TelME: Teacher-leading Multimodal Fusion Network for Emotion Recognition in Conversation

Anonymous ACL submission

Abstract

Emotion Recognition in Conversation (ERC) plays a crucial role in enabling dialogue systems to effectively respond to user requests. The emotions in a conversation can be identified by the representations from various modalities, such as audio, visual, and text. However, due to the weak contribution of non-verbal modalities to recognize emotions, multimodal ERC has always been considered a challenging task. In this paper, we propose Teacher-leading Multimodal fusion network for ERC (TelME). TelME incorporates cross-modal knowledge distillation to transfer information from a language model acting as the teacher to the non-verbal students, thereby optimizing the efficacy of the weak modalities. We then combine multimodal features using a shifting fusion approach in which student networks support the teacher. TelME achieves state-of-the-art performance in MELD, a multi-speaker conversation dataset for ERC. Finally, we demonstrate the effectiveness of our components through additional experiments.

1 Introduction

Emotion recognition holds paramount importance, enhancing the engagement of conversations by providing appropriate responses to the emotions of users in dialogue systems (Ma et al., 2020). The application of emotion recognition spans various domains, including chatbots, healthcare systems, and recommendation systems, demonstrating its versatility and potential to enhance a wide range of applications (Poria et al., 2019). Emotion Recognition in Conversation (ERC) aims to identify emotions expressed by participants at each turn within a conversation. The dynamic emotions in a conversation can be detected through multiple modalities such as textual utterances, facial expressions, and acoustic signals (Baltrušaitis et al., 2018; Liang et al., 2022; Majumder et al., 2019; Hu et al., 2022b; Chudasama et al., 2022). Figure 1 illustrates an example of a multimodal ERC.

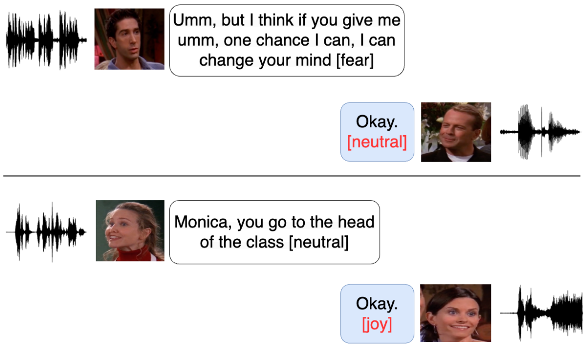
Umm, but I think if you give me umm, one chance I can, I can change your mind [fear]
Okay. [neutral]
Monica, you go to the head of the class [neutral]
Okay. [joy]

Figure 1: Examples of multimodal ERC. Even the same "Okay" answer varies depending on the conversation situation and captures emotions in various modalities.

Much research on ERC has mainly focused on context modeling from text modality, disregarding the rich representations that can be obtained from audio and visual modalities. Text-based ERC methods have demonstrated that contextual information derived from text data is a powerful resource for emotion recognition (Kim and Vossen, 2021; Lee and Lee, 2021; Song et al., 2022a,b). However, non-verbal cues such as facial expressions and tone of voice, which are not covered by text-based methods, provide important information that needs to be explored in the field of ERC. Multimodal approaches demonstrate the possibility of integrating features from three modalities to improve the robustness of ERC systems (Mao et al., 2021; Chudasama et al., 2022; Hu et al., 2022b). Nevertheless, these frameworks frequently ignore the varying degrees of impact the individual modalities have on emotion recognition and instead treat them as homogeneous components. This implies a promising opportunity to improve the ERC system by differentiating the level of contribution made by each modality.

In this paper, we propose Teacher-leading Multimodal fusion network for the ERC task (TelME) that strengthens and fuses multimodal information by accentuating the powerful modality while bolstering the weak modalities. Knowledge Distillation (KD) can be extended to transfer knowledge across modalities, where a powerful modality can play the role of a teacher to transfer knowledge to a weak modality (Hinton et al., 2015; Xue et al., 2022). As shown in Figure 2, in ERC tasks, information from the text modality is stronger compared to the other two modalities. Thus TelME enhances the representations of the two weak modalities through KD utilizing the text encoder as the teacher. The comparison experiments in Appendix A.1 also support our decision to set the text model as the teacher. Our approach aims to adjust the discrepancies in predictions between the teacher and non-verbal students while generating emotional features that are exceptionally well-suited for fusion. TelME then incorporates Attention-based modality Shifting Fusion, where the student networks strengthened by the teacher at the distillation stage assist the robust teacher encoder in reverse. Specifically, our fusion method creates displacement vectors from non-verbal modalities, which are used to shift the emotional embeddings of the teacher.

We conduct experiments on two widely used benchmark datasets and compare our proposed method with existing ERC methods. Our results show that TelME performs well on both datasets and particularly excels in multi-party conversations, achieving state-of-the-art performance. The ablation study also demonstrates the effectiveness of our knowledge distillation strategy and its interaction with our fusion method. Our contributions can be summarized as follows:

- We propose Teacher-leading Multimodal fusion network for Emotion Recognition in Conversation (TelME). The proposed method considers different contributions of text and non-verbal modalities to emotion recognition for better prediction.
- To the best of our knowledge, we are the first to enhance the effectiveness of weak non-verbal modalities for the ERC task through cross-modal distillation.
- TelME shows exceptional performance in two widely used benchmark datasets and especially achieves state-of-the-art in multi-party conversational scenarios.

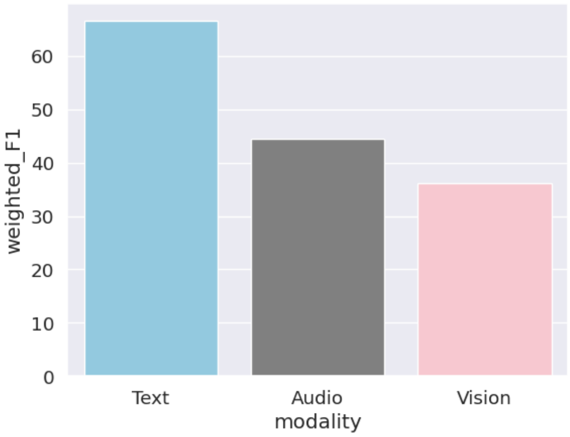

Figure 2: Unimodal Performance on MELD dataset

2 Related Work

2.1 Emotion Recognition in Conversation

Recently, ERC has become an increasingly attention in the field of emotion analysis. ERC can be categorized into text-based and multimodal methods, depending on the input format. Text-based methods primarily focus on context modeling and speaker relationships (Jiao et al., 2019; Li et al., 2020; Hu et al., 2021a). In recent studies (Lee and Lee, 2021; Song et al., 2022a), context modeling has been carried out to enhance the understanding of contextual information by pre-trained language models using dialogue-level input compositions. Additionally, there are graph-based approaches (Zhang et al., 2019; Ishiwatari et al., 2020; Shen et al., 2021; Ghosal et al., 2019) and approaches that utilize external knowledge (Zhong et al., 2019; Ghosal et al., 2020; Zhu et al., 2021).

Multimodal methods (Poria et al., 2017; Hazarika et al., 2018a,b; Majumder et al., 2019) reflect dialogue-level multimodal features through recurrent neural network-based models. Other multimodal approaches (Mao et al., 2021; Chudasama et al., 2022) integrate and manipulate utterance-level features through hierarchical structures to extract dialogue-level features from each modality. EmoCaps (Li et al., 2022) considers both multimodal information and contextual emotional tendencies to predict emotions. UniMSE (Hu et al., 2022b) proposes a framework that leverages complementary information between Multimodal Sentiment Analysis and ERC. Unlike these methods, our proposed TelME is a method in which the strong teacher leads emotion recognition while reinforcing features from weak modalities to support the

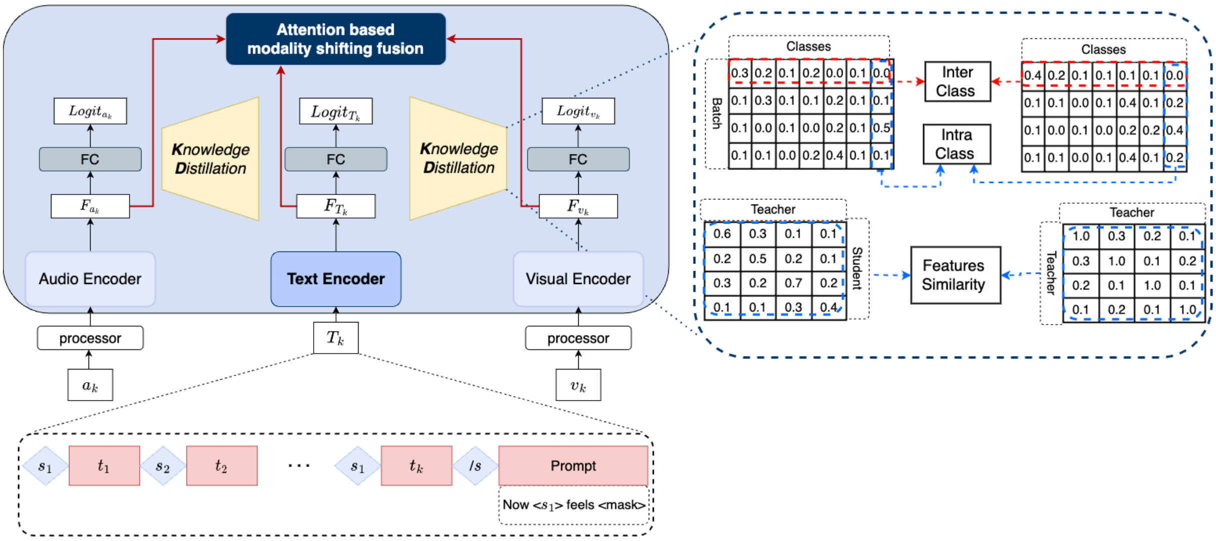

Figure 3: The overview of TelME

teacher.

2.2 Knowledge Distillation

The initial proposition of KD (Hinton et al., 2015) involves transferring knowledge by reducing the KL divergence between the prediction logits of teachers and students. Subsequently, KD method has been extended to distillation between intermediate features (Heo et al., 2019). KD approaches (Gupta et al., 2016; Jin et al., 2021; Tran et al., 2022) are also utilized for transferring knowledge between modalities in multimodal studies. Li et al. (2023b) mitigate multimodal heterogeneity by constructing dynamic graphs in which each vertex exhibits modality and each edge exhibits dynamic KD. However, since this work is not a study of the ERC task and is based on graph distillation, there is an intrinsic difference from our KD strategy. Ma et al. (2023) proposes a transformer-based model utilizing self-distillation for ERC. Our proposed method, in contrast, uses response and feature-based distillation at the same time to focus on maximizing the effectiveness of two other modalities by the teacher network based on text modality.

3 Method

3.1 Problem Statement

Given a set of conversation participants S, utterances U, and emotion labels Y, a conversation consisting of k utterances is represented as $[(s_i, u_1, y_1), (s_j, u_2, y_2), ..., (s_i, u_k, y_k)]$, where $s_i, s_j \in S$ are the conversation participants. If $i = j$, then s_i and s_j refer to the same speaker. $y_k \in Y$ is the emotion of the k-th utterance in a conversation, which belongs to one of the predefined emotion categories. Additionally, $u_k \in U$ is the k-th utterance. u_k is provided in the format of a video clip, speech segment, and text transcript. i.e. $u_k = \{t_k, a_k, v_k\}$, where $\{t, a, v\}$ denotes a text transcript, speech segment, and a video clip. The objective of ERC is to predict y_k, the emotion corresponding to the k-th utterance in a conversation.

3.2 TelME

3.2.1 Model Overview

We propose Teacher-leading Multimodal fusion network for ERC (TelME), as illustrated in Figure 3. This framework is devised based on the hypothesis that by exploiting the varying levels of modality-specific contributions to emotion recognition, there is a potential to enhance the overall performance of an ERC system. Therefore, we introduce a strategic approach that focuses on accentuating the powerful modality while bolstering the weak modalities. We first extract powerful emotional representations through context modeling from text modality while capturing auditory and visual features of the current speaker from non-verbal modalities. However, due to the limited emotional recognition capability of audio and visual features as well as the heterogeneity between the modalities, effective multimodal interactions cannot be guaranteed (Zheng et al., 2022). We thus mitigate the heterogeneity between modalities while maximizing the effectiveness of non-verbal modalities by distilling emotion-relevant knowledge of the teacher model into non-verbal students. We also use a fusion method in which strong emotional features from the teacher encoder are shifted by referring to representations of students strengthened in reverse. In the subsequent sections, we discuss the three components of TelME: Feature Extraction, Knowledge Distillation, and Attention-based modality Shifting Fusion.

3.2.2 Feature Extraction

Figure 3 visually illustrates how each modality encoder receives its corresponding input to extract emotional features. In this section, we explain the methodologies employed to generate emotional features that correspond to the input signals of each modality.

Text: Following previous research (Lee and Lee, 2021; Song et al., 2022a), we conduct context modeling, considering all utterances from the inception of the conversation up to the k-th turn as the context. To handle speaker dependencies and differentiate between speakers, we represent speakers using the special token, $< s_i >$. Additionally, we construct the prompt, "Now $< s_i >$ feels $< mask >$" to emphasize the emotion of the most recent speaker. We report the effect of the prompt in Appendix A.2. The emotional features are derived from the embedding of the special token, $< mask >$. For our text encoder, we employ the modified Roberta (Liu et al., 2019). Roberta is a language model that has undergone extensive pretraining on a large-scale text corpus and has exhibited its efficacy across various natural language processing tasks. We can extract emotional features from the text encoder as follows.

$$C_k = [< s_i >, t_1, < s_j >, t_2, ..., < s_i >, t_k] \quad (1)$$

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

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

where $< s_i >$ is the special token indicating the speaker and $< /s >$ is the separation token of Roberta. $F_{T_k} \in R^{1\times d}$ is the embeddings of the mask token, $< mask >$ and d is the dimension of the encoder.

Audio: Self-supervised learning using Transformer has witnessed remarkable achievement, not only within the field of natural language processing but also in the realms of audio and video (Bertasius et al., 2021; Baevski et al., 2022). In line with this trend, we set the initial state of our audio encoder with data2vec (Baevski et al., 2022). To focus solely on the voice of the current speaker, we only utilize a speech segment of the k-th utterance, denoted as a_k. This speech segment is processed according to the pre-trained processor. The audio encoder then extracts emotional features from the processed input as follows.

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

where $F_{a_k} \in R^{1\times d}$ is the embeddings of a_k and d is the dimension of the encoder.

Visual: Following the same reasoning as the audio modality, we configure the initial state of our visual encoder using Timesformer (Bertasius et al., 2021). In order to concentrate exclusively on the facial expressions of the current speaker, we solely utilize a video clip of the k-th utterance, denoted as v_k. We extract the frames corresponding to the k-th utterance from the video and construct v_k through image processing. The visual encoder then extracts emotional features from the processed input as follows.

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

where $F_{v_k} \in R^{1\times d}$ is the embedding of v_k and d is the dimension of the encoder.

3.2.3 Knowledge Distillation

Addressing the challenge of heterogeneity between modalities and low emotional recognition contributions of non-verbal modalities holds great potential in facilitating satisfactory multimodal interactions (Zheng et al., 2022). Thus, we distill strong emotion-related knowledge of a language model that understands linguistic contexts, thereby augmenting the emotional features extracted from the other two modalities that have comparatively lower contributions. We employ two distinct types of knowledge distillation concurrently: response and feature-based distillation. The overall loss for the student can be composed of the classification loss, response-based distillation loss, and feature-based distillation loss, i.e.,

$$L_{student} = L_{cls} + \alpha L_{response} + \beta L_{feature} \quad (6)$$

where α and β are the factors for balancing the losses.

$L_{response}$ utilizes DIST (Huang et al., 2022), a technique originally used in image networks, as a cross-modal distillation for ERC. As shown in Figure 2, due to the significant gap between the text modality and the other two modalities, effective knowledge distillation can be challenging. Therefore, unlike conventional KD methods, we use a KD approach($L_{response}$) that utilizes Pearson correlation coefficients instead of KL divergence as follows.

$$d(\mu, \upsilon) = 1 - \rho(\mu, \upsilon) \tag{7}$$

where $\rho(\mu, \upsilon)$ is the Pearson correlation coefficient between two probability vectors μ and υ.

Specifically, $L_{response}$ aims to distill preferences (relative rankings of predictions) by teachers through the correlations between teacher and student predictions, which can usefully perform knowledge distillation even in the extreme differences between teacher and student. We gather the predicted probability distributions for all instances within a batch and calculate the Pearson correlation coefficient between the teacher and student for inter-class and intra-class relations (Figure 3). Subsequently, we transfer the inter-class and intra-class relation to the student. The specific formulation of the response-based distillation can be described as follows.

$$Y^t_{i,:} = softmax(Z^t_{i,:}/\tau) \tag{8}$$

$$Y^s_{i,:} = softmax(Z^s_{i,:}/\tau) \tag{9}$$

$$L_{inter} = \frac{\tau^2}{B} \sum_{i=1}^{B} d(Y^s_{i,:}, Y^t_{i,:}) \tag{10}$$

$$L_{intra} = \frac{\tau^2}{C} \sum_{j=1}^{C} d(Y^s_{:,j}, Y^t_{:,j}) \tag{11}$$

$$L_{response} = L_{inter} + L_{intra} \tag{12}$$

Given a training batch B and the emotion categories C, $Z^s \in R^{B \times C}$ is the prediction matrix of the student and $Z^t \in R^{B \times C}$ is the prediction matrix of the teacher. $\tau > 0$ is a temperature parameter to control the softness of logits.

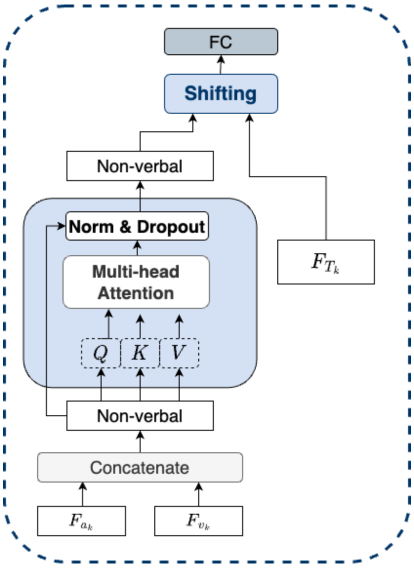

Figure 4: Attention-based modality Shifting Fusion

However, rather than relying solely on $L_{response}$, we introduce $L_{feature}$ as an additional distillation loss to better leverage the embedded information in the teacher network. $L_{feature}$ aims to mitigate the heterogeneity between the representations of the teacher and student models, allowing us to distill richer knowledge from the teacher compared to using only $L_{response}$. Through this, the features of the students can faithfully support the teacher during the multimodal fusion stage. $L_{feature}$ leverages the similarity among normalized representation vectors of the teacher and the student within a batch (Figure 3). We construct the target similarity matrix by performing a dot product between the representation matrix of the teacher and its transposition matrix. By applying the softmax function to this matrix, we derive the target probability distribution as follows.

$$P_i = \frac{exp(M_{i,j}/\tau)}{\sum_{l=1}^{B} exp(M_{i,l}/\tau)}, \forall i, j \in B \tag{13}$$

where B is a training batch and $M \in R^{B \times B}$ is the target similarity matrix. $\tau > 0$ is a temperature parameter controlling the smoothness of the distribution. P_i is the target probability distribution.

Similarly, we can compute the similarity matrix between the teacher and the student by taking the dot product of their representations. Subsequently, we can calculate the similarity probability distribution as follows.

$$Q_i = \frac{exp(M'_{i,j}/\tau)}{\sum_{l=1}^{B} exp(M'_{i,l}/\tau)}, \forall i, j \in B \tag{14}$$

where $M' \in R^{B \times B}$ is the similarity matrix of student and teacher. Q_i is the similarity probability distribution of teacher and student.

With these two probability distributions, we compute the KL-divergence as the loss for the feature-based distillation.

$$L_{feature} = \frac{1}{B}\sum_{i=1}^{B} KL(P_i \parallel Q_i) \qquad (15)$$

where KL is the Kullback–Leibler divergence.

3.2.4 Attention-based modality Shifting Fusion

The emotional features from the enhanced student networks have the potential to impact the emotion-relevant representations of the teacher model, providing information that may not be captured from the text. To fully utilize these features, we adopt a multimodal fusion approach where feature vectors from the student models manipulate the representation vectors from the teacher, effectively incorporating non-verbal information into the representation vector. To highlight non-verbal characteristics, we concatenate the vectors of the student models and perform multi-head self-attention. The vectors of non-verbal information generated through the multi-head self-attention process and emotional features of the teacher encoder enter the input of the shifting step (Figure 4). We are inspired by Rahman et al. (2020) to construct the shifting step. In the shifting step, a gating vector is generated by concatenating and transforming the vector of the teacher model and the vector of the non-verbal information.

$$g^k_{AV} = R(W_1 \cdot < F_{T_k}, F^k_{attention} > + b_1) \qquad (16)$$

where <,> is the operation of vector concatenation, $R(x)$ is a non-linear activation function, W_1 is the weight matrix for linear transform, and b_1 is scalar bias. $F_{attention}$ is the emotional representation vectors of non-verbal information. g_{AV} is the gating vector. The gating vector highlights the relevant information in the non-verbal vector according to the representations of the teacher model. We define the displacement vector by applying the gating vector as follows.

$$H_k = g^k_{AV} \cdot (W_2 \cdot F^k_{attention} + b_2) \qquad (17)$$

where W_2 is the weight matrix for linear transform and b_2 is scalar bias. H is the non-verbal information-based displacement vector.

We subsequently utilize the weighted sum between the representation vector of the teacher and the displacement vector to generate a multimodal vector. Finally, we predict emotions using the multimodal vector.

$$Z_k = F_{T_k} + \lambda \cdot H_k \qquad (18)$$

$$\lambda = min(\frac{\|F_k\|_2}{\|H_k\|_2} \cdot \theta, 1) \qquad (19)$$

where Z is the multimodal vector. We apply the scaling factor λ to control the magnitude of the displacement vector and θ as a threshold hyperparameter. $\|F_k\|_2, \|H_k\|_2$ denote the L2 norm of the F_k and H_k vectors respectively.

Dataset	IEMOCAP			MELD		
	train	dev	test	train	dev	test
Dialogue	108	12	31	1038	114	280
Utterance	5163	647	1623	9989	1109	2610
Classes		6			7	

Table 1: Statistics of the two benchmark datasets.

4 Experiments

4.1 Datasets

We evaluate our proposed network on MELD (Poria et al., 2018) and IEMOCAP (Busso et al., 2008) that include text, audio and visual modalities. The statistics are shown in Table 1.

MELD is a multi-party dataset consisting of over 1400 dialogues and over 13,000 utterances extracted from the TV series Friends. This dataset contains seven emotion categories for each utterance: neutral, surprise, fear, sadness, joy, disgust, and anger.

IEMOCAP consists of a total of 7433 utterances and 151 dialogues in 5 sessions, each involving two speakers per session. Each utterance is labeled as one of six emotional categories: happy, sad, angry, excited, frustrated and neutral. The train and development datasets consist of the first four sessions randomly divided at a 9:1 ratio. The test dataset consists of the last session.

4.2 Experiment Settings

We evaluate all experiments using the weighted average F1 score on two class-imbalanced datasets. We use the initial weight of the pre-trained models from Huggingface's Transformers (Wolf et al., 2019). The output dimension of all encoders is unified to 768. The optimizer is AdamW and the initial learning rate is 1e-5. We use a *linear schedule with*

Models	MELD: Emotion Categories								IEMOCAP
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	F1	F1
DialogueRNN (Majumder et al., 2019)	73.50	49.40	1.20	23.80	50.70	1.70	41.50	57.03	62.75
ConGCN (Zhang et al., 2019)	76.70	50.30	8.70	28.50	53.10	10.60	46.80	59.40	64.18
MMGCN (Hu et al., 2021b)	-	-	-	-	-	-	-	58.65	66.22
DialogueTRM (Mao et al., 2021)	-	-	-	-	-	-	-	63.50	69.23
DAG-ERC (Shen et al., 2021)	-	-	-	-	-	-	-	63.65	68.03
MM-DFN (Hu et al., 2022a)	77.76	50.69	-	22.94	54.78	-	47.82	59.46	68.18
M2FNet (Chudasama et al., 2022)	-	-	-	-	-	-	-	66.71	69.86
EmoCaps (Li et al., 2022)	77.12	**63.19**	3.03	42.52	57.50	7.69	**57.54**	64.00	**71.77**
UniMSE (Hu et al., 2022b)	-	-	-	-	-	-	-	65.51	70.66
GA2MIF (Li et al., 2023a)	76.92	49.08	-	27.18	51.87	-	48.52	58.94	70.00
FacialMMT (Zheng et al., 2023)	80.13	59.63	19.18	41.99	64.88	18.18	56.00	66.58	-
TelME	**80.22**	60.33	**26.97**	**43.45**	**65.67**	**26.42**	56.70	**67.37**	70.48

Table 2: Performance comparisons on MELD (7-way) and IEMOCAP

Dataset	ASF	L_response	L_feature	F1
IEMOCAP	✗	✗	✗	63.33
	✓	✗	✗	68.19
	✓	✓	✗	69.42
	✓	✓	✓	**70.48**
MELD	✗	✗	✗	67.04
	✓	✗	✗	66.75
	✓	✓	✗	67.23
	✓	✓	✓	**67.37**

Table 3: Results of ablation study. Here, $L_{response}$ is our response-based distillation, $L_{feature}$ is our feature-based distillation and ASF is our fusion method.

warmup for the learning rate scheduler. All experiments are conducted on a single NVIDIA GeForce RTX 3090. More details are in Appendix A.3.

4.3 Main Results

We compare TelME with various multimodal-based ERC methods (explained in Appendix A.4) on both datasets in Table 2. TelME demonstrates robust results in both datasets and achieves state-of-the-art performance on MELD. Specifically, TelME exhibits a substantial 3.37% difference compared to EmoCaps with the current state-of-the-art performance in IEMOCAP and improves 0.66% over the previous state-of-the-art method (M2FNet) in MELD. Previous methods such as EmoCaps and UniMSE have shown effectiveness in IEMOCAP but exhibit somewhat weaker performance on MELD. This makes our findings particularly significant.

As shown in Table 2, we report the performance of various methods for emotion labels in MELD. TelME outperforms other models in all emotions except Surprise and Anger. However, assuming that Surprise and Fear, as well as Disgust and Anger, are similar emotions, Emocaps shows a bias towards Surprise and Anger during inference, only achieving 3.03% and 7.69% in F1 score for Fear and Disgust, respectively. On the other hand, TelME distinguishes these similar emotions better, bringing the scores for Fear and Disgust up to 26.97% and 26.42%. We speculate that our framework predicts minority emotions more accurately as the non-verbal modality information (e.g., intensity and pitch of an utterance) enhanced through our KD strategy better assists the teacher in judging the confusing emotions.

4.4 Ablation Study

We conduct an ablation study to validate our knowledge distillation and fusion strategies in Table 3. The initial row for each dataset represents the outcome of training each modality encoder using cross-entropy loss and concatenating the embeddings without incorporating distillation loss and our fusion method.

Using our fusion method alone, IEMOCAP showed performance improvement, but MELD showed poor performance. The effectiveness of our fusion method in achieving optimal modality interaction cannot be guaranteed without knowledge distillation. Because each encoder is trained independently, focusing solely on improving its performance without considering the multimodal interaction. On the other hand, As our knowledge distillation components are added, these bring about consistent improvements for both datasets.

When we examine the specific effects of the KD strategy, we observe performance improvements for both datasets, even when using only $L_{response}$. From these results, we confirm that $L_{response}$ is a knowledge distillation approach capable of ad-

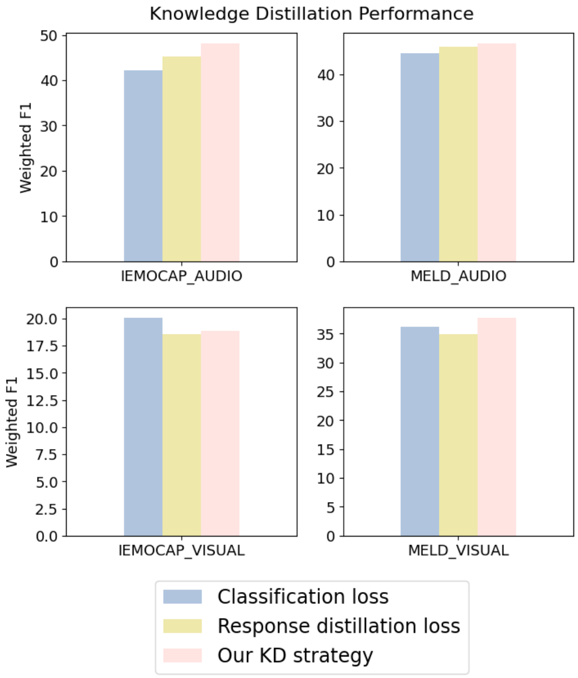

Figure 5: Individual performance of audio and visual modalities according to knowledge distillation type.

dressing the between modalities. Furthermore, adding $L_{feature}$ aimed to leverage the richer knowledge of the teacher is more effective in IEMOCAP and shows marginal performance enhancements in MELD. However, we speculate that the slight improvement in MELD may be attributed to the fundamental issue of class imbalance, limiting the effectiveness of the overall architecture. We show an analysis of this problem in Appendix A.5 as well as an error analysis of the emotion classes in Appendix A.6.

Figure 5 shows the individual performance of the audio and visual modalities based on the distillation loss. We observe that applying both types of distillation loss is more effective compared to not applying them. The performance of visual modality on the IEMOCAP dataset has declined, possibly because facial expressions are not effectively captured in the limited image frames of a short utterance. However, even with lower individual performance, all modalities have been shown to contribute to the improvement of emotion recognition performance through our approach (Table 3, 4).

In summary, we demonstrate that by applying both types of knowledge distillation, we can maximize the effectiveness of non-verbal modalities and effectively interact with our fusion method.

Methods	Remarks	IEMOCAP	MELD
Audio	KD	48.11	46.60
Visual	KD	18.85	36.72
Text	-	66.60	66.57
Text + Visual	ASF	67.94	67.05
Text + Audio	ASF	69.26	67.19
TelME		**70.48**	**67.37**

Table 4: Performance comparison for single modality and multiple multimodal combinations

4.5 The Impact of Each Modality

Table 4 presents the results for single-modality and multimodal combinations. The single-modality performances for audio and visual are the results after applying our knowledge distillation method, and the same fusion approach as TelME is used for dual-modality results. The text modality performs the best among the single-modality, which supports our decision to use the text encoder as the teacher model. In addition, the combination of non-verbal modalities and text modality achieves superior performance compared to using only text. Our findings indicate that the audio modality significantly contributes more to emotion recognition and holds greater importance compared to the visual modality. We speculate this can be attributed to its ability to capture the intensity of emotion through variations in the tone and pitch of the speaker. Overall, our method achieves 3.52% improvement in IEMOCAP and 0.8% in MELD over using only text.

5 Conclusion

This paper proposes Teacher-leading Multimodal fusion network for ERC (TelME), a novel multimodal ERC framework. TelME incorporates a cross-modal distillation that transfers the knowledge of text encoders trained in linguistic contexts to enhance the effectiveness of non-verbal modalities. Moreover, we employ the fusion method that shifts the features of the teacher model by referring to non-verbal information. We show through experiments on two benchmarks that our approach is practical in ERC. TelME delivers robust performance in both datasets and especially achieves state-of-the-art in the MELD dataset consisting of multi-party conversational scenarios. We believe that this research presents a new direction that can incorporate multimodal information for ERC.

Limitations

This study has a limitation wherein the visual modality shows a lower capability to recognize emotions compared to the audio modality. To address this limitation, future research should focus on developing techniques to accurately capture and interpret the facial expressions of the speaker during brief utterances. By improving the extraction of visual features, the effectiveness of knowledge distillation can be significantly enhanced, thus showcasing its potential to make a more substantial contribution to emotion recognition.

References

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

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

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. 2019. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. *arXiv preprint arXiv:2306.01505*.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022a. 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)*, pages 7037–7041. IEEE.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021a. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022b. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021b. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779*.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2022. Knowledge distillation from a stronger teacher. *arXiv preprint arXiv:2205.10536*.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. *arXiv preprint arXiv:1904.04446*.

Woojeong Jin, Maziar Sanjabi, Shaoliang Nie, Liang Tan, Xiang Ren, and Hamed Firooz. 2021. Msd: Saliency-aware knowledge distillation for multimodal understanding. *arXiv preprint arXiv:2101.01881*.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Joosung Lee and Wooin Lee. 2021. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. *arXiv preprint arXiv:2108.11626*.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023a. Ga2mif: Graph and attention based two-stage multi-source information fusion for conversational emotion detection. *IEEE Transactions on Affective Computing*.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023b. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. Emocaps: Emotion capsule based model for conversational emotion recognition. *arXiv preprint arXiv:2203.13504*.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2023. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. Dialoguetrm: Exploring multimodal emotional dynamics in a conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2694–2704.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022a. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*.

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022b. Emotionflow: Capture the dialogue level emotion transitions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8542–8546. IEEE.

Vinh Tran, Niranjan Balasubramanian, and Minh Hoai. 2022. From within to between: Knowledge distillation for cross modality retrieval. In *Proceedings of the Asian Conference on Computer Vision*, pages 3223–3240.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. 2022. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421.

Jiahao Zheng, Sen Zhang, Zilu Wang, Xiaoping Wang, and Zhigang Zeng. 2022. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE Transactions on Multimedia*.

Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arXiv preprint arXiv:2106.01071*.

A Appendix

A.1 Study on Teacher Modality

	MELD	IEMOCAP
TelME (Audio Teacher)	56.28	49.36
TelME (Visual Teacher)	56.85	56.78
TelME (Text Teacher)	**67.37**	**70.48**

Table 5: TelME Performance by Teacher Modality

We conduct comparative experiments by setting each modality as the teacher modality. Table 5 shows the performance of our framework based on the teacher modality setting. Our study shows that the TelME framework performs best with the text encoder as the teacher, while treating the other modalities as the teacher significantly hinders model performance.

	Audio Student	Visual Student	Text Student
Audio Teacher	44.55	34.86	54.83
Visual Teacher	40.18	36.14	59.72
Text Teacher	**46.60**	**36.72**	**66.60**

Table 6: Teacher Modality Study on MELD

	Audio Student	Visual Student	Text Student
Audio Teacher	42.24	20.45	57.42
Visual Teacher	44.13	**22.06**	63.94
Text Teacher	**48.11**	18.85	**66.57**

Table 7: Teacher Modality Study on IEMOCAP

Additionally, Tables 6 and 7 report the individual performance of the student models based on the teacher modality. The diagonals (cases where the teacher and student modalities are the same) in Tables 6 and 7 represent results without performing Knowledge Distillation (KD). Through our comparative experiment results, we observe that a robust text encoder can most effectively serve as the teacher. Specifically, designating the text encoder as the teacher enhances the performance of all student models except for the visual student in IEMOCAP. On the other hand, it is evident that treating a weak non-verbal model as the teacher impairs student performance. We believe this provides significant evidence for why the text encoder should be assigned the role of the teacher.

A.2 Effect of the prompt

	MELD	IEMOCAP
w/o prompt ([cls]+context)	65.25	66.48
context + prompt	**66.57**	**66.60**

Table 8: Comparison of the teacher performance based on the use of the prompt

Table 8 shows an ablation experiment on the prompt. We remove the prompt and use the CLS token to compare emotion prediction results with the results using the prompt. We observe from the results that the prompt helps to infer the emotion of a recent speaker from a set of textual utterances.

A.3 Hyperparameter Settings

Hyperparameter	IEMOCAP	MELD
Knowledge distillation		
Balance factors for $L_{student}$	α=0.1	1
Temperature for $L_{response}$	4	2
Temperature for $L_{feature}$	1	1
Attention modality Shifting Fusion		
Threshold parameter	0.01	0.1
Dropout	0.2	0.1
The number of heads for multi-head attention	4	3

Table 9: hyperparameter settings of TelME on two datasets

Through our KD strategy, audio and visual encoders are trained using the loss functions mentioned in Equation 6. In $L_{student}$, the balancing factors are all set to 1 excluding α for IEMOCAP. The temperature parameter for the $L_{response}$ function is adjusted to 4 for MELD and 2 for IEMOCAP. The temperature parameter for $L_{feature}$ is set to 1 regardless of the dataset. We also use a fusion method that shifts vectors in the teacher model, where the threshold parameter is set to 0.01 for IEMOCAP and 0.1 for MELD. Furthermore, Dropout is adjusted to 0.2 for MELD and 0.1 for IEMOCAP. The number of heads used in the multi-head attention process is 4 for IEMOCAP and 3 for MELD.

A.4 Compared Models

We compare TelME against the following models: DialogueRNN (Majumder et al., 2019) employs Recurrent Neural Networks (RNNs) to capture the speaker identity as well as the historical context and the emotions of past utterances to capture the nuances of conversation dynamics. ConGCN (Zhang et al., 2019) utilizes a Graph Convolutional Network (GCN) to represent relationships within a graph that incorporates both context and speaker information of multiple conversations. MMGCN (Hu et al., 2021b) also proposes a GCN-based approach, but captures representations of a conversation through a graph that contains long-distance flow of information as well as speaker information. DialogueTRM (Mao et al., 2021) focuses on modeling both local and global context of conversations to capture the temporal and spatial dependencies. DAG-ERC (Shen et al., 2021) studies how conversation background affects information of the surrounding context of a conversation. MMDFN (Hu et al., 2022a) proposes a framework that aims to enhance integration of multimodal features through dynamic fusion. EmoCaps (Li et al., 2022) introduces an emotion capsule that fuses information from multiple modalities with emotional tendencies to provide a more nuanced understanding of emotions within a conversation. UniMSE (Hu et al., 2022b) seeks to unify ERC with multimodal sentiment analysis through a T5-based framework. GA2MIF (Li et al., 2023a) introduces a two-stage multimodal fusion of information from a graph and an attention network. FacialMMT (Zheng et al., 2023) focuses on extracting the real speaker's face sequence from multi-party conversation videos and then leverages auxiliary frame-level facial expression recognition tasks to generate emotional visual representations.

A.5 Class Imbalance

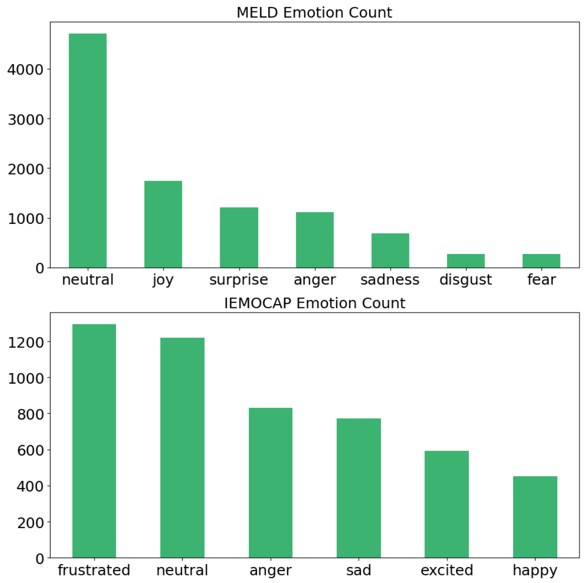

Figure 6: Count distribution of emotion classes for both MELD and IEMOCAP datasets

Figure 6 illustrates the label distribution within the MELD and IEMOCAP datasets. Notably, the MELD dataset exhibits a pronounced imbalance, with the "neutral" class comprising the majority at 47% of the data, followed by "joy" with 17% and "surprise" with 12%. This substantial class imbalance presents a challenge in the context of distillation, specifically for the teacher encoder to initially identify the minority classes and subsequently transfer this information to the non-verbal student encoders. We believe that this class imbalance is a contributing factor to the limited observed improvements associated with $L_{feature}$ in the MELD dataset as compared to the IEMOCAP dataset.

A.6 Error Analysis

Figure 7 shows the normalized confusion matrices of the TelME and the understated model for two datasets. We can evaluate the quality of the emotion prediction through the confusion matrix. TelME shows better True Positive results in almost all emotion classes. This suggests that TelME is extracting and fusing finer-grained features to infer emotions without bias. TelME better classifies similar emotions compared to the understated model(e.g., ex-

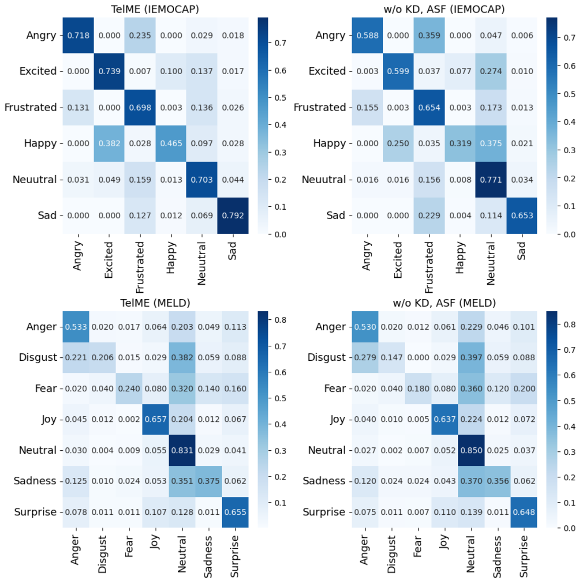

Figure 7: Confusion Matrices on IEMOCAP and MELD

cited and happy, angry and frustrated). However, the result of misclassifying happy as exciting is a little high. This result is due to the lowest percentage of happy in IEMOCAP with unbalanced classes. Even in the case of MELD, the emotion in which most emotion classes are misclassified is neutral, with the highest count. We can observe a similar misclassification tendency in other research (Chudasama et al., 2022; Hu et al., 2023) as well. Hence, we suspect that the cause of misclassification is not a problem with the method we proposed but rather stems from a class imbalance issue.