PEMV: Improving Spatial Distribution for Emotion Recognition in Conversation Using Proximal Emotion Mean Vectors

Chen Lin, Fei Li, Donghong Ji, Chong Teng*
Key Laboratory of Aerospace Information Security
and Trusted Computing, Ministry of Education,
School of Cyber Science and Engineering, Wuhan University, Wuhan, China
{2019302180106, lifei_csnlp, dhji, tengchong}@whu.edu.cn

Abstract

Emotion Recognition in Conversation (ERC) aims to identify the emotions expressed in each utterance within a dialogue. Existing research primarily focuses on the analysis of contextual structure in dialogue and the interactions between different emotions. Nonetheless, ERC datasets often contain difficult-to-classify samples and suffer from imbalanced label distributions, which pose challenges to the spatial distribution of dialogue features. To tackle this issue, we propose a method that generates **P**roximal **E**motion **M**ean **V**ectors (**PEMV**) based on emotion feature queues to optimize the spatial representation of text features. We design a Center Loss based on PEMVs to pull hard-to-classify samples closer to their respective category centers and employ Angle Loss to maximize the angular separation between different PEMVs. Furthermore, we utilize PEMV as a classifier to better adapt to the spatial structure of dialogue features. Extensive experiments on three widely used benchmark datasets demonstrate that our method achieves state-of-the-art performance and validates its effectiveness in optimizing feature space representations.

1 Introduction

With the rapid development of online social networks, capturing and understanding emotions in conversations has become a widely studied research field in both academia and industry (Li et al., 2020). The task of Emotion Recognition in Conversation (ERC) aims to identify the emotional attributes of each utterance within a dialogue (Zahiri and Choi, 2018; Zhao et al., 2022; Yu et al., 2024). Figure 1 illustrates an example of the ERC task, where each dialogue involves two or more participants, and each utterance carries distinct emotional attributes.

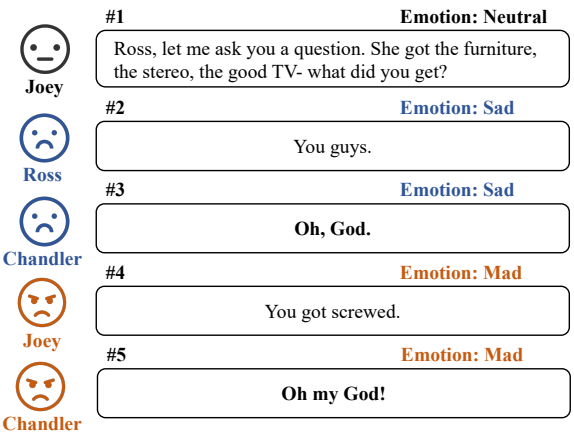


Figure 1: An example from the EmoryNLP dataset illustrates a dialogue involving multiple participants, where similar phrases such as "Oh, God" and "Oh my God" convey distinct emotions.

As shown in Figure 1, even within the same dialogue, similar expressions such as “Oh, God” and “Oh my God” can convey distinct emotions, making certain samples inherently difficult to classify. This challenge is further exacerbated by the need for effective context modeling (Zhang et al., 2023a), and the intricate dynamics of emotion interactions (Yang and Shen, 2021). Together, these factors significantly influence the accurate classification of challenging samples. Additionally, many commonly used datasets suffer from class imbalance issues, further complicating the emotion classification task (Song et al., 2022).

To optimize the spatial representation of utterances, Supervised Contrastive (SupCon) Learning (Khosla et al., 2020) has been widely applied in ERC task. Yu et al. (2024) propose the Emotion-Anchored Contrastive Learning Framework, which uses label encodings as anchors to enhance the distinguishability of utterance representations, particularly in handling similar emotions. However, despite the strong performance of SupCon Learn-

* Corresponding author

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 345–357

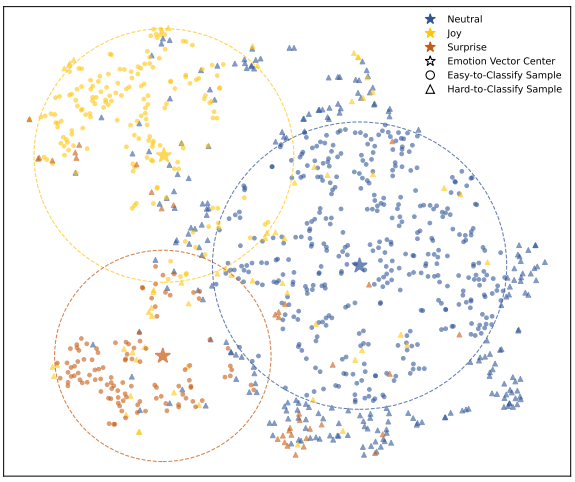


Figure 2: The utterance feature vectors from the MELD dataset, obtained by the model trained with SupCon Learning, are visualized using t-SNE dimensionality reduction. The radius of the circles represents the mean distance from all sample points to the center, plus two standard deviations. The triangular points indicate extreme outliers that are difficult to classify, positioned far from the center.

ing in ERC task, certain limitations still exist. As shown in Figure 2, even after training the model with SupCon Learning, there are still some extremely challenging outliers far from the class centers in the generated spatial representations of dialogue samples, significantly impacting the model's classification performance.

To address this issue, we propose a method for optimizing dialogue text feature space representation based on **P**roximal **E**motion **M**ean **V**ectors (**PEMV**). Inspired by Song et al. (2022), we maintain a dynamic emotion feature queue for each emotion category, recording the most recent text features. By averaging these vectors, we generate the PEMV to represent the typical feature of each category and guide the learning process. Specifically, we design **Center Loss** and use **Angle Loss** to pull within-category vectors closer and maximize the angular separation between different categories. Rather than using the emotion feature queue for curriculum learning as in prior work, we leverage it to directly optimize utterance representation. Additionally, after training the model, we employ PEMV as a classifier to adjust decision boundaries, further aligning PEMV with sample feature vectors, drawing inspiration from Yu et al. (2024). In contrast, our PEMVs not only serve as anchors but also refine the spatial distribution of utterance representations, addressing hard-to-classify samples and enhancing the overall distinction between emotion categories.

To validate the effectiveness of PEMV in guiding feature learning, we conducted extensive experiments on three widely-used ERC benchmark datasets, achieving state-of-the-art performance. The main contributions of this paper are summarized as follows:

- We propose a novel method that utilizes PEMV to optimize the spatial distribution of feature vectors, addressing the challenges of hard-to-classify samples and class imbalance in ERC tasks.
- To the best of our knowledge, this is the first work to consider optimizing the spatial representation of hard-to-classify samples.
- Extensive experiments on three ERC benchmark datasets demonstrate the effectiveness of our model, achieving state-of-the-art performance.

2 Related work

Existing research methods for ERC primarily focus on the structural modeling of dialogues and the interaction between different emotions. These methods can be broadly categorized into four main types: sequence modeling methods, graph-based methods, knowledge-enhanced methods, and large language model (LLM) methods.

(1) **Sequence modeling methods** typically treat the utterances in a conversation as sequential inputs, using recurrent neural networks or pre-trained language models to capture contextual information. Early works, such as ICON (Hazarika et al., 2018) and HiGRU (Jiao et al., 2019), employed gated recurrent units (GRU) to capture contextual information in conversations. CoMPM (Lee and Lee, 2022) integrates pre-trained language models to model contextual and speaker memory information.

(2) **Graph-based methods** construct graphs to model the relationships between utterances and speakers in a conversation. DialogGCN (Ghosal et al., 2019) represents utterances as nodes and uses different types of edges to model relationships within and between speakers. SIGAT (Jia et al., 2023) introduces a dual-connection graph attention network to model the interactive influence of speaker-aware and sequence-aware information, enhancing contextual representation.

(3) **Knowledge-enhanced methods** introduce external knowledge into ERC tasks to improve the accuracy of emotion recognition. COSMIC (Ghosal et al., 2020) incorporates social commonsense to help identify hidden emotional information. EmoTransKG (Zhao et al., 2024) models emotion transformations via a knowledge-enhanced Emotion Graph.

(4) **LLM methods** have emerged with the advent of LLMs. InstructERC (Lei et al., 2023) and DialogueLLM (Zhang et al., 2023b) leverage instruction-based construction and fine-tuning of large language models for ERC tasks. CKERC (Fu, 2024) and BiosERC (Xue et al., 2024) introduce tasks related to identifying speaker-related content into LLM training to capture implicit speaker cues.

In recent years, contrastive learning has been increasingly integrated into ERC methods as an effective technique to enhance feature representations and address challenges in emotion recognition. EACL (Yu et al., 2024) leverages label encodings as anchors to guide utterance representations and enhance the separation between similar emotions. In contrast, our method focuses on addressing hard-to-classify samples within the same emotion category, utilizing PEMV to directly guide the spatial distribution of utterances. SPCL (Song et al., 2022) introduces a prototypical contrastive loss to tackle imbalanced classification, incorporating a curriculum learning strategy to improve robustness in challenging samples, whereas our approach emphasizes refining the representation of utterances by utilizing PEMV for optimization.

3 Methodology

3.1 Problem Definition

In the task of ERC, given a conversation $C = [u_1, u_2, \ldots, u_N]$ consisting of N consecutive utterances and M speakers $S = [s_1, s_2, \ldots, s_M]$ (where $M \geq 2$), each utterance u_i is spoken by a specific speaker s_j. The goal of ERC is to predict the emotion label e_i for each utterance u_i, that is, to identify the emotional state of the corresponding speaker at each turn of the conversation. In this paper, we focus on the real-time setting of ERC, where the model can only utilize a portion of the previous turns $[(s_1, u_1), (s_2, u_2), \ldots, (s_t, u_t)]$ as input to predict the emotion label e_t for the current utterance u_t.

3.2 Model Overview

As illustrated in Figure 3, our proposed model comprises four main components: Utterance Text Feature Extraction, Proximal Emotion Mean Vectors Generation, Proximal Emotion Mean Vectors Guide Feature Learning, and Proximal Emotion Mean Vectors Adaptation. We first extract features from each utterance and store them in dynamic queues corresponding to their respective labels (referred to as emotion feature queues). Subsequently, we compute the mean of all feature vectors within each emotion feature queue to obtain the **P**roximal **E**motion **M**ean **V**ectors (**PEMV**). Next, we utilize PEMVs to guide feature learning, focusing on two primary objectives: first, to pull feature vectors within the same category that are distant from the PEMV closer to it, and second, to maximize the angular differences between PEMVs of different categories. Finally, we employ the generated PEMVs as classifiers, adjusting their decision boundaries to further align the utterance features with the PEMVs.

3.3 Utterance Feature Extraction

We utilize the SimCSE-Roberta-Large model (Gao et al., 2021) to extract text feature vectors for each utterance. To predict the emotion of a given utterance, we use the current utterance along with its preceding utterances as input. Specifically, for a given utterance u_t at timestamp t, we take the preceding k utterances as the context, forming the input sequence:

$$x_t = [s_{t-k}, u_{t-k}, \ldots, s_t, u_t, \text{Prompt}] \quad (1)$$

where s_t represents the speaker of utterance u_t. The prompt used to align the downstream task with the large semantic information learned by the language model during the pre-training stage (Liu et al., 2023) is: "For utterance u_t, the speaker s_t feels <mask>". We take the hidden state of the <mask> token from the final layer of the SimCSE-Roberta-Large model as the feature representation of the utterance.

3.4 PEMV-Guided Learning

3.4.1 PEMV Generation

We consider the PEMV as a prototypical feature of an emotion category, which is used to guide the spatial distribution of sample vectors within that category. In the process of generating PEMV, we maintain a fixed-size emotion feature queue for

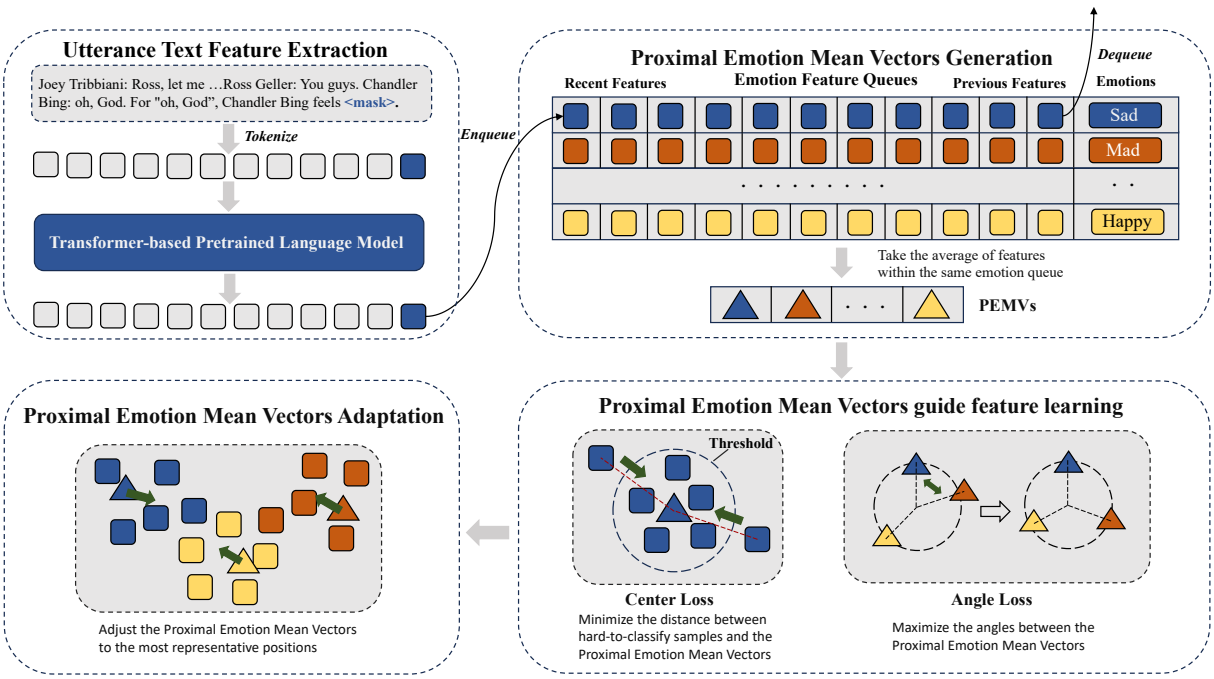


Figure 3: The overall structure of the proposed model. First, utterance text features are extracted using a Transformer-based pretrained language model and stored in dynamic emotion feature queues corresponding to their labels. These features are then used to generate Proximal Emotion Mean Vectors (PEMV). PEMVs guide the distribution of features in the spatial representation by encouraging tighter clustering within the same category and greater separation between different categories. After model training, PEMVs also function as a classifier to further align dialogue text features with their respective emotion categories.

each emotion category, which stores the most recent textual feature representations of that category. Specifically, the queue size is denoted as L_i, used to store feature vectors for each category, with the number of categories adjusted according to different datasets. For the i-th category, the feature queue is defined as:

$$Q_i = [z_i^1, z_i^2, \cdots, z_i^{L_i}] \tag{2}$$

where z_i^j represents the j-th feature vector belonging to category i, L_i denotes the length of the feature queue for category i. Specifically, the feature vector is derived as:

$$z_i^j = \text{Encoder}(x_t^j) \tag{3}$$

Each time a new feature representation z_i is generated, if the size of the queue Q_i has reached L_i, the oldest element in the queue is removed, and the gradient of z_i is detached before it is pushed into the queue. To generate the PEMV for the i-th category, we compute the mean of all the samples in Q_i, yielding the PEMV:

$$T_i = \frac{1}{L_i} \sum_{j=1}^{L_i} z_i^j \tag{4}$$

where vector T_i represents the prototypical emotional features of category i.

3.4.2 Optimizing Spatial Representation

The generated PEMV is employed to guide the spatial distribution of sample vectors, focusing on two key aspects: First, within the same category, PEMV addresses vectors that deviate significantly from their category center, employing a penalization mechanism to pull these outliers closer. Second, across different categories, PEMV aims to enhance the angular separation between vectors from distinct categories, thereby ensuring improved inter-class distinction.

Center Loss is utilized to address vectors within the same category that deviate substantially from the PEMV. This deviation is quantified by computing the Euclidean distance between a sample feature f_i and its category's PEMV T_i . Specifically, the distance between a sample feature f_i and the category PEMV T_i is calculated as:

$$d(f_i, T_i) = \sqrt{\sum_{j=1}^{D} (f_i^j - T_i^j)^2} \tag{5}$$

where D denotes the feature dimension. Samples

with a distance exceeding a predefined threshold are penalized. Samples with a distance exceeding a predefined threshold are penalized. A dynamic threshold μ is computed using the mean and the variance of the distances:

$$\overline{d} = \frac{1}{N}\sum_{i=1}^{N} d(f_i, T_i) \tag{6}$$

$$\mu = \overline{d} + \frac{c \cdot \left(\sum_{i=1}^{N}(d(f_i, T_i) - \overline{d})^2\right)}{N} \tag{7}$$

where $\overline{d}$ is the average distance, N represents the batch size and c is a scaling factor applied to the variance term. The final Center Loss is then calculated as:

$$\mathcal{L}_{cen} = \frac{1}{N}\sum_{i=1}^{N} max(d(f_i, T_i) - \mu, 0) \tag{8}$$

Center loss penalizes the sample features f_i whose distances from their respective category centers T_i exceed the threshold μ, thereby encouraging the feature vectors to be closer to their corresponding PEMV.

Angle Loss is employed to enhance inter-class differentiation by maximizing the angular separation between categories. For each category's PEMV T_i , we first compute the mean vector g as:

$$g = \frac{1}{C}\sum_{i=1}^{C} T_i \tag{9}$$

where C is the total number of categories. The PEMVs are then centered and normalized to facilitate the computation of cosine similarity:

$$T_i' = \frac{T_i - g}{|T_i - g|} \tag{10}$$

We compute the maximum cosine similarity between all categories and use it to derive the Angle Loss:

$$\mathcal{L}_{ang} = -\frac{1}{C}\sum_{i=1}^{C} \arccos\left(\max_{i \neq j}\left(T_i' \cdot T_j'\right)\right) \tag{11}$$

Minimizing this loss function maximizes the angular separation between PEMVs of different categories, thereby enhancing inter-class distinction.

In conjunction with Center Loss and Angle Loss, we employ the **Cross-Entropy Loss** to ensure accurate emotion classification. The Cross-Entropy Loss is defined as:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{12}$$

where $y_{i,c}$ is the true label indicator (1 if sample i belongs to class c, otherwise 0), and $p_{i,c}$ represents the predicted probability of sample i belonging to class c.

Additionally, we incorporate **SupCon Loss** to further enhance the model's feature learning capabilities. The SupCon Loss is defined as:

$$C(z_i, z_j) = \frac{sim(z_i, z_j)}{\tau} \tag{13}$$

$$\mathcal{L}_{sp} = -\sum_{i=1}^{N} \frac{\sum_{j \in P(i)} \log \frac{e^{C(z_i, z_j)}}{\sum_{k \neq i} e^{C(z_i, z_k)}}}{|P(i)|} \tag{14}$$

where $sim_(z_i, z_j)$ is the cosine similarity between the feature representations of z_i and z_j, and τ is a temperature scaling parameter used to control the sharpness of the similarity distribution.

The overall loss function is defined as a weighted sum of the four components:

$$\mathcal{L} = \sum_{i=1}^{4} \lambda_i \mathcal{L}_i \tag{15}$$

where $\mathcal{L}_i$ represents one of the four loss components: $\mathcal{L}_{cen}$, $\mathcal{L}_{ang}$, $\mathcal{L}_{ce}$, and $\mathcal{L}_{sp}$. The corresponding weights λ_i are hyperparameters (λ_1, λ_2, λ_3, λ_4) that control the trade-off between these four loss components.

By optimizing this composite loss, we guide the spatial distribution of sample vectors, ensuring that they are closer to their respective category centers, well-separated from other categories, and accurately classified.

3.4.3 PEMV Adaptation

After the PEMV-guided learning process, following the method of Yu et al. (2024), we further adapt the PEMV to enhance its classification performance. During the adaptation process, we freeze the parameters of the language model and treat the PEMV T_i $(i = 1, \ldots, s)$, as trainable parameters to be optimized. To ensure effective alignment of the PEMV, we use alignment loss function $\mathcal{L}_{\text{ada}}$, which is de-

fined as:

$$\mathcal{L}_{\text{ada}} = -\frac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{s} y_{i,j}\log \hat{y}_{i,j}$$
$$= -\frac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{s} y_{i,j}\log \frac{e^{C_{i,j}}}{\sum_{k=1}^{s} e^{C_{i,k}}} \quad (16)$$

where $C_{i,j}$ represents the cosine similarity between the i-th feature vector and the j-th PEMV.

In the prediction phase, for each utterance representation r_i , we calculate its similarity with every PEMV T_j using the cosine similarity. The predicted emotion label is determined by selecting the PEMV with the highest similarity score:

$$\hat{y}_i = \arg\max_j \text{sim}(r_i, T_j) \quad (17)$$

where r_i represents the feature vector of the utterance x_i, and T_j corresponds to the PEMV of class j.

4 Experimental Setting

4.1 Datasets

We conducted experiments on three widely used ERC benchmark datasets: MELD(Poria et al., 2019), EmoryNLP(Zahiri and Choi, 2018) and IEMOCAP(Busso et al., 2008). The statistical information for the three datasets is presented in Table 1.

(1) **MELD**: This dataset is derived from the TV show Friends, containing 1,433 dialogues and 13,708 utterances. Each utterance is labeled with one of seven emotions: surprise, neutral, anger, sadness, disgust, joy, and fear.

(2) **EmoryNLP**: Also sourced from Friends, this dataset includes 897 dialogues and 12,606 utterances. Unlike MELD, EmoryNLP uses a different set of emotion labels, including neutral, joyful, peaceful, powerful, scared, mad, and sad.

(3) **IEMOCAP**: IEMOCAP is a two-speaker conversation dataset comprising 151 dialogues and a total of 7,433 utterances. Each utterance is annotated with one of six emotions: excited, frustrated, sad, neutral, angry, and happy. Since there is no official validation set for this dataset, we followed the method of Sun et al. (2021) and used the last 20 dialogues from the training set as the validation set.

Notably, all three datasets suffer from class imbalance, with further details provided in Appendix A.

Dataset	Dialogues			Utterances			CLS
	train	dev	test	train	dev	test	
MELD	1038	114	280	9989	1109	2610	7
EmoryNLP	659	89	79	9934	1344	1328	7
IEMOCAP	100	20	31	4890	920	1623	6

Table 1: The statistics of three datasets. CLS represents the number of emotion categories in each dataset.

4.2 Baseline Models

To evaluate the effectiveness of the proposed model, we conduct a comparative analysis against several existing methods. A detailed description of these comparative models is provided in Appendix B.

4.3 Implementation Details

The utterance feature extraction model we employed is initialized with SimCSE-Roberta-Large (Gao et al., 2021) parameters. In all experiments presented in this paper, we select the optimal checkpoint based on performance on the development set, and subsequently use this checkpoint to evaluate and report results on the test set. All experiments were conducted on a single NVIDIA A100-SXM4-80GB GPU, utilizing the PyTorch 2.0 framework. Further experimental details are provided in Appendix C.

4.4 Metrics

Following previous works (Zhao et al., 2022; Song et al., 2022), we choose weighted-F1 score as the metric for all experiments.

5 Results and Analysis

5.1 Main Results

In Table 2, we present the weighted-F1 scores for the MELD, EmoryNLP, and IEMOCAP datasets, comparing the performance of our PEMV model with several strong baselines, including sequence modeling and graph-based approaches. The results suggest that PEMV delivers competitive performance, consistently yielding strong outcomes across all datasets.

On the MELD dataset, PEMV achieves a weighted-F1 score of 67.95, reflecting a modest improvement of 0.70% over SPCL+CL. For the EmoryNLP dataset, PEMV attains a weighted-F1 of 40.97, surpassing EACL by 0.73%. These results suggest that PEMV is effective in addressing more challenging classification cases, despite the inherent complexity of the dataset. On the IEMO-

Methods	MELD	EmoryNLP	IEMOCAP	Average
Sequence modeling methods				
SPCL+CL (Song et al., 2022)	67.25	40.94	69.74	59.31
MuCDN (Zhao et al., 2022)	65.37	40.09	-	-
ChatGPT 3-shot (Zhao et al., 2023)	58.35	35.92	48.58	47.62
ERNetCL (Li et al., 2024)	66.31	39.71	69.73	58.58
COSMIC(2020) (Ghosal et al., 2020)	65.21	38.11	65.28	56.20
+E-TransKG (Zhao et al., 2024)	-	39.06	68.39	-
CoMPM (Lee and Lee, 2022)	66.52	37.37	66.33	56.74
+CLED (Kang and Cho, 2024)	66.00	38.76	67.65	57.47
EACL(Yu et al., 2024)	67.12	40.24	70.41	59.26
Graph-based methods				
DAG-ERC (Shen et al., 2021)	63.65	39.02	68.03	56.90
DAG-ERC+HCL(Yang et al., 2022)	63.89	39.82	68.73	57.48
SIGAT (Jia et al., 2023)	66.18	39.95	70.17	58.77
DialogueGCN (Ghosal et al., 2019)	64.09	38.23	65.30	55.87
+E-TransKG (Zhao et al., 2024)	-	38.80	67.11	-
PEMV(ours)	**67.95**	**40.97**	**70.65**	**59.86**

Table 2: Comparison of Weighted-F1 between PEMV and Baseline Models. The bold value represents the **best performance**, while the underlined value indicates the second-best performance. Through the t-test, the experimental results on all three datasets demonstrate statistical significance ($p < 0.05$).

CAP dataset, PEMV records a weighted-F1 score of 70.65, slightly exceeding EACL by 0.24%.

Furthermore, Table 3 provides a more detailed breakdown of performance across different emotion categories for each dataset. On the MELD dataset, PEMV achieves consistent performance across emotion categories, with particularly notable improvements in "Neutral" and "Anger". This is partly due to the effect of the PEMV's Distance Penalty, as shown in Table 5, where the number of samples from these two emotion categories that were too distant from the PEMV has decreased after applying the Distance Penalty, indicating its ability to pull back hard-to-classify samples. Similarly, on the EmoryNLP dataset, PEMV exhibits competitive performance in the "Mad" and "Scared" categories, slightly surpassing SPCL+CL and EACL. On the IEMOCAP dataset, PEMV demonstrates its strength in capturing complex emotions such as "Frustration", showing an overall improvement over both SPCL+CL and EACL.

(a) MELD

Methods	Fear	Neu	Ang	Sad	Dis	Surp	Joy	Avg	W-f1
SPCL+CL	**26.59**	77.92	54.40	43.53	30.94	59.26	60.34	50.43	65.74
EACL	23.54	80.44	54.01	42.41	**33.86**	**60.48**	**65.22**	51.42	67.12
PEMV	22.78	**81.18**	**56.94**	**45.76**	32.38	59.76	64.95	**52.67**	**67.95**

(b) EmoryNLP

Methods	Joy	Sad	Pow	Mad	Neu	Pea	Sca	Avg	W-f1
SPCL+CL	53.52	31.61	10.28	44.21	**51.40**	16.83	39.51	35.34	39.52
EACL	52.73	30.77	15.27	41.97	49.76	23.48	41.18	36.45	40.24
PEMV	**53.59**	**33.04**	**15.38**	**47.39**	47.51	**24.33**	**44.09**	**37.90**	**40.97**

(c) IEMOCAP

Methods	Exc	Fru	Sad	Neu	Ang	Hap	Avg	W-f1
SPCL+CL	66.72	63.96	80.03	72.29	64.82	43.96	65.30	67.19
EACL	**71.27**	67.76	**81.80**	73.32	67.54	**51.29**	**68.81**	**70.41**
PEMV	64.91	**71.69**	81.64	**77.30**	**70.36**	43.71	68.27	**70.65**

Table 3: Weighted-F1 scores for each class between PEMV, SPCL+CL, and EACL across three benchmark datasets

5.2 Ablation Study

To evaluate the contributions of various components in our PEMV framework, we conducted ablation studies on the MELD, EmoryNLP, and IEMOCAP datasets, as shown in Table 4. The results suggest that removing any component leads to a decline in performance. For instance, the exclusion of PEMV-Guided Learning results in a decrease of 1.06% on MELD, indicating its importance in guiding feature learning and supporting emotion separation. Similarly, omitting the PEMV Distance Penalty corresponds to a drop of 0.58%, which points to its role in reducing within-class variance. Although the removal of Angle Adjustment leads to only a minor decrease of 0.13% on EmoryNLP, it still suggests its contribution to refining the spatial distribution of emotion vectors.

Moreover, the absence of the Classification Objective results in a reduction of 0.63% on MELD, highlighting its necessity for aligning the feature space with the emotion categories, consistent with the findings of Gunel et al. (2020). The impact of excluding SupCon Learning is also noticeable, with

Methods	MELD	EmoryNLP	IEMOCAP
PEMV	67.95	40.97	70.65
w/o PEMV-Guided Learning	66.89 (1.06 ↓)	40.22 (0.75 ↓)	69.81 (0.84 ↓)
w/o PEMV Distance Penalty	67.37 (0.58 ↓)	40.60 (0.37 ↓)	70.22 (0.43 ↓)
w/o Angle Adjustment	67.51 (0.44 ↓)	40.84 (0.13 ↓)	70.29 (0.36 ↓)
w/o Classification Objective	67.32 (0.63 ↓)	40.52 (0.45 ↓)	70.25 (0.40 ↓)
w/o SupCon Learning	67.40 (0.55 ↓)	40.38 (0.59 ↓)	70.33 (0.32 ↓)
w/o PEMV Adaptation	67.12 (0.83 ↓)	40.45 (0.52 ↓)	70.51 (0.14 ↓)

Table 4: Results of ablation study on three benchmark datasets

a modest decline of 0.55% on MELD, which reflects its potential in enhancing emotion distinction through better clustering of similar representations. Lastly, not incorporating PEMV Adaptation leads to a 0.83% drop on MELD, underscoring its value in fine-tuning the positions of emotion vectors relative to the data distribution.

5.3 Effect of PEMV-Guided Spatial Distribution

We conducted additional experiments to explore the impact of the PEMV distance penalty and angle adjustment on optimizing the distribution of feature vectors in the embedding space.

For the three benchmark datasets, we applied the distance threshold μ provided in Equation 7 and the variance scaling factor c from Table 7 in Appendix C. We calculated the percentage of samples exceeding the threshold μ in each emotion category, as well as the total number of such samples across the dataset, for both the fully trained PEMV model and the model trained with cross entropy and supervised contrastive learning. The results are presented in Table 5. The proportion of hard-to-classify samples in almost all emotion categories has decreased, which aligns with the trend shown in Table 3 for fine-grained emotion classification results. For instance, on the MELD dataset, "Neutral" and "Anger" performed better compared to other well-performing models, and the proportion of hard-to-classify samples in these two categories also showed a noticeable reduction. We present several examples of hard-to-classify samples in Appendix D.

We visualized the spatial distribution of the PEMVs for different emotion categories before and after training, as presented in Appendix E. After the angle adjustment, the PEMVs' distribution in the feature space becomes more dispersed.

To further analyze the spatial distribution, we visualized the emotion category distributions on the MELD dataset after training with both the CE+SupCon and PEMV methods, as shown in Figure 4. The PEMV approach resulted in fewer samples exceeding the distance threshold τ (Eq. 7) within the same category, while achieving better separation between different categories.

(a) MELD

Methods	Fear	Neu	Ang	Sad	Dis	Surp	Joy	Cnt
CE + SupCon	4.74	4.91	4.91	4.44	4.33	5.56	5.30	510
PEMV	4.48	4.01	4.06	4.23	4.43	5.23	5.34	440

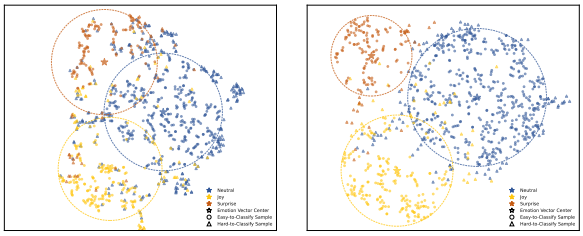
(b) EmoryNLP

Methods	Joy	Sad	Pow	Mad	Neu	Pea	Sca	Cnt
CE + SupCon	5.68	5.96	3.70	5.30	3.13	4.56	4.75	447
PEMV	4.99	5.66	3.06	4.09	2.57	3.67	4.44	383

(c) IEMOCAP

Methods	Exc	Fru	Sad	Neu	Ang	Hap	Cnt
CE + SupCon	8.05	8.54	8.29	10.01	8.26	9.08	597
PEMV	7.68	7.77	7.03	9.29	8.68	9.52	482

Table 5: The percentage of samples exceeding the threshold μ for each emotion category and the total number of such samples across different methods on three benchmark datasets.



(a) CE+SupCon (b) PEMV

Figure 4: Comparison of sample vector space distributions trained using the CE+SupCon method and the PEMV method. Triangles represent hard-to-classify samples, while the radius of the circles corresponds to the value of the parameter τ.

6 Conclusion

We propose a method that aims to more effectively tackle the persistent challenges of difficult-to-classify samples and imbalanced label distributions in ERC tasks. Our approach leverages **P**roximal **E**motion **M**ean **V**ectors (**PEMV**) to optimize the spatial distribution of feature representations, ensuring that they more accurately align with their corresponding emotion categories. By capturing the inherent structure of these categories, PEMV aids in refining the positioning of feature vectors within the space, specifically by reducing the distance between each sample and its corresponding PEMV through Center Loss, particularly for those samples

that are significantly distant from the PEMV. Simultaneously, Angle Loss maximizes the angular separation between the PEMVs of different emotion categories, thereby maintaining clear boundaries among them. Moreover, employing PEMV as a classifier enhances the model's capacity to adapt to the evolving feature space, leading to more stable and precise classification. Our extensive experiments on multiple benchmark datasets confirm the effectiveness of this approach, showcasing consistent improvements in both feature representation quality and the accuracy of emotion recognition in dialogues.

Limitations

Our model is trained and evaluated solely on the text modality derived from the dataset. However, as illustrated in Figure 2, nearly identical utterances can convey different emotions depending on their contextual nuances. This variability underscores the inherent limitations of relying exclusively on text for ERC task.

The complexities of human communication involve various modalities that significantly contribute to the emotional depth and intent behind spoken language. Elements such as tone of voice, facial expressions, and even accompanying visual cues can alter the perceived emotion of an utterance. (Bansal et al., 2022) Therefore, the sole dependence on textual data restricts our model's ability to fully grasp the multifaceted nature of emotions in dialogue.

To address this limitation, future research could explore the integration of multimodal features, such as image and audio, alongside text. For example, incorporating audio signals could capture the intonation and stress patterns that convey subtle emotional cues, while visual inputs could provide insights into the speaker's expressions and gestures.

References

Keshav Bansal, Harsh Agarwal, Abhinav Joshi, and Ashutosh Modi. 2022. Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 44–56.

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.

Yumeng Fu. 2024. Ckerc: Joint large language models with commonsense knowledge for emotion recognition in conversation. *arXiv preprint arXiv:2403.07260*.

T Gao, X Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. 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.

Zhaohong Jia, Yunwei Shi, Weifeng Liu, Zhenhua Huang, and Xiao Sun. 2023. Speaker-aware interactive graph attention network for emotion recognition in conversation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(12):1–18.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 397–406.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron

Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Joosung Lee and Wooin Lee. 2022. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5669–5679.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Jiang Li, Xiaoping Wang, Yingjian Liu, and Zhigang Zeng. 2024. Ernetcl: A novel emotion recognition network in textual conversation based on curriculum learning strategy. *Knowledge-Based Systems*, 286:111434.

Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *arXiv preprint arXiv:2003.01478*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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*, pages 527–536.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2949–2958.

Jieying Xue, Minh Phuong Nguyen, Blake Matheny, and Le Minh Nguyen. 2024. Bioserc: Integrating biography speakers supported by llms for erc tasks. *arXiv preprint arXiv:2407.04279*.

Haiqin Yang and Jianping Shen. 2021. Emotion dynamics modeling via bert. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. 2022. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11595–11603.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2024. Emotion-anchored contrastive learning framework for emotion recognition in conversation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4521–4534.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Mian Zhang, Xiabing Zhou, Wenliang Chen, and Min Zhang. 2023a. Emotion recognition in conversation from variable-length context. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yazhou Zhang, Mengyao Wang, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023b. Dialoguellm: Context and emotion knowledge-tuned llama models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*.

Huan Zhao, Xupeng Zha, and Zixing Zhang. 2024. Emotranskg: An innovative emotion knowledge graph to reveal emotion transformation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12098–12110.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Weixiang Zhao, Yanyan Zhao, and Bing Qin. 2022. Mucdn: Mutual conversational detachment network for emotion recognition in multi-party conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7020–7030.

Appendix

A Dataset Details

The number of samples for each category in the MELD, EmoryNLP, and IEMOCAP datasets is shown in Table 6. It can be observed that all three datasets exhibit varying degrees of class imbalance, with the MELD dataset showing the most prominent class imbalance.

(a) MELD

	Fear	Neu	Ang	Sad	Dis	Surp	Joy	CIR
Train	268	4710	1109	271	1743	1205	683	17.57
Dev	40	470	153	22	163	150	111	21.36
Test	50	1256	345	68	402	281	208	25.12

(b) EmoryNLP

	Joy	Sad	Pow	Mad	Neu	Pea	Sca	CIR
Train	2184	671	784	1076	3034	900	1285	4.52
Dev	289	75	134	143	393	132	178	5.24
Test	282	98	145	113	349	159	182	3.56

(c) IEMOCAP

	Exc	Fru	Sad	Neu	Ang	Hap	CIR
Train	705	1215	636	1140	786	408	2.98
Dev	37	253	203	184	147	96	6.84
Test	299	381	245	384	170	144	2.67

Table 6: Statistical Analysis of MELD, EmoryNLP, and IEMOCAP Datasets. The Class Imbalance Ratio (CIR) refers to the ratio of the sample size between the most frequent and least frequent classes.

B Baseline Models

The detailed descriptions of the comparative models we have selected are provided as follows:

(1) Sequence modeling methods: **SPCL+CL** (Song et al., 2022) uses prototypical contrastive learning to address class imbalance, enhanced by curriculum learning for improved performance. **MuCDN** (Zhao et al., 2022) introduces a Mutual Conversational Detachment Network to effectively model emotional dynamics in multi-party conversations. **ChatGPT** (Zhao et al., 2023) presents findings on its performance in the 3-shot setting. **ERNetCL** (Li et al., 2024) employs a curriculum learning strategy to optimize emotion recognition in conversation by capturing contextual cues effectively. **COSMIC** (Ghosal et al., 2020) leverages commonsense knowledge to enhance utterance-level emotion recognition, effectively addressing context propagation and emotion shift detection challenges. **CoMPM** (Lee and Lee, 2022) integrates pre-trained memory with context modeling, improving emotion recognition in conversation without relying on structured data, enabling cross-linguistic adaptability. **EACL** (Yu et al., 2024) introduces an Emotion-Anchored Contrastive Learning framework, utilizing label encodings as anchors to enhance distinguishability of similar emotions.

(2) Graph-based methods: **DAG-ERC** (Shen et al., 2021) utilizes a directed acyclic graph to model conversational context, improving emotion recognition by capturing long-distance and nearby context relationships effectively. **DAG-ERC+HCL** (Yang et al., 2022) employs a hybrid curriculum learning framework to improve ERC by addressing emotion shifts and confusing emotions progressively. **SIGAT** (Jia et al., 2023) introduces a speaker-aware interactive graph attention network to capture both sequence and speaker information, enhancing performance with richer contextual representations. **DialogueGCN** (Ghosal et al., 2019) employs graph convolutional networks to capture speaker dependencies, enhancing emotion recognition by improving conversational context modeling.

E-TransKG (Zhao et al., 2024) establishes an innovative Emotion Knowledge Graph to model emotion transformations in conversations. **CLED** (Kang and Cho, 2024) employs supervised contrastive learning and emotion dynamics augmentation to address imbalanced emotion distribution in ERC, particularly enhancing neutral emotion recognition.

C Implementation Details

PEMV model is initialized with parameters from SimCSE-Roberta-Large (Gao et al., 2021). We employ a grid-search method to optimize the hyperparameters, specifically setting λ_1 in {0.01, 0.03, 0.05}, λ_2 in {0.05, 0.1, 0.15}, λ_3 in {0.1, 0.3, 0.5, 0.9}, λ_4 in {0.1, 0.3, 0.5, 0.9}, and c in {1, 1.5, 2}. The specific hyperparameter settings for the model on the three datasets are presented in Table 7.

Hyperparameters	MELD	EmoryNLP	IEMOCAP
λ_1	0.05	0.03	0.01
λ_2	0.1	0.1	0.1
λ_3	0.9	0.3	0.1
λ_4	0.1	0.3	0.9
Temperature τ	0.1	0.1	0.1
variance scaling factor c	2	2	1.5
Maximum length	256	256	256
Batch Size	64	64	64
Epochs	8	8	8
SimCSE Learning Rate	1e-5	1e-5	1e-5
FFN Learning Rate	1e-3	1e-3	1e-3
Dropout	0.1	0.1	0.1

Table 7: Hyperparameters of PEMV on three benchmark datasets

D Examination of Hard-to-Classify Instances

We employed the threshold μ from Equation 7 and the variance scaling factor c from Table 7 to train

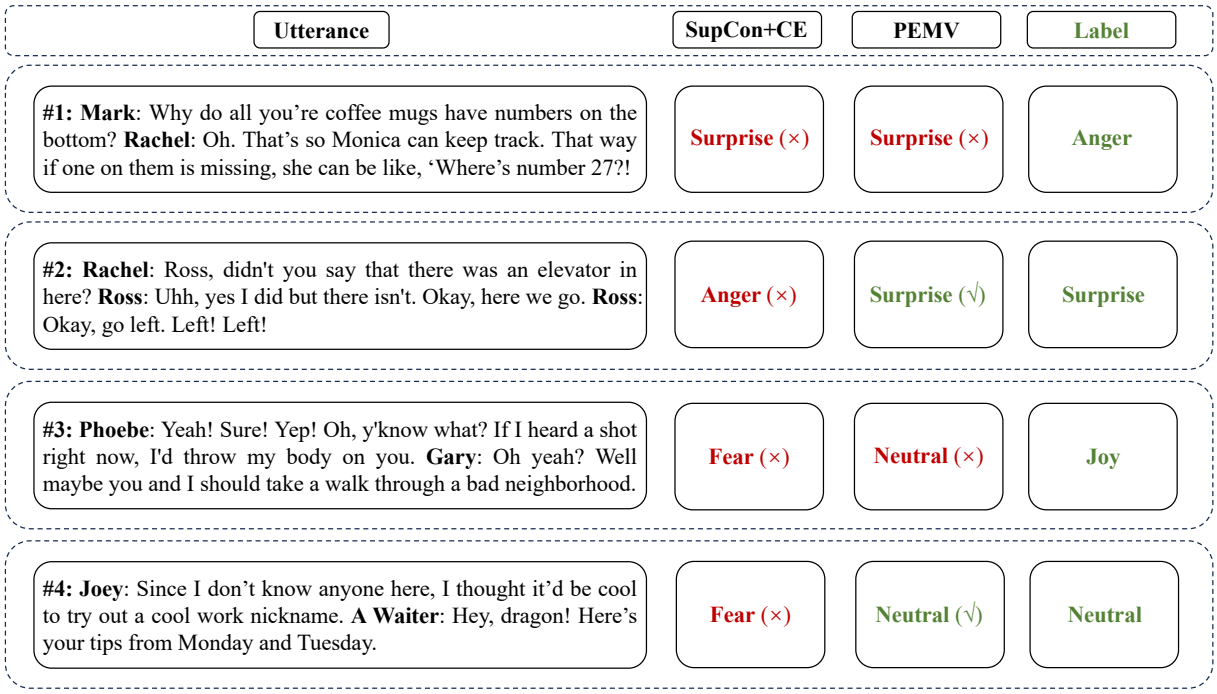


Figure 5: Comparison of Hard-to-Classify Sample Instances between the SupCon+CE Model and the PEMV Model on the MELD Test Set.

the SupCon+CE model and the PEMV model on the MELD dataset, respectively. Figure 5 presents several instances of hard-to-classify samples from the test set for both models.

From Sample #1, it can be observed that emotional expression in certain dialogue contexts is not always direct. In this case, Rachel's words appear lighthearted and humorous, yet are labeled as "anger". When emotional cues are not overt, the model struggles to accurately extract the sentiment from the literal information, potentially causing its embedding to deviate from the class center. Both models misclassified this sample as "surprise".

Sample #2 indicates that some samples require strong contextual associations. The emotional label for this sample is "surprise", but the isolated sentence lacks a clear expression of surprise. The SupCon+CE model incorrectly categorized it as "anger", while our PEMV model was able to classify it correctly.

Sample #3 demonstrates that some dialogues contain a mix of emotions. This sample features a complex emotional expression of irony and humor, making it challenging for the model to identify the core sentiment of "joy", leading to a deviation in embedding. The SupCon+CE model misclassified it as "fear", whereas the PEMV model misclassified it as "neutral".

Sample #4 reveals that some neutral emotions can be ambiguous. Due to the lack of distinct emotional characteristics, their embeddings are prone to drifting away from the class center. In certain cases, neutral sentences do not significantly differ from other emotional categories, making it difficult for the model to make clear judgments, resulting in deviations. In this sample, the PEMV model classified it correctly, which partially demonstrates its effectiveness in guiding spatial distribution.

These examples illustrate the challenges faced by emotion recognition models when dealing with difficult-to-classify samples. They highlight how indirect emotional expressions, the necessity for contextual understanding, the presence of mixed emotions, and the ambiguity of neutral sentiments contribute to misclassification. Notably, the PEMV model demonstrates an improved ability to accurately classify certain samples, underscoring its effectiveness in addressing the complexities of emotion recognition in dialogue contexts.

E Effect of Angle Adjustment

We provide a visualization of the PEMV for various emotion categories on the MELD dataset, showcasing their spatial distribution before and after the training process, As illustrated in Figure 6. This visual representation highlights the impact of angle adjustment, demonstrating that the distribution

of PEMVs in the feature space becomes more dispersed, thereby enhancing the clarity and separability of different emotion categories.

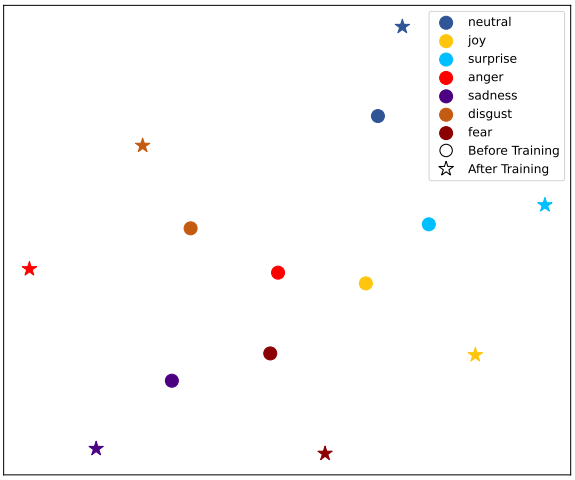


Figure 6: Visualization of PEMVs. Circles represent the positions before training, while pentagrams denote the positions after training.