Emotion Recognition in Conversation Based on the Fine-grained Multidimensional Emotion Representation Learning

Anonymous ACL submission

Abstract

Traditional emotion recognition in conversation (ERC) studies usually are designed to predict a fixed set of predetermined emotion categories. This limited supervision diminishes the expressive power of the data, resulting in failing to capture the complexity of human emotions in conversation. Learning from a well-designed fine-grained representation of emotions offers a promising alternative that utilizes a wider range of supervision. In this paper, the proposed Fine-grained Multidimensional Emotion Representation Learning (FMERL) framework integrates multitask learning and contrastive learning, and extends the emotion representation of valence, arousal and dominance (VAD) from psychological field to both continuous and discrete forms. Firstly, the emotion features from text, audio and visual modalities are extracted. Then, the multimodal features are fused by a transformer-based model. The multitask learning contains three feedforward networks (FFNs), the valence net, arousal net, and dominance net, for learning the continuous fine-grained emotion representations from the fused multimodal features. The contrastive learning aligns fused multimodal features with the discrete fine-grained emotion representations derived through prompt engineering applied to a large language model. The transferable ability of contrastive learning enables FMERL to map the semantic information of emotion representation and fused multimodal features into a shared embedding space, thereby understanding their semantic relationships and enabling zero-shot learning. Experimental results on the IEMOCAP and MELD datasets have shown that FMERL achieves state-of-the-art performance in emotion recognition and implements zero-shot learning in the field of ERC.

1 Introduction

In recent years, the field of emotion recognition in conversation (ERC) has garnered significant attention from researchers, driven by the increasing demand for more sophisticated human-computer interaction systems(Zhou et al., 2020). The purpose of ERC is to analyze and interpret the emotion content embedded in conversational exchanges, leveraging the advancements of multimodal data analysis. Traditional ERC studies have focused mainly on predicting a limited set of emotion categories, which limits their capability to capture more complex human emotions and leads to their poor generality and usability.

The reason lies in the poor representation capability of hard labels as supervision signals. Emotions in conversation are inherently nuanced and vary significantly in intensity and context. By constraining the model to the predefined limited fixed set of emotion categories, such as happiness, sadness, anger, and fear, some important emotional subtleties in a conversation may inevitably be ignored, which undoubtedly results in failing to capture the complexity of human emotions. The lack of emotional granularity hinders the model's ability to provide meaningful insights or responses, ultimately reducing its effectiveness in emotion recognition.

To improve the generalization ability of models, recent studies on Emotion Recognition in Conversations (ERC) have increasingly focused on representation learning by contrastive learning. These studies can be mainly categorized into two kinds of approaches: (1) data-side representation learning (needs to be defined), CKCL (Tu et al., 2023) leverages contrastive learning between context and knowledge to refine emotion vector representations, while CLED (Kang and Cho, 2024) enhances emotion recognition by performing emotion interpolation data augmentation in the hidden space of pre-trained language models, combined with a reinforced contrastive objective for neutral emotions. (2) label-side representation learning (needs to be defined), SACL (Hu et al., 2023) adopts a supervised adversarial contrastive learning framework

to learn emotion representations. However, most of these works predict emotions in a fixed set of emotion categories. Our work belongs to label-side representation learning, but is different from above mentioned methods, inspired by psychological studies, fine-grained multidimensional emotion representation of valence, arousal and dominance(VAD) is learned, where valence refers to the positivity or negativity of an emotion, arousal indicates the level of physiological activation or intensity, and dominance reflects the sense of control or power in a situation.

In this paper, we propose a fine-grained emotion representation learning (FMERL) framework, The emotion features from text, audio and visual modalities are extracted at first. Then, the multimodal features are fused by a transformer-based model. FMERL integrates the multitask learning and the contrastive learning. The multitask learning contains three feedforward networks (FFNs), the valence net, arousal net, and dominance net, for learning the continuous fine-grained emotion representations from the fused multimodal features. The contrastive learning aligns fused multimodal features with the discrete fine-grained emotion representations derived through prompt engineering applied to a large language model. Experimental results have shown that FMERL achieves state-of-the-art performance in emotion recognition.

Our contributions are summarized below. (1) We propose a fine-grained multidimensional emotion representation by extending the emotion representation of valence, arousal and dominance (VAD) from psychological field to both continuous and discrete forms. (2) We propose a Fine-grained Multidimensional Emotion Representation Learning framework (FMERL), and it integrates the multitask learning and the contrastive learning. The multitask learning learns continuous fine-grained emotion representations, while the contrastive learning learns discrete fine-grained emotion representations. (3) Our method achieves state-of-the-art performance in ERC and implements zero-shot learning in ERC, which has been evidenced by experiments on the IEMOCAP and MELD datasets.

2 Related Work

2.1 Multidimensional Emotion Representation Model

In sentiment analysis research, several well-known sentiment classification models are available. The Ekman model (Ekman et al., 1969) classifies emotions into six basic emotions: happiness, sadness, anger, fear, surprise, and disgust. (Plutchik, 2003) proposes an emotion wheel to classify emotions.

However, the discrete models above categorize sentiment into a limited number of fixed categories, which may overlook similar sentiments and subtle variations. The dimensional emotion model offers a nuanced description and measurement of emotions, treating them as points in a multidimensional space and mapping them to a continuous frequency spectrum. One of the famous dimensional emotion models is the PAD model (Mehrabian, 1974). Within this model, pleasure denotes the positive or negative valence of emotion experiences, arousal reflects the intensity or level of emotion activation, and dominance indicates the extent to which emotions influence individual behavior. (Russell, 1980) introduces the widely recognized arousal-valence model, where valence indicates positive or negative evaluations of emotion intensity, while arousal measures an individual's energy level, with low arousal signifying less energy or lower emotion intensity. In this paper, due to the semantic similarity between pleasure and valence, we abandon pleasure and adopt valence, arousal, and dominance to build the multidimensional emotion model.

2.2 Contrastive Learning

SimCLR (Chen et al., 2020) is a milestone in the field of contrastive learning, which employs diverse image augmentation techniques to generate positive and negative samples from a single image for visual representation. CLIP (Radford et al., 2021) introduces a contrastive learning approach that embeds images and text into the same feature space, enabling cross-modal understanding and zero-shot learning, demonstrating strong performance across various visual tasks. In the field of ERC, (Li et al., 2022) uses the supervised contrastive learning (SCL) to distance utterances with different emotions for better emotion identification. (Tu et al., 2023) uses contrastive learning scenarios among context and knowledge to learn the better representations of emotions. EACL (Yu et al., 2024) employs label encodings as anchors to guide the learning of utterance representations. This paper adopts the CLIP (Radford et al., 2021) approach of embedding emotion representations from labels and multimodal features into a unified feature space for alignment through training.

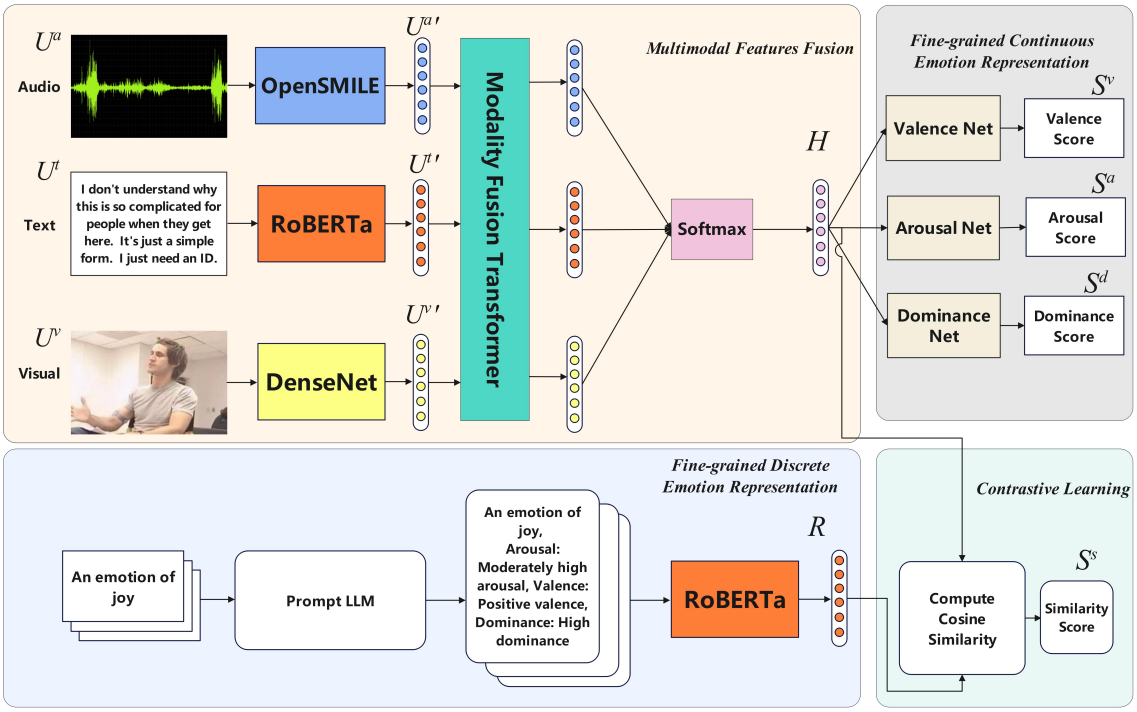

Figure 1: The overall architecture of FMERL.

3 Method

3.1 Task-definition

A conversation consists of N sequential utterances $\{u_1, u_2, \cdots, u_N\}$ from M speakers $\{s_1, s_2, \cdots, s_M\}$. Each utterance u_i is delivered by speaker $s_{\phi(u_i)}$, where ϕ maps each utterance to its corresponding speaker index. Each u_i includes textual (t), acoustic (a), and visual (v) modalities, which are represented as $u_i^t \in \mathbb{R}^{d_t}$, $u_i^a \in \mathbb{R}^{d_a}$, and $u_i^v \in \mathbb{R}^{d_v}$.

The sequences of modalities for all utterances are denoted as $U^t = [u_1^t; u_2^t; \cdots; u_N^t] \in \mathbb{R}^{N \times d_t}$, $U^a = [u_1^a; u_2^a; \cdots; u_N^a] \in \mathbb{R}^{N \times d_a}$, and $U^v = [u_1^v; u_2^v; \cdots; u_N^v] \in \mathbb{R}^{N \times d_v}$.

Our contrastive learning ERC task aims to identify the emotion representation corresponding to each utterance.

3.2 The Model Architecture and Prompt Engineering Framework

Figure 1 gives an overview of our proposed FMERL framework. In the part of multimodal features fusion module, the given input is an utterance U including three modal features: U^a, U^t, and U^v. After extracting utterance-level unimodal features to $U^{a\prime}$, $U^{t\prime}$, and $U^{v\prime}$, we use the Intra- and Inter-modal Transformers from SDT (Ma et al., 2023) as the backbone to fuse multimodal features H. In the fine-grained emotion continuous representation module, emotion scores are predicted across the fine-grained multidimensional scales of valence S^v, arousal S^a, and dominance S^d. In the fine-grained emotion discrete representation module, the traditional emotion category is extened to a fine-grained discrete representation by prompting LLM. Then, the discrete representation is embedded into a high-dimensional vector R. Furthermore, we introduce a contrastive learning module to align the multimodal features H with high-dimensional vector R by maximizing the similarity score S^s.

Figure 2 shows the prompt engineering framework. We extend the fine-grained multidimensional emotion representation into continuous and discrete forms by prompting LLMs. Then, we utilize continuous representation for multitask regression and discrete representation for contrastive learning. Combining these strategies, we finally implement the emotion recognition task.

3.3 Multi-task Regression Learning Based on Continuous Emotion Representaion

To enhance the representational ability of emotions, we propose using a continuous fine-grained features to describe emotions. Rather than the hard labels, we use the labels from three dimensions: valence, arousal, and dominance scores. The scores of these dimensions serve to provide a more fine-

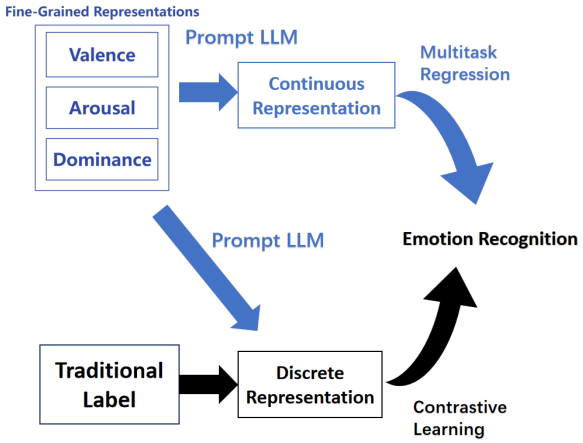

Figure 2: An overview of the prompt engineering framework. The blue part is the continuous representation for multitask regression, and the black part is the discrete representation for contrastive learning.

grained and comprehensive understanding of emotions. In this section, we present a continuous emotion representaion as supervision signal based on prompting with GPT (Brown et al., 2020) to score these three dimensions above. Then, we introduce how the fine-grained emotion representaion facilitate multitask regression learning.

Continuous Emotion Representaion. The multidimensional emotion representation model we used includes valence, arousal, and dominance scores, which is also adopted by IEMOCAP dataset. However, not all datasets include the scores of these three dimensions. The traditional method of manually labeling datasets requires a high cost. Prompting GPT to score these dimensions of text in the dataset can significantly reduce the workload of labeling the dataset. Specifically, given a text of an utterance, we prompt GPT with $Query$, $Metric$, $Utterance$

$$S_i = GPT(Query) \tag{1}$$

where $Query$ is "In the field of emotion recognition, score this utterance in the dimension of $Metric_i$ from 0 to 5. The utterance is : $Utterance$. Hence, the score is:", $Metric$ = {valence, arousal, dorminance}, and $Utterance$ is the current utterance.

Multitask Regression. As shown in Figure 1, the emotion recognition task and regression tasks share the same architecture and weights in the part of multimodal features fusion. For the multitask learning part, we use three feedforward neural networks: valence net, arousal net, and dominance net to predict the valence score, arousal score, and dominance score, respectively. These nets have the same input $H \in \mathbb{R}^{N\times d}$. The output of the regression network n is its predicted emotion scores:

$$\hat{S_n} = \sigma(W_n \cdot H + b_n) \in \mathbb{R}^N \tag{2}$$

where $n \in \{\text{valence}, \text{arousal}, \text{dominance}\}$, $W_n \in \mathbb{R}^{N\times d}$, and $b_n \in \mathbb{R}^N$ are trainable parameters, and σ is the sigmoid function. For these three regression tasks, we use the same optimization policy, which minimizes the mean squared error loss $\mathcal{L}_{\text{mse}}$.

$$\mathcal{L}_{\text{mse}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{S}_i - S_i)^2 \tag{3}$$

where $\hat{S}$ is the predicted emotion score, S is the ground-truth score. During training, in addition to the parameters of three regression networks, the parameters of the modality fusion backbone are also updated.

Although these regression tasks do not contribute directly to the output of the predictions, they enable the model to learn additional representations, thereby enhancing its generalization ability.

3.4 Discrete Fine-grained Emotion Representation

To improve the distinction of various emotions that can be difficult to differentiate by category names alone, we propose utilizing multidimensional emotion information in psychology to improve emotion representations. In this subsection, we describe how to generate discrete emotion representations for humans in conversations based on prompting GPT. We also explain how these representations are integrated into our model to improve recognition.

Prompt with GPT. To harness deeper meanings tied to emotion expression, we employ GPT (Brown et al., 2020) to produce descriptions of different emotions. Nevertheless, directly querying GPT for emotion descriptions may yield verbose descriptions. Hence, we implement a prompt engineering based mechanism to query GPT for emotion descriptions. Following Section 3.3, we selected and integrated valence, arousal, and dominance as the emotional dimensions ED. Then, we query GPT with the emotion dimensions ED to get the descriptions D_i:

$$D_i = GPT(ED_i, Query_i') \tag{4}$$

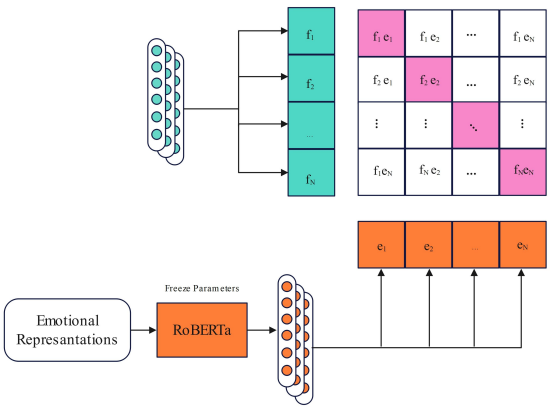

Figure 3: The contrastive learning method to match the feature embeddings (green part) and emotion embeddings (orange part)

where $Query'$ is "Use a brief phrase to describe the level of an emotion of e_i in the dimension of ED." to acquire the level of each emotion dimension ED. For example, if the e_i is joy, the returned D_i is "Arousal: Moderately high arousal, Valence: Positive valence, Dominance: High dominance"

Finally, discrete multi-dimensional emotion model is proposed to describe emotions, encompassing the levels of arousal, valence, and dominance. The descriptions of emotions provide a more fine-grained and comprehensive understanding of emotions concepts.

Enhance Emotion Representation. Our ERC research aims to transform textual descriptions of emotions into feature space representations and align multi-modal feature representations with them. We use the RoBERTa Large model to obtain representations of textual descriptions, matching the multi-modal feature representations to ensure both are in a common embedding space.

3.5 Contrastive Learning Method

The CLIP framework, introduced by (Radford et al., 2021), encodes image and text features into a shared space, maximizing cosine similarity for matching pairs (positives) and minimizing it for non-matching pairs (negatives) through cross-entropy loss. We use this framework to align multi-modal features with emotion representations as shown in Fig 3.

Specifically, we employ two distinct encoders to fuse multimodal features and represent emotions. For a feature-emotion pair $x = \{x^f, x^e\}$, we derive the embeddings using the feature encoder E_f and emotion representation encoder E_e: $f = E_f(x^f)$ and $e = E_e(x^e)$, where $f, e \in \mathbb{R}^D$. The fused features f and emotion representations e for each feature-emotion pair in the mini-batch N are used to create an $N \times N$ matrix of cosine similarities. The diagonal elements of the matrix represent the N positive pairings, while the other elements denote $N^2 - N$ negative pairings. During the training process, the emotion representations encoder E_e is frozen. We propose a triple-loss to maximize the similarity among the diagonal positives and minimize the similarity of the negatives. The first part of our triple-loss is a contrastive learning loss:

$$\mathcal{L}_1 = -\sum_i \log\left(\frac{\exp(S_{ii})}{\sum_j \exp(S_{ij})}\right) \quad (5)$$

$$S_{ij} = \frac{f_i \cdot e_j}{\|f_i\|\|e_j\|} \quad (6)$$

$\mathcal{L}_1$ is designed to maximize intra-class similarity while minimizing inter-class similarity. S_{ii} denotes the similarity score for positive pairs, and S_{ij} spans all potential pairs for a given sample i. This formulation uses a log-softmax approach to convert similarity scores into probabilities, emphasizing the correct class alignment. The similarity metric S_{ij} based on cosine similarity measures the angular separation between vectors, focusing on orientation rather than magnitude. To further differentiate similar emotions, we introduce the cosine similarity loss for negative samples as the second component of our triple loss:

$$\mathcal{L}_2 = \frac{1}{N}\sum_i \sum_j (1 + \cos(\theta_{ij})) \quad (7)$$

$$\cos(\theta_{ij}) = \frac{f_i \cdot n_{ij}}{\|f_i\|\|n_{ij}\|} \quad (8)$$

The supplementary cosine loss $\mathcal{L}_2$ penalizes the alignment of feature vectors f_i with negative samples n_{ij}, thereby promoting the differentiation of similar emotions. To keep the $\mathcal{L}_2$ positive, we add 1 to the cosine similarity. We also take the positive value of $\mathcal{L}_2$ to maximize the differences between features and negative samples. The cosine similarity $\cos(\theta_{ij})$ quantifies the angular difference between the feature vector f_i and the negative samples n_{ij}. We derive our triple-loss function by combining $\mathcal{L}_1$ and $\mathcal{L}_2$ as follows:

$$\mathcal{L}_{\text{tri}} = \mathcal{L}_1 + \lambda_{neg}\mathcal{L}_2 \quad (9)$$

where λ_{neg} is a hyperparameter that regulates the weight of the cosine similarity loss for negative

samples. Equation 9 contributes to a comprehensive loss that enhances the granularity of emotion differentiation within our model.

3.6 Training and Inference

In this subsection, we detail the training and inference processes of our model.

Training. During the training stage, two parts of parameters of the model are updated: multitask regression learning and contrastive learning. The multitask regression learning cost is $\mathcal{L}_{\text{mse}}$, which is formulated in Section 3.3. Then considering the triple loss $\mathcal{L}_{\text{tri}}$ of contrastive learning introduced in Section 3.6, the final loss can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda_{tri}\mathcal{L}_{\text{tri}} \tag{10}$$

where λ_{tri} is the hyper-parameter weight for the additional constraint.

Inference. For each emotion prediction, we compute the the similarity S_{ij} between the feature embedding f_i and emotion embedding e_j by Equation 6. Then, we do the matching task to find the emotion e_j that is most similar to f_i:

$$j^* = \arg\max_j S_{ij} \tag{11}$$

where j^* is the index of the emotion most similar to f_i.

4 Experimental Settings

4.1 Datasets

The experiments were carried out on two datasets: IEMOCAP and MELD.

IEMOCAP.(Busso et al., 2008) IEMOCAP has already used the scores of valence, arousal and dominance as continuous representations to describe emotions. It includes around 12 hours of dyadic conversation videos, segmented into 7,433 utterances and 151 dialogues. Each utterance is categorized into one of six emotions: happy, sad, neutral, angry, excited, or frustrated. There are five sessions in this dataset. We choose the former four sessions to be used for training, while the last one is for testing.

MELD.(Poria et al., 2019) MELD has no scores of valence, arousal and dominance. The prompt engineering of LLM is to extend its labels to these three dimensions. Specifically, we prompt the LLM with each utterance within the MELD and the query of the scores of these three dimensions to get the continuous representation. This multi-party dataset contains 13,708 utterances and 1,433 dialogues from the TV series Friends, with each utterance classified into one of seven emotions: anger, disgust, fear, joy, neutral, sadness, and surprise.

For the discrete form, we prompt the LLM with each emotion category within these two datasets and query the levels of these three dimensions to get the text description of each emotion category. Then, we use a pre-trained text encoder to embed these discrete representations to a vector space.

4.2 Unimodal Feature Extraction

Following (Ma et al., 2023), we use openSMILE (Eyben et al., 2013), DenseNet (Huang et al., 2017), RoBERTa (Liu, 2019) to extract utterance-level features of acoustic, visual, and textual modality respectively.

4.3 Baseline Methods

We compare our proposed model with the following ERC methods.

Supervised learning baseline: **CoMPM** (Lee and Lee, 2021) combines the speaker's pre-trained memory with the context model to improve performance. **SDT** (Ma et al., 2023) is a transformer-based model with a self-distillation mechanism. **MultiEMO** (Shi and Huang, 2023) is an attention-based correlation-aware framework to fuse multimodal features. **MPLP** (Zhang et al., 2023) mimics the thinking process when modeling complex factors of emotion. **SACL-LSTM** (Hu et al., 2023) uses supervised adversarial contrastive learning to learn class-spread structured representations. **CKCL** (Tu et al., 2023) is a contrastive learning framework with context and knowledge that can distinguish the utterances for better vector representations. **Beyond Linguistic Cues** (Xu et al., 2024) incorporates both belief and desire to recognize emotion. **CLED** (Kang and Cho, 2024) is a supervised contrastive learning method with data augmentation method emulating the emotion dynamics.

Unsupervised learning baseline: **Qwen-7B**(Bai et al., 2023), **Llama2-7B**(Touvron et al., 2023), and **Llama3.2-3B**.

4.4 Implementation Details

We use Adam optimizer(Kingma and Ba, 2014) with an initial learning rate of 1.0×10^{-4} for IEMOCAP and 1.0×10^{-5} for MELD. The batch size is 16 for IEMOCAP and 256 for MELD. Following (Ma et al., 2023), the 1D convolutional layers have

Models	IEMOCAP		MELD	
	ACC	w-F1	ACC	w-F1
CoMPM	-	66.33	-	66.52
SDT	73.95	74.08	67.55	66.60
MultiEMO	-	72.84	-	66.74
MPLP	-	66.65	-	66.51
SACL-LSTM	69.08	69.22	67.51	66.45
CKCL	-	67.16	-	66.21
Beyond Linguistic Cues	-	68.22	-	64.27
CLED	-	62.77	-	66.24
FMERL	**74.92**	**75.14**	**67.66**	**66.78**

Table 1: Performance metrics for different models on IEMOCAP and MELD datasets, including total ACC and weighted F1 scores.

IEMOCAP					
Models	Unseen		Full		Params Cont
	ACC	F1	ACC	F1	(Million)
Qwen-7B	13.14	23.22	33.42	30.67	7720
Llama2-7B	11.71	20.97	32.00	29.60	6740
Llama3.2-3B	6.04	11.40	32.98	28.43	3210
FMERL	**36.81**	**27.32**	**58.38**	**57.30**	**83**
MELD					
Models	Unseen		Full		Params Cont
	ACC	F1	ACC	F1	(Million)
Qwen-7B	3.60	6.96	32.38	28.46	7720
Llama2-7B	3.28	6.35	33.98	35.19	6740
Llama3.2-3B	1.56	3.08	42.11	45.55	3210
FMERL	**11.86**	**21.21**	**61.38**	**60.30**	**83**

Table 2: Performance metrics (ACC and F1) for different models across various datasets.

input channels of 1024, 1582, and 342 for textual, acoustic, and visual modalities on IEMOCAP, and 1024, 300, and 342 on MELD. All modalities feature an output channel and kernel size of 1024 and 1, respectively, for both datasets. The transformer encoder has a hidden size, number of attention heads, feed-forward size, and number of layers set to 1024, 8, 1024, and 1. The three multitask learning nets have 2 hidden layers and the dimensions are 1024. The hyper-parameter weights λ_{neg} and λ_{tri} are set to 0.5 and 1. For zero-shot learning, we set the long-tail classes in IEMOCAP (happy) and MELD (fear) as the unseen classes. For the baseline LLMs, we use the few-shot predict for the seen classes and zero-shot for the unseen classes. The results are averages of 5 runs.

5 Results and Analysis

5.1 Main Results

We compare the performance of our proposed model and the state-of-the-art approaches. Table 1 presents the results of supervised learning on IEMOCAP and MELD, whereas the results of zero-shot learning are detailed in Table 2. On the IEMOCAP dataset, FMERL surpasses all baselines, exceeding SDT by 0.97% in overall accuracy and 1.06% in weighted F1-score. On the MELD dataset, FMERL outperforms all baselines, achieving the highest overall accuracy and weighted F1-score, exceeding SDT by 0.11% and 0.18%, respectively.

For the zero-shot learning results, FMERL also achieves state-of-the-art performance. FMERL outperforms all LLMs on IEMOCAP, surpassing Qwen-7B by 29% in overall accuracy and 31.84% in weighted F1-score for all classes, and by 8.62% and 14.25% respectively for unseen classes. Similarly, on MELD, FMERL exceeds Qwen-7B by 24.96% in overall accuracy and 26.63% in weighted F1-score for all classes, and by 23.67% and 4.1% respectively for unseen classes. Additionally, the model parameter count of FMERL is significantly small compared to LLMs. In comparison to the smallest LLM we use as the baseline, Llama3.2-3B, which has 3210 million parameters, our model reduces the parameter count by 97.4%, demonstrating a more compact architecture.

Models	IEMOCAP		MELD	
	ACC	w-F1	ACC	w-F1
FMERL	**74.92**	**75.14**	**67.66**	**66.78**
w/o Triple-loss	74.68	74.91	67.20	65.96
w/o Fine-grained Enhancement	74.18	74.37	66.74	65.11
w/o Dominance	74.86	74.96	67.24	66.16
w/o Arousal	73.44	73.68	67.51	66.37
w/o Valence	74.31	74.38	67.43	66.41
w/o Multitask Regression	72.46	72.75	67.09	65.87

Table 3: Ablation results on IEMOCAP and MELD.

5.2 Ablation Study

We conduct a series of experiments to confirm the effectiveness of components in our method. The results are shown in Table 3. Removing any element of FMERL makes the overall performance worse.

To validate the effects of components in contrastive learning, we remove the triple loss, which encourages the samples to stay away from negative samples. We can find that the lack of triple loss results in a decline in the accuracy of 0.35% and in the weighted F1-score of 0.53% on average of two datasets. Also, without the fine-grained enhancement, it causes a decline in the accuracy of 0.83% and in the weighted F1-score of 1.22% on average.

In multitask regression, ablating dimensions reveals their importance. For IEMOCAP, arousal is most critical; removing it decreases accuracy by 1.48% and weighted F1-score by 1.46%. For MELD, dominance is key; its removal reduces ac-

curacy by 0.46% and weighted F1-score by 0.62%. Removing multitask regression entirely significantly degrades performance, with IEMOCAP experiencing a 2.46% accuracy drop and a 2.39% weighted F1-score decline, while MELD sees a 0.57% accuracy drop and a 0.91% weighted F1-score decline.

5.3 Fine-grained Multidimensional Emotion Representation Learning Analysis

In this section, we conducted a comparison of the multidimensional scores of emotions between the model non-converged and converged stages in the process of training to prove that multidimensional representation learning is critical to the generalization of the model. We randomly select one epoch from each of the two stages, the model's non-converged stage and converged stage, and freeze the model parameters at that point. We assume that the samples within each emotion class follow a multivariate Gaussian distribution. Based on this assumption, we evaluate both models separately on the test set to predict multidimensional scores in Figure 4 and get the confusion matrices of emotion classes in Figure 5. Figure 4 and Figure 5 show that during the non-converged stage, classes are closely packed and hard to distinguish. Once the model converges, the classes spread out, making them easily distinguishable. Our analysis of the above result is as follows: (1) The shared model parameters and representations across multiple tasks enhance generalization by leveraging common features and task correlations. The regression tasks and contrastive learning task share parts of the model, enabling the model to learn more features from different tasks. There are correlations between multitasks, the model can capture these correlations through shared representations, thereby improving the performance. (2) The regularization effect reduces overfitting. Multitask learning, by sharing parameters, acts as a form of regularization. Single emotion recognization task learning may overfit to the training data of this specific task, while multitask regression and contrastive learning forces the model to learn more general features from emotion by optimizing multiple objectives simultaneously, thereby reducing overfitting.

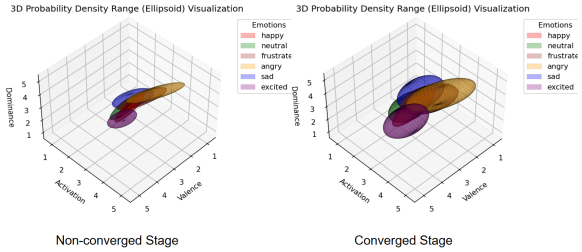

Figure 4: Comparison of multidimensional scores: valence, arousal, and dominance. Left side is the non-converged stage. Right side is the converged stage

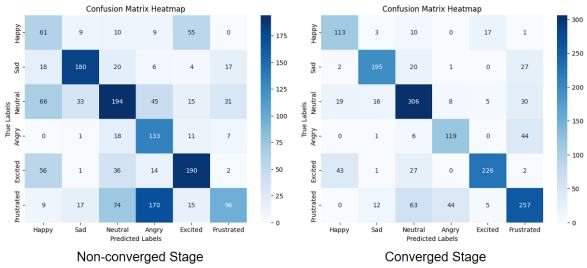

Figure 5: Comparison of confusion matrices. The left side is the non-converged stage. The right side is the converged stage.

6 Conclusion and Future Work

This paper presents a novel framework for emotion recognition in conversation, named Fine-grained Multidimensional Emotion representation learning (FMERL). FMERL utilizes multidimensional emotion representations—valence, arousal, and dominance—as supplementary supervision signals to improve the learning of recognize utterance representations. In addition, we propose a labeling method based on prompt engineering of LLM to provide the supervision signal to the model. Our extensive experiments on two benchmark datasets demonstrate that our approach achieves the state-of-the-art performance in both supervised and unsupervised learning. Ablation studies and evaluations demonstrate that the FMERL framework has excellent data representation capabilities and outstanding emotion recognition abilities.

The zero-shot learning of FMERL is based on the representation learning of the seen classes and inference of the unseen classes, which is following the idea of CLIP. In the future work, we will research the zero-shot learning for emotion recognition without relying on seen classes.

Limitations

Our proposed FMERL shows less performance improvement on the MELD dataset compared to the IEMOCAP dataset. This is because the IEMOCAP dataset has the multidimensional emotion scores meticulously labeled by humans, while these scores of the MELD dataset are labeled by our proposed prompt engineering labeling method. Recent LLMs are still unable to achieve a level of fine-grained emotion recognition comparable to that of humans.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*, page 335–359.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Paul Ekman, E Richard Sorenson, and Wallace V Friesen. 1969. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838.

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*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Yujin Kang and Yoon-Sik Cho. 2024. Improving contrastive learning in emotion recognition in conversation via data augmentation and decoupled neutral emotion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2194–2208.

DiederikP. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv: Learning,arXiv: Learning*.

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*.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11002–11010.

Yinhan Liu. 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*.

Albert Mehrabian. 1974. An approach to environmental psychology. *Massachusetts Institute of Technology*.

Robert Plutchik. 2003. Emotions and life: Perspectives from psychology, biology, and evolution. *American Psychological Association*, 19.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 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*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Tao Shi and Shao-Lun Huang. 2023. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Geng Tu, Bin Liang, Ruibin Mao, Min Yang, and Ruifeng Xu. 2023. Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14054–14067.

Bo Xu, Longjiao Li, Wei Luo, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2024. Beyond linguistic cues: Fine-grained conversational emotion recognition via belief-desire modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2318–2328.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2024. Emotion-anchored contrastive learning framework for emotion recognition in conversation. *arXiv preprint arXiv:2403.20289*.

Ting Zhang, Zhuang Chen, Ming Zhong, and Tieyun Qian. 2023. Mimicking the thinking process for emotion recognition in conversation with prompts and paraphrasing. *arXiv preprint arXiv:2306.06601*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.