sigma(W_{n\to m} \cdot H_{n\to m} + b_{n\to m}) \tag{21}$$

$$H^{'}_{n\to m} = H_{n\to m} \otimes \alpha_{n\to m} \tag{22}$$

where $W_{n\to m}$ is a weight matrix, $b_{n\to m}$ is a bias parameter, $\alpha_{n\to m}$ represents gate, and $\otimes$ is the element-wise product.

3.5 Emotion Classifier

The emotion classification task is performed using a linear layer. The features t', extracted from the Multi-modal Attention and Graph-based Transformer (MAGT), are transformed into the emotion label p_i corresponding to each utterance u_i.

The classification process can be formalized as follows:

$$p_i = \text{argmax}(\text{softmax}(W \cdot t' + b)), \tag{23}$$

where $W \in \mathbb{R}^{C\times d}$ and $b \in \mathbb{R}^C$ are the weight matrix and bias vector of the linear layer, respectively. Here, C denotes the

number of emotion classes, and d represents the dimension of the feature vector t'. The softmax function ensures that the outputs are normalized probabilities across all emotion classes, and argmax selects the class with the highest probability as the predicted label p_i.

3.6 Training

In the first stage, the utterance-level feature extraction is optimized using the loss function defined in Equation 14. In the second stage, the multi-modal fusion is performed using the knowledge distillation (KD) loss functions defined in Equations 9 and 12. These can be formalized as:

$$L_{KD}^{i} = L_f + L_s \tag{24}$$

where i refers to either the audio or video modality.

The total loss function in the second stage is the sum of the cross-entropy loss and the two distillation loss functions, weighted by their respective coefficients. This can be expressed as:

$$L_{\text{total}} = L_{\text{CE}} + \alpha L_{\text{KD}}^{\text{a}} + \beta L_{\text{KD}}^{\text{v}} \tag{25}$$

where α and β are the coefficients for the distillation losses of the audio and video modalities, respectively.

4 Experiments

4.1 Datasets

In this section, we introduce two widely adopted benchmark datasets: MELD and IEMOCAP. Following descriptions for specific details of these two datasets:

MELD [Poria *et al.*, 2019] is a multiparty conversation dataset containing over 1,400 dialogues and more than 13,000 utterances extracted from the TV show "Friends." This dataset includes seven emotion categories: neutral, surprise, fear, sadness, joy, disgust, and anger.

IEMOCAP [Busso *et al.*, 2008] consists of 7,433 utterances and 151 dialogues, divided into five sessions, each involving two speakers. Each utterance is labeled with one of six emotion categories: happiness, sadness, anger, excitement, frustration, and neutral. The training and development datasets are randomly split from the first four sessions in a 9:1 ratio. The test dataset comprises the last session.

4.2 Experimental Setup

We evaluate the performance of our model on two datasets using Accuracy (Acc) and Weighted F1-score (W_F1) as metrics. We designed a two-stage experimental process. The first stage focuses on extracting utterance-level features from different modalities, while the second stage involves multi-modal feature fusion using a dataset constructed at the conversation level. The hyperparameter settings are shown in Table 1. All experiments are conducted on a single NVIDIA GeForce RTX 2080 Ti GPU.

Stage	Parameter	IEMOCAP	MELD
Feature Extraction	Dimensions	768	768
	Learning rate	1e-5	1e-5
	Batch	4	4
	Epochs	10	10
Multi-modal Fusion	Learning rate	1e-5	1e-4
	Batch	16	16
	Epochs	30	30
	α, β	$0.7, 0.8$	$0.01, 0.09$

Table 1. Hyperparameters used in the experiments.

Model	IEMOCAP		MELD	
	Acc	W_F1	Acc	W_F1
DialogueRNN	63.4	62.75	60.31	57.66
DialogueGCN	65.25	64.18	-	58.1
MMGCN	66.22	-	58.65	-
DialogueCRN	66.05	66.2	60.73	58.39
A-DMN	64.6	64.3	-	<u>60.45</u>
DialogueINAB	67.32	67.22	60.52	57.78
SACCMA	67.41	67.1	62.3	59.3
Ada2I	68.76	<u>68.97</u>	<u>63.03</u>	60.38
GraphCFC	<u>69.13</u>	68.91	61.42	58.86
Ours	**69.38**	**69.59**	**66.36**	**65.32**

Table 2. Performance Comparison on IEMOCAP and MELD. Best results are in bold, second-best are underlined.

4.3 Baselines

We compare our proposed model, MAGTKD, against classic baselines, including DialogueRNN [Majumder *et al.*, 2019], DialogueGCN [Ghosal *et al.*, 2019], MMGCN [Hu *et al.*, 2021b], and DialogueCRN [Hu *et al.*, 2021a], as well as state-of-the-art models such as A-DMN [Xing *et al.*, 2022], DialogueINAB [Ding *et al.*, 2023], SACCMA [Guo *et al.*, 2024], Ada2I [Nguyen *et al.*, 2024], and GraphCPC [Li *et al.*, 2024].

4.4 Comparative Experiments

Table 2 compares our model with prior works on IEMOCAP and MELD. Our model achieves the best performance on both datasets, setting new state-of-the-art (SOTA) results. On IEMOCAP, we achieve 69.38% accuracy and 69.59% weighted F1 (W_F1), outperforming GraphCFC by 0.99% and 1.56%, respectively. On MELD, our model achieves 66.36% accuracy and 65.32% W_F1, with improvements of 5.17% and 6.83% over Ada2I, the previous SOTA. These results demonstrate our model's ability to effectively integrate multi-modal features and handle challenges in conversational emotion recognition through prompt learning, knowledge distillation, and advanced fusion techniques.

4.5 Visualization and Analysis

Figure 4 shows t-SNE visualizations of the feature representations from the IEMOCAP and MELD datasets. We visualize single-modal features (text, audio, and video) as well as features enhanced by knowledge distillation, where the text modality guides the audio and video modalities.

The visualizations indicate that the text modality has the strongest discriminative power across both datasets, followed by audio, while video shows the least discriminative ability. After applying knowledge distillation, the audio modality improves significantly, benefiting from the text modality's guid-

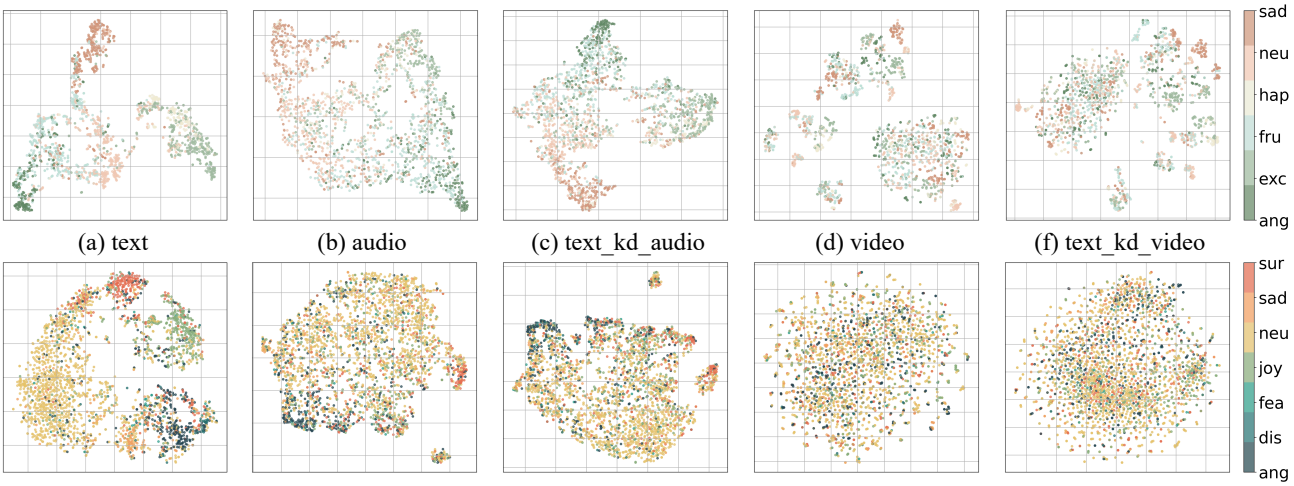

Figure 4: t-SNE visualization of feature representations for IEMOCAP (top row) and MELD (bottom row) datasets. "kd" refers to knowledge distillation.

ance. However, the video modality struggles to learn effectively, highlighting the challenge of transferring knowledge to modalities with weaker feature representations.

Furthermore, the audio modality in IEMOCAP shows stronger discriminative power compared to MELD, which results in better performance when learning from the text modality. In the next sections, we will provide quantitative results to further explore this trend.

4.6 Ablation Studies

Table 3 shows the results of ablation studies on feature extraction and multi-modal fusion. We evaluate individual modalities and the effects of knowledge distillation from the text modality to audio and video modalities.

The results highlight that knowledge distillation improves the performance of the audio modality but has little effect on the video modality. This aligns with the observations in Figure 4, where the video modality struggles to learn effectively from the text modality due to its weaker feature representation. On the other hand, the audio modality benefits more from the distillation process, particularly in the IEMOCAP dataset where it has a stronger feature representation compared to MELD.

For multi-modal fusion, we first tested the Concat method. The best performance was achieved by fusing the strongest modality features (text, distilled audio, and undistilled video). However, adding the distilled video modality degraded performance. We then tested the MAGT fusion method, which also performed best when fusing the strongest modalities. Notably, MAGT maintained stable performance even when the weakest modality, the distilled video, was included.

These results demonstrate that MAGT effectively integrates emotional cues from different modalities, even when some modalities contribute less useful information.

4.7 Complexity Analysis

We define the feature dimensions of the prior frame-level model as (S, L, D), and for the proposed utterance-level model as (C, U, D), where:

- S: number of samples,
- L: frame sequence length,
- D: hidden feature dimension,
- C: number of conversations,
- U: number of utterances per conversation.

Assuming $C \cdot U = S$ and L is consistent across modalities.

Spatial Complexity The frame-level model has spatial complexity:

$$O(S \cdot L \cdot D) \quad \text{(linear with respect to } S \text{ and } L\text{)}.$$

while the utterance-level model has:

$$O(C \cdot U \cdot D) \quad \text{(linear with respect to } C \text{ and } U\text{)}.$$

which reduces by a factor of L since $C \cdot U = S$.

Temporal Complexity The frame-level model's temporal complexity is:

$$O(S \cdot L^2 \cdot D) \quad \text{(quadratic with respect to } L\text{)}.$$

whereas the utterance-level model's complexity is:

$$O(C \cdot U^2 \cdot D) \quad \text{(quadratic with respect to } U\text{)}.$$

Thus, the relative temporal complexity is:

$$\frac{O(C \cdot U^2 \cdot D)}{O(S \cdot L^2 \cdot D)} = \frac{U}{L^2}.$$

For shorter dialogues ($U \ll L$), the proposed model has a significant reduction in complexity. Additionally, spatial complexity is reduced by a factor of L. In summary, the proposed model is more efficient, with linear spatial and quadratic temporal complexity in U, compared to the frame-level model's quadratic temporal complexity in L.

Module	IEMOCAP		MELD	
	Acc	W_F1	Acc	W_F1
Feature Extraction				
T	67.09	67.46	62.79	62.99
A	47.01	45.91	50.38	44.80
V	27.84	26.28	40.91	36.84
A_{KD}	50.03	49.65	49.08	45.69
V_{KD}	25.63	20.21	40.45	36.06
Multi-modal Fusion				
Concat				
T+A+V	68.52	68.64	65.71	65.06
T+A+V_{KD}	64.70	64.38	60.46	55.74
T+A_{KD}+V	68.64	68.70	65.86	65.18
T+A_{KD}+V_{KD}	65.37	65.19	60.61	55.67
MAGT				
T+A+V	68.08	68.29	66.32	65.30
T+A+V_{KD}	68.08	68.18	65.79	64.73
T+A_{KD}+V	**69.38**	**69.59**	**66.36**	**65.32**
T+A_{KD}+V_{KD}	68.88	69.02	65.79	64.73

Table 3. Ablation studies on different modalities and fusion methods for IEMOCAP and MELD. "KD" indicates knowledge distillation, "Concat" is the simple fusion method, and "MAGT" is our proposed fusion method. Best results are in bold.

4.8 Hyper-parametric Analysis

Figure 5 shows the effect of varying the hyperparameters α (audio distillation coefficient) and β (video distillation coefficient) on model performance for IEMOCAP and MELD. For IEMOCAP, changing α significantly affects performance when β is fixed, while adjusting β with a fixed α results in smaller variations. This suggests that the audio modality better benefits from knowledge distillation, whereas the video modality shows weaker learning. This aligns with the observations in Figure 4, where the audio modality has better feature discriminability than the video modality. In MELD, setting $\alpha = 0$ and increasing β initially improves performance, but further increases lead to a decline. Similarly, setting $\beta = 0$ and varying α shows an initial performance boost

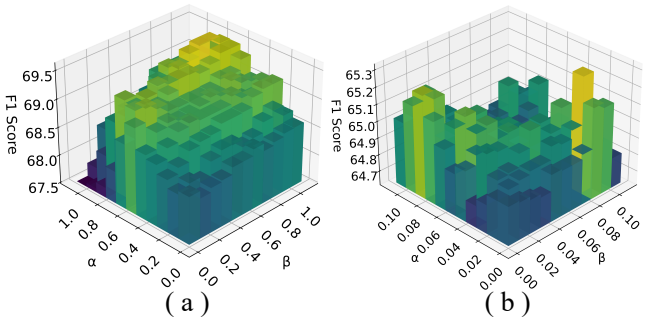

Figure 5: (a) Hyperparametric analysis for IEMOCAP, (b) Hyperparametric analysis for MELD. Here, α and β are the coefficients for audio and video knowledge distillation losses, respectively.

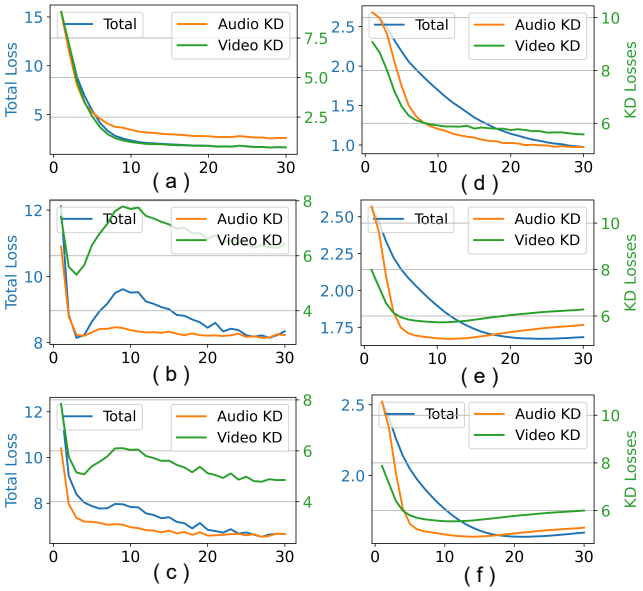

Figure 6: (a)-(b) show the variations in Total Loss and Knowledge Distillation (KD) Loss over training epochs for IEMOCAP across training, validation, and test sets. (d)-(f) show the same for MELD.

followed by a decrease, confirming that the KD loss improves model performance but requires careful tuning.

4.9 Convergence Analysis

Figure 6 presents the convergence behavior of the model on IEMOCAP and MELD datasets. For IEMOCAP (a)-(c), as the number of epochs increases, both the total loss and the KD loss for audio and video modalities converge. The KD loss for the audio modality converges steadily, while the video modality experiences fluctuations before stabilizing, which is consistent with the lower discriminability of video features, as shown in Figure 4. Similar patterns are observed for the validation and test sets. For MELD (d)-(f), the trends are similar, with both the total loss and the KD losses for audio and video modalities converging as epochs increase. These results confirm the effectiveness of our model and the positive impact of knowledge distillation on training stability and performance.

5 Conclusion

The proposed MAGTKD model effectively addresses the challenges of multi-modal ERC by leveraging prompt learning to extract robust textual representations and employing knowledge distillation to enhance weaker modalities. The subsequent use of MAGT enables efficient aggregation of emotional information across modalities, resulting in state-of-the-art performance on both the MELD and IEMOCAP datasets. Future work will explore extending MAGTKD to handle more complex multi-modal scenarios, such as incorporating dynamic contextual information in real-time conversations or addressing challenges posed by highly imbalanced datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant no.62276265 and 62406326).

References

[Albanie *et al.*, 2018] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *ACM MM*, pages 292–301, 2018.

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NIPS*, volume 33, pages 12449–12460, 2020.

[Baevski *et al.*, 2022] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, pages 1298–1312, 2022.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[Busso *et al.*, 2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.*, 42:335–359, 2008.

[Chung *et al.*, 2020] Inseop Chung, SeongUk Park, Jangho Kim, and Nojun Kwak. Feature-map-level online adversarial knowledge distillation. In *ICML*, pages 2006–2015, 2020.

[Ding *et al.*, 2023] Junyuan Ding, Xiaoliang Chen, Peng Lu, Zaiyan Yang, Xianyong Li, and Yajun Du. Dialogueinab: an interaction neural network based on attitudes and behaviors of interlocutors for dialogue emotion recognition. *J. Supercomput.*, 79(18):20481–20514, 2023.

[Gao *et al.*, 2021] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL/IJCNLP*, pages 3816–3830, 2021.

[Ghosal *et al.*, 2019] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP/IJCNLP*, pages 154–164, 2019.

[Guo *et al.*, 2022] Lili Guo, Longbiao Wang, Jianwu Dang, Yahui Fu, Jiaxing Liu, and Shifei Ding. Emotion recognition with multimodal transformer fusion framework based on acoustic and lexical information. *IEEE MultiMedia*, 29(2):94–103, 2022.

[Guo *et al.*, 2024] Lili Guo, Yikang Song, and Shifei Ding. Speaker-aware cognitive network with cross-modal attention for multimodal emotion recognition in conversation. *Knowl.-Based Syst.*, 296:111969, 2024.

[Guo *et al.*, 2025] Lili Guo, Jie Li, Shifei Ding, and Jianwu Dang. Apin: Amplitude- and phase-aware interaction network for speech emotion recognition. *Speech Commun.*, 169:103201, 2025.

[He *et al.*, 2025] Dongxiao He, Yongqi Huang, Jitao Zhao, Xiaobao Wang, and Zhen Wang. Str-gcl: Structural commonsense driven graph contrastive learning. In *WWW*, page 1129–1141, 2025.

[Heinzerling and Inui, 2021] Benjamin Heinzerling and Kentaro Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *EACL*, pages 1772–1791, 2021.

[Hinton *et al.*, 2015] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[Hu *et al.*, 2021a] Dou Hu, Lingwei Wei, and Xiaoyong Huai. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *ACL/IJCNLP*, pages 7042–7052, 2021.

[Hu *et al.*, 2021b] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *ACL/IJCNLP*, pages 5666–5675, 2021.

[Hu *et al.*, 2023] Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. Supervised adversarial contrastive learning for emotion recognition in conversations. In *ACL*, pages 10835–10852, 2023.

[Khattak *et al.*, 2023] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023.

[Latif *et al.*, 2023] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W. Schuller. Multitask learning from augmented auxiliary data for improving speech emotion recognition. *IEEE Trans. Affective Comput.*, 14(4):3164–3176, 2023.

[Lee and Lee, 2022] Joosung Lee and Wooin Lee. CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *NAACL*, pages 5669–5679, 2022.

[Li *et al.*, 2020] Tianhong Li, Jianguo Li, Zhuang Liu, and Changshui Zhang. Few sample knowledge distillation for efficient network compression. In *CVPR*, pages 14639–14647, 2020.

[Li *et al.*, 2023] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, volume 37, pages 1504–1512, 2023.

[Li *et al.*, 2024] Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Trans. Multimed.*, 26:77–89, 2024.

[Lin *et al.*, 2022] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang

Wang. Knowledge distillation via the target-aware transformer. In *CVPR*, pages 10915–10924, 2022.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[Ma *et al.*, 2024] Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Trans. Multimed.*, 26:776–788, 2024.

[Majumder *et al.*, 2019] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *AAAI*, volume 33, pages 6818–6825, 2019.

[Nguyen *et al.*, 2024] Cam-Van Thi Nguyen, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. Ada2i: Enhancing modality balance for multimodal conversational emotion recognition. In *ACM MM*, page 9330–9339, 2024.

[Passalis and Tefas, 2018] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, pages 268–284, 2018.

[Poria *et al.*, 2017] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *ACL*, pages 873–883, 2017.

[Poria *et al.*, 2019] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, pages 527–536, 2019.

[Romero *et al.*, 2015] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

[Son *et al.*, 2022] Junyoung Son, Jinsung Kim, Jungwoo Lim, and Heuiseok Lim. GRASP: Guiding model with RelAtional semantics using prompt for dialogue relation extraction. In *COLING*, pages 412–423, 2022.

[Song *et al.*, 2022] Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. Supervised prototypical contrastive learning for emotion recognition in conversation. In *EMNLP*, pages 5197–5206, 2022.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569, 2019.

[Tsimpoukelli *et al.*, 2021] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NIPS*, volume 34, pages 200–212, 2021.

[Wang *et al.*, 2025] Xiaobao Wang, Yujing Wang, Dongxiao He, Zhe Yu, Yawen Li, Longbiao Wang, Jianwu Dang, and Di Jin. Elevating knowledge-enhanced entity and relationship understanding for sarcasm detection. *IEEE Trans. Knowl. Data Eng.*, 37(6):3356–3371, 2025.

[Wei *et al.*, 2021] Jie Wei, Xinyu Yang, and Yizhuo Dong. User-generated video emotion recognition based on key frames. *Multimed. Tools Appl.*, 80(9):14343–14361, 2021.

[Wu *et al.*, 2025] Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. *AAAI*, 39(2):1601–1609, 2025.

[Xing *et al.*, 2022] Songlong Xing, Sijie Mai, and Haifeng Hu. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Trans. Affect. Comput.*, 13(3):1426–1439, 2022.

[Xu *et al.*, 2023] Yige Xu, Zhiwei Zeng, and Zhiqi Shen. Efficient cross-task prompt tuning for few-shot conversational emotion recognition. In *EMNLP*, pages 11654–11666, 2023.

[Yang *et al.*, 2024] Kailai Yang, Tianlin Zhang, and Sophia Ananiadou. Disentangled variational autoencoder for emotion recognition in conversations. *IEEE Trans. Affective Comput.*, 15(2):508–518, 2024.

[Yim *et al.*, 2017] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, pages 4133–4141, 2017.

[Yun *et al.*, 2024] Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. TelME: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *NAACL*, pages 82–95, 2024.

[Zhang *et al.*, 2018] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018.

[Zheng *et al.*, 2023] Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In *ACL*, pages 15445–15459, 2023.

[Zhong *et al.*, 2019] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations. In *EMNLP/IJCNLP*, pages 165–176, 2019.

[Zhu *et al.*, 2021] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *ACL/IJCNLP*, pages 1571–1582, 2021.

[Zhu *et al.*, 2023] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *CVPR*, pages 9516–9526, 2023.