Emotion-Anchored Contrastive Learning Framework for Emotion Recognition in Conversation

Anonymous ACL submission

Abstract

Emotion Recognition in Conversation (ERC) involves detecting the underlying emotion behind each utterance within a conversation. Effectively generating representations for utterances remains a significant challenge in this task. Recent works propose various models to address this issue, but they still struggle with differentiating similar emotions such as excitement and happiness. To alleviate this problem, We propose an **E**motion-**A**nchored **C**ontrastive **L**earning (EACL) framework that can generate more distinguishable utterance representations for similar emotions. To achieve this, we utilize label encodings as anchors to guide the learning of utterance representations and design an auxiliary loss to ensure the effective separation of anchors for similar emotions. Moreover, an additional adaptation process is proposed to adapt anchors to serve as effective classifiers to improve classification performance. Across extensive experiments, our proposed EACL achieves state-of-the-art emotion recognition performance and exhibits superior performance on similar emotions.

1 Introduction

Emotion Recognition in Conversation (ERC) aims to identify the emotions of each utterance in a conversation. It plays an important role in various scenarios, such as chatbots, healthcare applications, and opinion mining on social media. However, the ERC task faces several challenges. Depending on the context, similar statements may exhibit entirely different emotional attributes. Simultaneously, distinguishing conversation texts that contain similar emotional attributes is also extremely difficult (Ong et al., 2022; Zhang et al., 2023). Figure 1 is an example of a chat between a man and a woman. Differentiating between *happy* and *excited* can be challenging for machines due to their frequent occurrence in similar contexts. Appendix A exhibits quantitative analysis for emotions. This requires

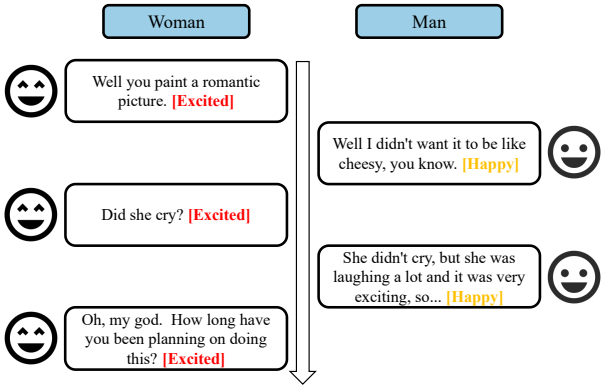


Figure 1: An example of a conversation in the IEMOCAP dataset.

the model to accurately distinguish different emotions based on the context.

Therefore, abundant efforts have been made implicitly to obtain distinguishable utterance representations from two lines, model design and representation learning. As the representative of the former line, DialogueRNN (Majumder et al., 2019) designs recurrent modules to track dialogue history for classification. Representation learning methods primarily exploit supervised contrastive learning (SupCon) (Khosla et al., 2020) for learning utterance representations. SPCL (Song et al., 2022) proposes a prototypical contrastive learning method to alleviate class imbalance problem and achieve state-of-the-art performance. However, as shown in Figure 2, our pilot fine-grained experimental results indicate that SPCL still struggles with effectively differentiating similar emotions.

To tackle the aforementioned issues, this paper presents a novel **E**motion-**A**nchored **C**ontrastive **L**earning framework (EACL). EACL utilizes textual emotion labels to generate anchors that are emotionally semantic-rich representations. These representations as anchors explicitly strengthen the distinction between similar emotions in the representation space. Specifically, we introduce a

Figure 2: Normalized confusion matrix of a pilot study for SPCL on the IEMOCAP dataset.

penalty loss that specifically targets emotions with the largest cosine similarity. This loss function encourages the corresponding emotion anchors to exhibit improved angular separation in the representation space. By doing so, more separated emotion anchors guide utterance representations with similar emotions to learn larger dissimilarities, leading to enhanced discriminability. After generating separable utterance representations, we aim to compute the optimal positions of emotion anchors to which utterance representations can be assigned for classification purposes. To achieve better assignment, inspired by the two-stage frameworks (Kang et al., 2019; Menon et al., 2020; Nam et al., 2023), we propose the second stage to shift the decision boundaries of emotion anchors with fixed utterance representations and achieve better classification performance, which is simple yet effective.

We conduct experiments on three widely used benchmark datasets, the results demonstrate that EACL achieves a new state-of-the-art performance. Moreover, EACL achieves a significantly higher separability in similar emotions, which validates the effectiveness of our method.

The main contributions of this work are summarized as follows:

- We propose a novel emotion-anchored contrastive learning framework for ERC, that can generate more distinguishable representations for utterances.
- To the best of our knowledge, our method is the first to explicitly alleviate the problem of emotion similarity by introducing label semantic information in modeling for ERC, which can effectively guide representation learning.
- Experimental results show that our proposed EACL achieves a new state-of-the-art performance on benchmark datasets.

2 Related Work

2.1 Emotion Recognition in Conversation

Most of the present works adopt graph-based and sequence-based methods. DialogueGCN (Ghosal et al., 2019) builds a graph treating utterances as nodes, and models intra-speaker and inter-speaker relationships by setting different edge types between two nodes. MMGCN (Hu et al., 2021b) fuses multi-modal utterance representations into a graph. Differently, DAG-ERC (Shen et al., 2021) exploits directed acyclic graphs to naturally capture the spatial and temporal structure of the dialogue. COGMEN (Joshi et al., 2022) combines graph neural network and graph transformer to leverage both local and global information respectively.

Another group of works exploits transformers and recurrent models to learn the interactions between utterances. DialogueRNN (Majumder et al., 2019) combines several RNNs to model dialogue dynamics. DialogueCRN (Hu et al., 2021a) introduces a cognitive reasoning module. Commensense Knowledge is explored by KET (Zhong et al., 2019) and COSMIC (Ghosal et al., 2020). Cog-BART (Li et al., 2022a) employs BART (Lewis et al., 2019) to simultaneously generate responses and detect emotions with the auxiliary of contrastive learning. EmoCaps (Li et al., 2022c) and DialogueEIN (Liu et al., 2022) design several modules to explicitly model emotional tendency and inertia, local and global information in dialogue. The power of the language models is utilized by CoMPM (Lee and Lee, 2021) which learns and tracks contextual information by the language model itself and SPCL (Song et al., 2022), a prototypical supervised contrastive learning method to alleviate the data imbalance problem. SACL (Hu et al., 2023)introduces adversarial examples to learn robust representations. Our EACL goes along this track. Unlike the above approaches, HCL (Yang et al., 2022) comes up with a general curriculum learning paradigm that can be applied to all ERC models.

2.2 Supervised Contrastive Learning

Recent works (Chen et al., 2020; He et al., 2020a) in unsupervised contrastive learning provide a

similarity-based learning framework for representation learning. These methods maximize the similarity between positive samples while minimizing the similarity between negative sample pairs. To make use of supervised information, supervised contrastive learning(SupCon) (Gunel et al., 2020) aims to make the data that have the same label to be closer in the representation space and push away those that have different labels. However, SupCon works poorly in data imbalance settings. To mitigate this problem, KCL (Kang et al., 2021) explicitly pursues a balanced representation space. TSC (Li et al., 2022b) uniformly set targets in the hypersphere and enforce data representations to close to the targets. BCL (Zhu et al., 2022) regards classifier weights as prototypes in the representation space and incorporates them in the contrastive loss. LaCon (Zhang et al., 2022) incorporates label embedding for better language understanding. Our method is inspired by TSC, differently, we incorporate emotion semantics in the representation space and dynamically adjust the emotion anchors for better classification.

3 Methodology

3.1 Problem Definition

A conversation can be denoted as a sequence of utterances $\{u_1, u_2, u_3, ..., u_n\}$, each utterance u_t is uttered by one of the conversation speakers s_j. There are $m \quad (m \geq 2)$ speakers in the conversation, denoted as $\{s_1, s_2, ..., s_m\}$. Given the set of emotion labels $\mathcal{E}$ and conversation context $\{(u_1, s_{u_1}), (u_2, s_{u_2}), ..., (u_t, s_{u_t})\}$, the ERC task aims to predict emotion $e_t (e_t \in \mathcal{E})$ for current utterance u_t. $\mathcal{E}$ is a set of emotions. For instance, in the IEMOCAP dataset, $\mathcal{E}$ = {*excited*, *frustrated*, *sad*, *neutral*, *angry*, *happy*}.

3.2 Model Overview

The overview of our model is shown in Figure 3. The encoding strategy of our model adopts the paradigm of prompt learning (Section 3.3). Our training process is composed of two stages.

The first stage (Section 3.4) is called representation learning, which aims to learn more distinctive representations with emotion anchors. Concretely, we incorporate anchors containing semantic information into the contrastive learning framework and utilize them to guide the learning of utterance representations. Our objectives are (1) to bring utterances with the same emotion closer to their corresponding anchors and push utterances with different emotions farther away, and (2) to achieve a more uniform distribution of anchors in the hyperspace for better classifying different emotions.

The second stage (Section 3.5) is called emotion anchor adaptation, which aims to further improve classification performance by slightly adjusting anchors. The anchors in the first stage can help the model learn separable representations of utterances. However, separated emotion anchors are not accurately located in the optimal positions to serve as nearest-neighbor classifiers for utterances. Therefore, we design the second stage to slightly adjust the positions of emotion anchors to shift the decision boundaries for better classification performance. In this stage, we freeze the parameters of the language model and only fine-tune the emotion anchors, as shown on the right side of Figure 3. Lastly, EACL matches the utterance representations with the most similar emotion anchors to make predictions.

3.3 Prompt Context Encoding

Following previous work (Song et al., 2022), we employ pre-trained language models and adopt prompt tuning to transform the classification into masked language modeling. An effective prompt template aligns the downstream task with the large semantic information learned by the language model in the pre-training stage, which boosts the model's performance in downstream tasks.

To predict the emotion of utterance u_t, we take k utterances before timestamp t as the context to predict e_t. Formally, the input for the language model is composed as:

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

where Prompt P = "For utterance u_t, speaker s_t feels [mask]" . We take the last hidden state of [mask] as utterance representation.

3.4 Stage One: Representation Learning

In this section, we will introduce two main components of EACL in stage one: utterance representation learning and emotion anchor learning.

3.4.1 Utterance Representation Learning

The objective in this section is to acquire discernible representations for each individual utterance. To accomplish this, we employ label encodings to generate emotion anchors and incorporate them into a contrastive learning framework. By

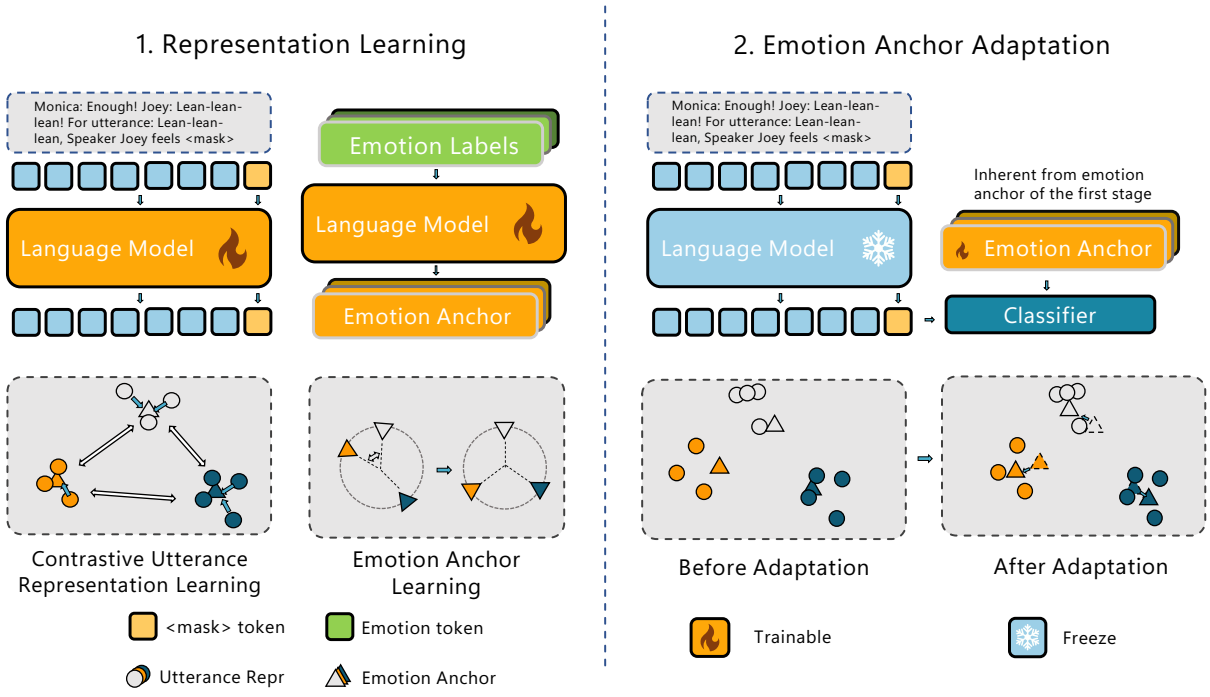


Figure 3: Overview of our proposed framework. **Left side** introduces representation learning, which is composed of utterance representation and emotion anchor learning. **Right side** describes the process of adapting emotion anchors to the optimal positions for classification.

utilizing these anchors, we can proficiently steer the process of representation learning.

Given a batch of samples $\mathcal{X} = \{x_1, x_2, \ldots, x_b\} \in \mathbb{R}^{b \times \ell}$, where b, ℓ are batch size and max length of input respectively. We feed $\mathcal{X}$ into the pre-trained language model and get the last hidden states $\mathcal{Z} = \text{Encoder}(\mathcal{X})$. Then we use the hidden state of [mask] token at the end of the sentence as the representation of utterance u_t. Finally, we obtain the representations of utterances with an MLP layer:

$$\mathcal{R} = \text{MLP}_{cl}(\mathcal{Z}_{[mask]}) \tag{2}$$

where $\mathcal{R} = \{r_1, r_2, \ldots, r_b\}$ and $\mathcal{R} \in \mathbb{R}^{b \times d}$, d is dimension of the encoder.

Similarly, we take textual emotion labels as the input of language models to obtain emotion anchors for all emotions $\mathcal{E} = \{e_1, e_2, \ldots, e_s\}$:

$$\begin{aligned} \mathcal{Z}^a &= \text{Encoder}(\mathcal{E}) \\ \mathcal{A} &= \text{MLP}_{cl}(\mathcal{Z}^a) \end{aligned} \tag{3}$$

where $\mathcal{A} \in \mathbb{R}^{s \times d}$, each row of which represents a emotion anchor. s represents the number of emotions. To ensure we get a stable anchor representation, $\mathcal{Z}_a$ is frozen in our training process.

We propose an emotion-anchored contrastive learning loss to utilize emotion label semantics for better representation learning. More specifically, in each mini-batch, we let $\mathcal{V} = \{v_1, v_2, \ldots, v_{b+s}\} = \mathcal{R} \cup \mathcal{A}$ and $\mathcal{V}_i^+$ represents the set of utterances or anchor representation that have the same label as utterance r_i except for itself. Finally, our emotion-anchored contrastive loss is as follows:

$$\begin{aligned} c_{ij} &= \text{sim}(v_i, v_j)/\tau \\ \mathcal{L}_{sup} &= \sum_{i=1}^{s+b} -\log \frac{\sum_{v_j \in \mathcal{V}_i^+} e^{c_{ij}}}{|\mathcal{V}_i^+| \sum_{v_j \in \mathcal{V}} e^{c_{ij}}} \end{aligned} \tag{4}$$

where $|\mathcal{V}_i^+|$ represents number of positive examples. τ is the temperature hyperparameter for the contrastive loss. sim represents a similarity function, we adopt cosine similarity here.

In equation 4, interactions between representations can be divided into three components: utterances-utterances, anchors-utterances, and anchors-anchors. Representations with the same label are brought closer to each other, while those with different labels are pushed farther apart. The utterances-utterances interactions are similar to traditional contrastive learning, while the anchors-utterances interactions represent the process of anchor-guided utterance representation learning. The anchors-anchors interaction ensures a better distinction between different emotions.

Recent research (Gunel et al., 2020) has indicated that combining cross-entropy loss with contrastive learning facilitates language models with more discriminative ability. Therefore cross-entropy loss is added to help improve representation learning. We additionally add a linear mapping for classification:

$$\hat{\mathcal{Y}} = \text{softmax}(\text{MLP}_{\text{ce}}(\mathcal{Z}_{[mask]})) \tag{5}$$

$$\mathcal{L}_{CE} = -\frac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{s} y_{ij}\log\hat{y}_{ij} \tag{6}$$

where $\hat{\mathcal{Y}} \in \mathbb{R}^{b\times s}$ represents the possibility distribution of b utterances over s emotions. y_{ij} represents the element in the i-th row and j-th column of $\hat{\mathcal{Y}}$. MLP_{ce} is a linear layer for classification.

3.4.2 Emotion Anchor Learning

Nevertheless, despite the implementation of the interaction between representations, the three types of interactions mentioned in Section 3.4.1 alone are insufficient to explicitly disperse the distance between the most similar emotion anchors. To further tackle the issue of similarity, we propose an anchor angle loss. This loss is designed to incentivize emotion anchors to maximize the angle between themselves and their most similar emotion anchors within the contrastive space:

$$\mathcal{L}_{Ag} = -\frac{1}{s}\sum_{i=1}^{s}\min_{j,i\neq j}\arccos\frac{\langle a_i, a_j\rangle}{\|a_i\|\|a_j\|} \tag{7}$$

where a_i represents i-th emotion anchor representation in $\mathcal{A}$.

$\mathcal{L}_{Ag}$ aims to minimize the maximal pairwise cosine similarity between all the emotion anchors. It is equivalent to maximizing the minimal pairwise angle. The more dispersed emotion anchors are, the better their capacity to recognize similar emotions.

Combining all the components mentioned in stage one, the overall loss is a weighted average of cross-entropy loss, anchor angle loss, and contrastive loss, as given in equation 8.

$$\mathcal{L} = \lambda_1(\mathcal{L}_{sup} + \lambda_2\mathcal{L}_{Ag}) + (1-\lambda_1)\mathcal{L}_{CE} \tag{8}$$

where λ_1 and λ_2 are hyper-parameters to balance loss terms.

3.5 Stage Two: Emotion Anchor Adaptation

In the first stage, we used emotion anchors generated from emotion labels to guide the convergence of utterance representations toward different emotion clusters. These emotion anchors serve as representatives for each emotion, which are suitable to function as effective nearest-neighbor classifiers for utterance representations. However, separated emotion anchors trained from stage one are not accurately located in the optimal positions which weakens the classification ability of emotion anchors. To ensure the alignment between utterance representations and emotion anchors, we propose the second stage to adapt the emotion anchors to shift the decision boundaries by training them with a small number of epochs. This approach aims to enhance the ability of emotion anchors for classification purposes.

To be more specific, we freeze the parameters of the language model and make the emotion anchors inherited from stage one $a_i (i = 1, ..., s)$ trainable parameters, which corresponds to the right side in figure 3. In order to be consistent with the representation learning, we still use the same similarity measure for adapting emotion anchors.

The loss function for emotion anchor adaptation:

$$\begin{aligned} c_{ij} &= \text{sim}(r_i, a_j)/\tau \\ \mathcal{L}_{ada} &= -\frac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{s} y_{ij}\log\hat{y_{ij}} \\ &= -\frac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{s} y_{ij}\log\frac{e^{c_{ij}}}{\sum_{k=1}^{s}e^{c_{ik}}} \end{aligned} \tag{9}$$

where c_{ij} means adjusted cosine similarity between the i-th utterance representation r_i and j-th emotion anchors a_j. τ is the same temperature hyper-parameter in stage one.

3.6 Emotion Prediction

During the inference stage, we predict emotion labels by matching each utterance representation with the nearest emotion anchor:

$$\hat{y}_i = \arg\max_j \text{sim}(r_i, a_j) \tag{10}$$

Where r_i is the representation of utterance x_i and a_j is the emotion anchor of class j.

Dataset	Dialogues			Utterances			CLS
	train	dev	test	train	dev	test	
IEMOCAP	100	20	31	4810	1000	1623	6
MELD	1038	114	280	9989	1109	2610	7
EmoryNLP	659	89	79	7551	954	984	7

Table 1: Statistics of the three datasets, where CLS is the number of classes.

4 Experiments

4.1 Experimental setup

Without specification, the language model loads the initial parameter by SimCSE-Roberta-Large (Gao et al., 2021). All experiments are conducted on a single NVIDIA A100 GPU 80GB and we implement models with PyTorch 2.0 framework. More experimental details are provided in Appendix B.

4.2 Datasets

In this section, we will introduce three adopted popular benchmark datasets: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2018) and EmoryNLP (Zahiri and Choi, 2017).

(1) **IEMOCAP**: consists of 151 videos of two speakers' dialogues with 7433 utterances. Each utterance is annotated by an emotion label from 6 classes, including *excited*, *frustrated*, *sad*, *neutral*, *angry*, and *happy*.

(2) **MELD**: is extracted from the TV show Friends. It contains about 13000 utterances from 1433 dialogues. Each utterance is labeled by one of the following 7 emotion labels: *surprise*, *neutral*, *anger*, *sadness*, *disgusting*, *joy*, and *fear*.

(3) **EmoryNLP**: contains 97 episodes, 897 scenes, and 12606 utterances from TV show Friends. It differs from MELD in that the emotional tags contained are: *joyful*, *sad*, *powerful*, *mad*, *neutral*, *scared*, and *peaceful*.

In our experiments, we only use textual modality. The detailed statistics of the three datasets are shown in Table 1.

4.3 Metrics

Following previous works (Lee and Lee, 2021; Song et al., 2022), we choose the weighted-average F1 score as the evaluation metric.

4.4 Baselines

For a comprehensive evaluation, we compare our method with the following baselines:

(1) Graph-based models: **DialogueGCN (Ghosal et al., 2019)** employs GCNs to gather context features for learning utterance representations, Shen (Shen et al., 2021) shows the performance of replacing the feature extractor with Roberta-Large. **RGAT (Ishiwatari et al., 2020)** proposes relational position encodings to model both speaker relationship and sequential information. **DAG-ERC (Shen et al., 2021)** utilizes an acyclic graph neural network to intuitively model a conversation's natural structure without introducing any external information. **DAG-ERC+HCL (Yang et al., 2022)** proposes a curriculum learning paradigm combined with DAG-ERC for learning from easy to hard.

(2) Sequence-based models: **COSMIC (Ghosal et al., 2020)** incorporates different elements of commonsense and leverages them to learn self-speaker dependency. **Cog-BART (Li et al., 2022a)** applies BART with contrastive learning to take response generation into consideration. **DialogueEIN (Liu et al., 2022)** designs emotion interaction and tendency blocks to explicitly simulate emotion inertia and stimulus. **CoMPM (Lee and Lee, 2021)** utilizes pretrained models directly learn contextual information and track dialogue history. **Emocaps (Li et al., 2022c)** devises transformer to a novel architecture, Emoformer, to extract the emotional tendency of utterance. **SACL (Hu et al., 2023)** proposes contrastive learning combined with adversarial training for robust representations. **SPCL+CL (Song et al., 2022)** combines prototypical contrastive learning and curriculum learning to tackle the emotional class imbalance issue. **ChatGPT (Zhao et al., 2023)** reports their pilot results in the 3-shot performance.

5 Results and Analysis

5.1 Main Results

Table 2 reports the result of our method and the baselines. Our model outperforms other baselines and achieves a new state-of-the-art performance on IEMOCAP, MELD, and EmoryNLP datasets. The results exhibit the effectiveness of our emotion-anchored contrastive learning framework.

Based on the results, we can observe that sequence-based methods have overall better performance than graph-based methods. Compared to the graph-based models, EACL improves a large margin over the DAG-ERC (Shen et al., 2021) which is the state-of-the-art graph-based method without introducing extra knowledge by 2.38%, 3.57%, and 1.22% on three benchmark datasets.

Compared to sequence-based methods, EACL

Methods	IEMOCAP	MELD	EmoryNLP	Average
Graph-based methods				
DialogueGCN (Ghosal et al., 2019)	64.91	63.02	38.1	55.34
RGAT (Ishiwatari et al., 2020)	66.36	62.80	37.89	55.68
DAG-ERC (Shen et al., 2021)	68.03	63.65	39.02	56.9
DAG-ERC+HCL (Yang et al., 2022)	68.73	63.89	39.82	57.48
Sequence-based methods				
COSMIC (Ghosal et al., 2020)	65.25	65.21	38.11	56.19
Cog-BART (Li et al., 2022a)	66.18	64.81	39.04	56.68
DialogueEIN (Liu et al., 2022)	68.93	65.37	38.92	57.74
CoMPM (Lee and Lee, 2021)	69.46	66.52	38.93	58.3
Emocaps (Li et al., 2022c)	69.49	63.51	-	-
SACL (Hu et al., 2023)	69.22	66.45	39.65	58.44
SPCL+CL (Song et al., 2022)	67.19	65.74	39.52	57.48
ChatGPT 3-shot (Zhao et al., 2023)	48.58	58.35	35.92	47.62
EACL (ours)	**70.41**†	**67.12**†	**40.24**†	**59.26**†

Table 2: Weighted-average F1 score of different models on benchmark datasets. Bold font and underlining indicate the best and second-best performance respectively. SPCL+CL is reproduced with the official code and uses SimCSE-Roberta-Large that EACL uses. † represents statistical significantly over baselines with t-test (p<0.05)

outperforms two contrastive learning methods, SACL and SPCL+CL by a large margin. Specifically, SPCL's use of a queue for storing class representations and prototype generation from small batches results in unstable representation learning. Significant movement of prototypes that undergo during training and the asynchronous update of queue representations with the language model's parameters lead to suboptimal utterance representations. EACL outperforms the state-of-the-art results on the IEMOCAP dataset by 0.92%, the MELD dataset by 0.6%, and the EmoryNLP dataset by 0.59%. Besides, EACL has an overwhelming performance advantage over ChatGPT, one possible reason is that the few-shot prompt setting may not be enough to achieve satisfactory performance.

Table 3 reports the fine-grained performance on benchmark datasets. EACL outperforms SPCL+CL which is the most relevant method to us in most emotion categories on all benchmark datasets. Specifically, in the IEMOCAP dataset, We have observed a significant improvement in performance on two pairs of similar emotions, *happy* and *excited* with an increase of 7.33% and 4.55%, *frustrated* and *angry* with an increase of 3.80% and 2.72% respectively. Detailed performance analysis is provided in Appendix C.

(a) 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**
Δ	+4.55	+3.80	+1.77	+1.03	+2.72	+7.33	+3.51	+3.22

(b) 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**
Δ	-3.05	+2.52	-0.39	-1.12	+2.92	+1.22	+4.88	+0.99	+1.38

(c) 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**
Δ	-0.79	-0.84	+4.99	-2.24	-1.64	+6.65	+1.67	+1.11	+0.72

Table 3: Fine-grained performance comparison between SPCL+CL and EACL for all emotions on three benchmark datasets, the F1-score is used for each class. Δ is the difference between the two models.

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 4. Removing any element of EACL makes the overall performance worse.

To validate the effects of components in the first stage, We remove the $\mathcal{L}_{Ag}$ which encourages the angle of different emotion anchors to be uniform. We can find that the lack of $\mathcal{L}_{Ag}$ results in a significant decline in the performance of nearly 0.5%, as reported in line 2 in Table 4, indicating that emotion anchor learning helps for separating utterance representations. Also, the removal of $\mathcal{L}_{CE}$ drops the performance by about 0.5% on average, the result demonstrates that supervised learning benefits the fine-tuning of language models.

Dataset	IEMOCAP	MELD	EmoryNLP
Original	70.41	67.12	40.24
w/o Emotion Anchor Learning	69.78 (0.63 ↓)	66.63(0.49 ↓)	39.90(0.34 ↓)
w/o Classification Objective	69.98(0.43 ↓)	66.24(0.88 ↓)	39.73(0.51 ↓)
w/o Anchor Inheritance	69.79(0.62 ↓)	67.03(0.09 ↓)	38.46 (1.78 ↓)
w/o Anchor Adaptation	69.67(0.74 ↓)	64.43(2.89 ↓)	39.98 (0.26 ↓)

Table 4: Ablation results on benchmark datasets.

In the second stage, We explore whether adapting emotion anchors and emotion semantics are necessary. Similar to classifier re-training (Kang et al., 2019; Nam et al., 2023), we randomly initialize emotion anchors that lie far from the data distribution after learning the utterance representations. Training from scratch is a cold start and cannot reach the optimal position. This result in Line 4 verifies the importance of inheriting emotion anchors and the result shows that the trained emotion anchors express a more powerful ability of recognition. When we remove the anchor adaptation, performance will degrade significantly, indicating the improper positions of emotion anchors weaken the classification performance and verifying the importance of stage two. Line 5 in Table 4 confirms our assumption. In summary, the components of our method contribute to the results substantially.

5.3 Analysis of Contrastive Learning

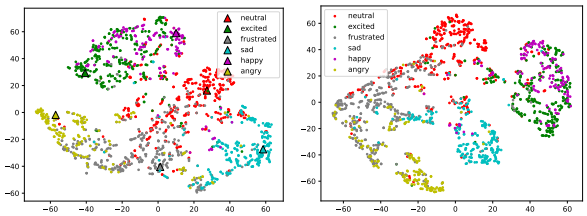
Equipped with emotion anchors, utterance representations move toward their own semantic position, whose cohesion ability is absent in vanilla-supervised contrastive learning. EACL achieves more separability for utterances that have similar emotions. In Figure 4(a), we can observe that the emotion anchors are distributed uniformly, *excited* and *happy*, *frustrated* and *angry* lie far. Meanwhile, utterance representations with other emotions also exhibit significant dispersion. Figure 4(b) shows that similar emotion representations obtained by SupCon lie closer than EACL and thus are harder to distinguish. The slight visual difference is due to the contrastive learning that had been employed in SimCSE, which helps learn distinct representations. The ablation study shows the superior performance of EACL over removing emotion anchors. Quantitative comparison recorded in Appendix D indicates that EACL alleviates emotion similarity to a large extent.

5.4 Performance on Different Language Models

To evaluate the versatility of our learning framework, we conducted experiments using different pretrained language models. Specifically, we examined the performance of our framework on two additional popular language models, namely Deberta-Large (He et al., 2020b) and Promcse-Roberta-Large (Jiang et al., 2022). The results, presented in Table 5, demonstrate that all the pretrained models deliver competitive performance. This observation serves as evidence for the robustness and effectiveness of our framework across various pre-trained language models. It further emphasizes the generalizability of our approach in conversational emotion recognition tasks. We report fine-grained performance in Appendix E.

Dataset	IEMOCAP	MELD	EmoryNLP
SimCSE-Roberta-Large	70.41	67.12	40.24
Deberta-Large	69.09	**67.80**	**41.09**
Promcse-Roberta-Large	**70.45**	67.38	40.93

Table 5: Performance under different language models.



(a) EACL (b) SupCon

Figure 4: The t-SNE visualization of representations on the IEMOCAP test dataset. Triangles represent the emotion anchors. Figures (a) and (b) depict the representation distribution of EACL and SupCon respectively.

6 Conclusion

This paper introduces a novel framework for conversational emotion recognition called emotion-anchored contrastive learning. The proposed EACL leverages emotion representations as anchors to enhance the learning process of distinctive utterance representations. Building upon this foundation, we further adapt the emotion anchors through fine-tuning, bringing them the optimal positions and more suitable for classification purposes. Through extensive experiments and evaluations on three popular benchmark datasets, our approach achieves a new state-of-the-art performance. Ablation studies and evaluations confirm that the proposed EACL framework significantly benefits dialogue modeling and enhances the learning of utterance representations for more accurate emotion recognition.

Limitations

It is important to note that our current method has limitations in tracking distant dialogue history. This constraint arises from the input length restriction of the language model we employ. However, we acknowledge the significance of addressing long-range dialogue modeling and consider it a promising avenue for future research.

Additionally, our method focuses solely on textual inputs and does not incorporate multi-modal settings. We recognize that complementing emotion recognition with facial expressions and tone can provide valuable information. Considering multi-modal inputs is an interesting direction for future enhancements, as it has the potential to improve the overall performance and richness of our emotion recognition framework.

Ethics Statement

The experiments conducted in this paper adopt open-source data for only research purposes. In this work, we try to facilitate machines with the ability to understand better human emotions which is beneficial for dialogue systems or robots. However, it is far from exceeding the understanding of humanity.

References

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.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

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*.

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*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020a. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020b. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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

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*.

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.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Improved universal sentence embeddings with prompt-based contrastive learning and energy-based learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3021–3035.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, and Ashutosh Modi. 2022. Cogmen: Contextualized gnn based multimodal emotion recognition. *arXiv preprint arXiv:2205.02455*.

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. 2021. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. 2022b. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928.

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

Yuchen Liu, Jinming Zhao, Jingwen Hu, Ruichen Li, and Qin Jin. 2022. Dialogueein: Emotion interaction network for dialogue affective analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 684–693.

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.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Giung Nam, Sunguk Jang, and Juho Lee. 2023. Decoupled training for long-tailed classification with stochastic representations. *arXiv preprint arXiv:2304.09426*.

Donovan Ong, Jian Su, Bin Chen, Anh Tuan Luu, Ashok Narendranath, Yue Li, Shuqi Sun, Yingzhan Lin, and Haifeng Wang. 2022. Is discourse role important for emotion recognition in conversation? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11121–11129.

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*.

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. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*.

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.

Sayyed M Zahiri and Jinho D Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint arXiv:1708.04299*.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023. Dualgats: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408.

Zhenyu Zhang, Yuming Zhao, Meng Chen, and Xiaodong He. 2022. Label anchored contrastive learning for language understanding. *arXiv preprint arXiv:2205.10227*.

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*.

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

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. 2022. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917.

Appendix

A Emotion Similarity Anlaysis

To better understand our motivation, we exhibit the emotion similarity in Figure 5. We split the emotions into 3 groups which are composed of positive emotions, negative emotions, and neutral, where positive emotions include *excited* and *happy*, negative emotions contain *frustrated*, *sad*, *angry*, and neutral. It is observed that *excited* and *happy* have a cosine similarity of 0.77, and for *frustrated* and *angry*, they have 0.84 cosine similarity. The similarity of the positive emotions group is higher than that of the negative emotions group. For *neutral*, it is almost equally similar to other emotions.

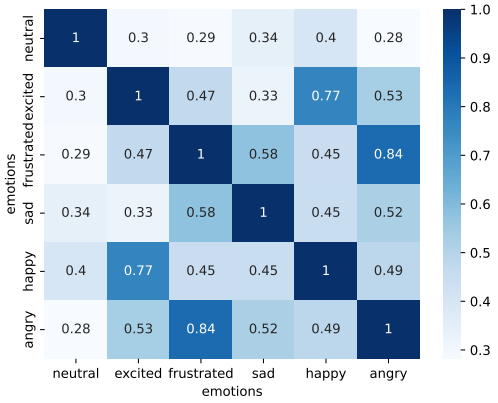


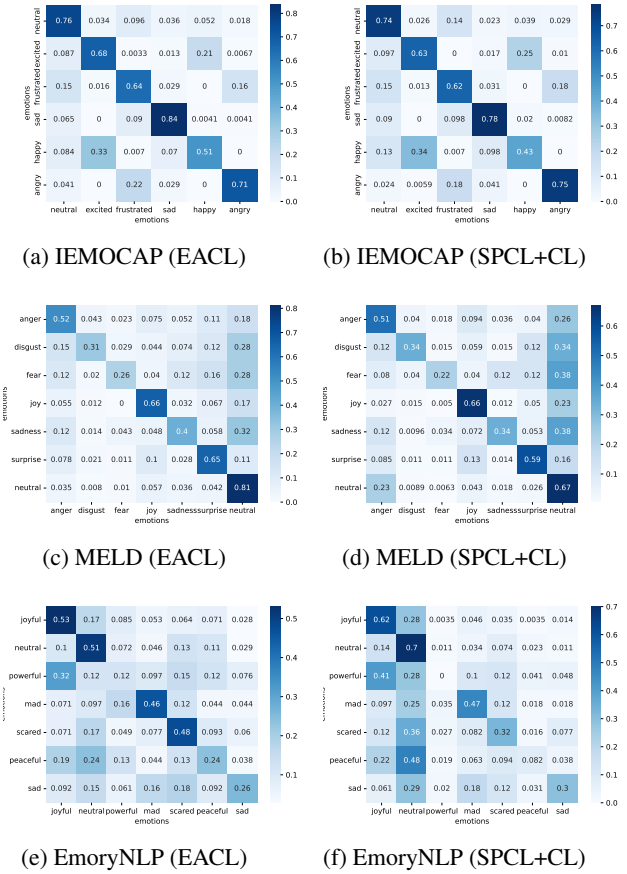
Figure 5: Cosine similarity between emotion word representations extracted from Roberta-Large-SimCSE.

B Experimental Setup

EACL loads the initial parameter by SimCSE-Roberta-Large (Gao et al., 2021) which is identical to the setting of SPCL. All the hyperparameters are reported in Table 6. We exploit grid-search for λ_1 in {0, 0.1, 0.3, 0.5, 0.7, 0.9}, λ_2 in {0, 0.01, 0.1, 1.0} and τ in { 0.05, 0.07, 0.1, 0.15, 0.2}.

Hyperparameters	IEMOCAP	MELD	EmoryNLP
λ_1	0.9	0.1	0.9
λ_2	0.01	0.1	0.01
Temperature τ	0.1	0.1	0.15
Epochs	8	8	8
Maximum length	256	256	256
Learning rate	1e-5	1e-5	1e-5
Dropout	0.1	0.1	0.1

Table 6: Hyperparameters of EACL on three benchmark datasets.



(a) IEMOCAP (EACL) (b) IEMOCAP (SPCL+CL)

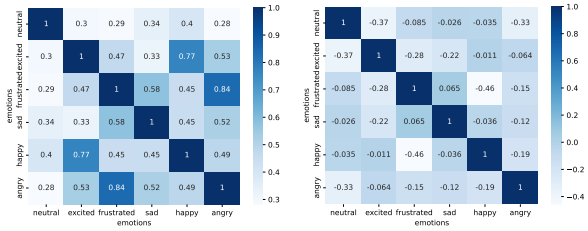
(c) MELD (EACL) (d) MELD (SPCL+CL)

(e) EmoryNLP (EACL) (f) EmoryNLP (SPCL+CL)

Figure 6: The normalized confusion matrix of three benchmark datasets, each row is the true classes and column is predictions. The Coordinate i, j means the percentage of emotion i predicted to be emotion j.

C Detailed Performance Analysis

In Figure 6, we provide the normalized confusion matrices for our EACL and SPCL+CL models across various datasets. These matrices serve as crucial tools for assessing the models' performance. Notably, when we examine the diagonal elements of these matrices, it becomes evident that EACL consistently outperforms the state-of-the-art method SPCL+CL in terms of true positives for most fine-grained emotion categories. This suggests that EACL excels at learning features that are more distinguishable. Particularly noteworthy is the performance of EACL in comparison to SPCL+CL when considering specific emotion pairs, such as *excited* and *happy*, as well as *frustrated* and *angry* on the IEMOCAP dataset. In these cases, EACL demonstrates superior performance. This underscores the effectiveness of the EACL framework in effectively addressing the challenge of misclassification, especially when deal-



(a) Before training (b) After training

Figure 7: The cosine similarity of pair-wise emotions. Figure (a) depicts cosine similarity between emotion anchors extracted from Roberta-Large-SimCSE. and (b) depicts that similarity after training with EACL.

ing with emotions that share similar characteristics. When we focus on the MELD and EmoryNLP datasets, we observe that EACL significantly reduces misclassifications between *neutral* emotions and other emotional states. This highlights EACL's capability to effectively mitigate misclassification issues not only for similar emotions but for all emotion categories.

D Emotion Similarity Comparison

In this section, we conducted a comparison of the similarity between pairs of emotions generated by Roberta-Large-SimCSE in Figure 7a and after training with EACL in Figure 7b. Figure 7 illustrates our findings, which reveal a significant decrease in similarity for emotions that are considered similar. For instance, the cosine similarity between *excited* and *happy* drops sharply from 0.77 to 0.08, while for *frustrated* and *angry*, it decreases from 0.84 to -0.3. Meanwhile, naturally dissimilar emotions are now positioned further apart. For instance, the similarity between *neutral* and other emotions also experiences a notable decline. These observations suggest that EACL effectively increases the separation between similar emotions, thereby enhancing the model's ability to distinguish between them.

E Fine-Grained Performance on Different Models

In this section, we report the fine-grained performance when using Deberta-Large (He et al., 2020b) and Promcse-Roberta-Large (Jiang et al., 2022) in Table 7. The results indicate that our learning framework is robust to different language models. Similar to the result under Roberta-SimCSE, these models can also effectively separate similar emotions and achieve state-of-the-art performance on the benchmark datasets.

(a) IEMOCAP

Model	Exc	Fru	Sad	Neu	Ang	Hap	Avg	W-f1
Deberta	68.55	69.74	80.17	70.18	65.41	50.96	67.50	69.09
PromCSE	68.64	67.19	80.81	74.66	69.11	53.41	68.97	70.45
SPCL+CL	66.72	63.96	80.03	72.29	64.82	43.96	65.30	67.19

(b) MELD

Methods	Fear	Neu	Ang	Sad	Dis	Surp	Joy	Avg	W-f1
Deberta	34.0	80.43	55.28	44.44	37.59	60.85	65.34	53.99	67.8
PromCSE	23.59	81.0	54.96	43.35	30.53	59.51	65.12	51.15	67.38
SPCL+CL	26.59	77.92	54.40	43.53	30.94	59.26	60.34	50.43	65.74

(c) EmoryNLP

Methods	Joy	Sad	Pow	Mad	Neu	Pea	Sca	Avg	W-f1
Deberta	54.04	28.74	21.54	41.73	51.75	18.12	42.52	36.92	41.09
PromCSE	54.42	28.33	14.21	43.35	51.64	23.42	41.30	36.68	40.93
SPCL+CL	53.52	31.61	10.28	44.21	51.40	16.83	39.51	35.34	39.52

Table 7: Fine-grained performance record on different language models for all emotions on three benchmark datasets, the F1-score is used for each class.