Emotion Transfer with Enhanced Prototype for Unseen Emotion Recognition in Conversation

Kun Peng[1,2], Cong Cao[1*], Hao Peng[3], Guanlin Wu[4], Zhifeng Hao[5], Lei Jiang[1], Yanbing Liu[1,2], Philip S. Yu[6],

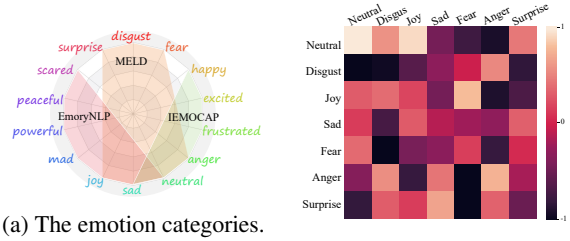
[1]Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
[3]Beihang University, China [4]National University of Defense Technology, China
[5]Shantou University, China [6]University of Illinois at Chicago, USA
{pengkun, caocong}@iie.ac.cn

Abstract

Current Emotion Recognition in Conversation (ERC) research follows a closed-domain assumption. However, there is no clear consensus on emotion classification in psychology, which presents a challenge for models when it comes to recognizing previously unseen emotions in real-world applications. To bridge this gap, we introduce the Unseen Emotion Recognition in Conversation (UERC) task for the first time and propose **ProEmoTrans**, a solid prototype-based emotion transfer framework. This prototype-based approach shows promise but still faces key challenges: First, implicit expressions complicate emotion definition, which we address by proposing an LLM-enhanced description approach. Second, utterance encoding in long conversations is difficult, which we tackle with a proposed parameter-free mechanism for efficient encoding and overfitting prevention. Finally, the Markovian flow nature of emotions is hard to transfer, which we address with an improved Attention Viterbi Decoding (AVD) method to transfer seen emotion transitions to unseen emotions. Extensive experiments on three datasets show that our method serves as a strong baseline for preliminary exploration in this new area.

1 Introduction

Emotion Recognition in Conversation (ERC) aims to predict the emotional state of each utterance in multi-turn conversations, holding significant research value in areas such as Conversational Sentiment Analysis (Li et al., 2023) and Empathetic Responses (Peng et al., 2022). However, in the field of psychology, existing research works (Ekman, 1999; Plutchik and Kellerman, 2013; Cowen and Keltner, 2017) feature a variety of emotion classification theories, yet they have not reached a clear consensus[1]. Due to the complex definitions and the various classification theories, in real-world applications, such as open-domain dialogue systems, it is likely to occur new emotions that are unseen in the training stage. As shown in Figure 1 (a), the emotion labels across three widely used datasets (Busso et al., 2008; Zahiri and Choi, 2018; Poria et al., 2019) exhibit significant non-overlapping portions. This makes it challenging to directly apply models trained on a single dataset to other datasets. For instance, a model trained on the MELD dataset may struggle to recognize the emotion *powerful* in the EmoryNLP dataset.



(a) The emotion categories.

(b) Transition scores.

Figure 1: (a) shows that the emotion categories in three foundational datasets vary significantly in the emotion labels. (b) shows the transition scores learned on the MELD dataset.

To bridge this gap, we introduce the Unseen Emotion Recognition in Conversation (UERC) task for the first time, which aims to predict unseen emotions by leveraging prior knowledge from seen emotions in training data. To address this task, we attempt the prototype-based approaches (Chen and Li, 2021; Zhao et al., 2023; Li et al., 2024) to learn a prototype vector for each emotion, helping the model capture the distinct meaning of emotions. However, three key challenges hinder progress. **Challenge 1: Implicit emotion expression**. Existing methods primarily rely on the provided label descriptions to enhance prototype semantics. How-

*Corresponding author.

[1]For instance, Plutchik and Kellerman (2013) categorizes emotions into 32 types, while Cowen and Keltner (2017) categorizes emotions into 27 types, including unusual emotions like *nostalgia* and *sexual desire*.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 597–609
November 4-9, 2025 ©2025 Association for Computational Linguistics

ever, the UERC task lacks emotion descriptions, and, even more critically, many complex emotions are hard to define clearly, and relying solely on descriptive information is insufficient to obtain robust and faithful prototypes. **Challenge 2: Hard utterance encoding**. Due to the extensive length of conversation texts, existing ERC methods (Majumder et al., 2019; Hu et al., 2021; Zhang et al., 2023; Yang et al., 2024) typically follow two steps: encoding utterance representations first, then modeling inter-utterance features with additional relation-learning modules. However, our preliminary experiments indicate that these additional modules can lead to overfitting the training data, compromising the model's ability to generalize to unseen emotions. Conversely, removing these modules results in losing valuable inter-utterance relations, creating a dilemma. **Challenge 3: Unadapted emotion transition**. It's found that emotions exhibit a Markov property (Song et al., 2022b), whereby the current utterance's emotion is influenced by preceding ones. As illustrated in Figure 1 (b), when the current emotion is *Disgust*, the transfer score for *Anger* in the subsequent utterance is notably highest, aligning with intuitive expectations. While the Markov property can effectively aid emotion prediction, the transfer score matrix for unseen emotions cannot be pre-learned.

To address these challenges, we propose a solid prototype-based emotion transfer framework called **ProEmoTrans**. Specifically, to address the **implicit emotion expression** challenge, we first employ a dictionary to obtain all the emotion descriptions. We then leverage the in-context learning capabilities of large language models (LLMs) to generate utterances that implicitly express these emotions, thereby enhancing the model's comprehension of complex emotions. To address the **hard utterance encoding** challenge, we refrain from using additional relation-learning modules to prevent the model from overfitting to seen emotions. Instead, we propose a Gaussian Self-Attention mechanism to capture inter-utterance relations. This parameter-free mechanism obtains utterance embeddings by using linear combinations of contextual representations, effectively leveraging relation information among utterances at varying distances. To leverage the **emotion transition**, we propose an improved Attention Viterbi Decoding (AVD) algorithm within the Conditional Random Field (CRF) framework, enabling the capture of transition probabilities for seen emotions between all adjacent utterances. Subsequently, we extend the transition probabilities of seen emotions to unseen emotions by utilizing prototype similarity. Our contributions can be summarized as follows:

1) We propose the UERC task for the first time and introduce a novel model called ProEmoTrans[2]. Extensive experiments on three datasets demonstrate that this method serves as a solid baseline.

2) We leverage the prior knowledge of LLMs to generate implicit contexts that enhance complex emotion prototypes.

3) We introduce a Gaussian self-attention mechanism that effectively utilizes inter-utterance relations while avoiding overfitting to seen emotions.

4) We improve the Viterbi decoding algorithm to extend the transition probabilities of seen emotions to unseen emotions.

2 Related Work

2.1 Emotion Recognition in Conversation

ERC in a text-modality setting is an active research topic. Early RNN-based (Jiao et al., 2019; Majumder et al., 2019; Hu et al., 2021) and GCN-based (Ghosal et al., 2019; Shen et al., 2021; Zhang et al., 2023) methods tried to model the temporal features or conversational structures. Some other studies (Ghosal et al., 2020; Ong et al., 2022) have also attempted to integrate more common-sense knowledge. The latest contrastive-based methods (Hu et al., 2023; Yang et al., 2023; Yu et al., 2024) focus on using contrastive learning to distinguish semantically similar emotions. While these additional modules can effectively help the model fit the distributions of seen emotions, in the UERC setting, they can impair the model's ability to generalize to unseen emotions.

2.2 Zero-shot Learning in ERC

Zero-shot Learning (ZSL) aims to train a model on one label set and then apply it to another set of previously unseen labels. Currently, research on ZSL in the ERC field is quite limited. A work that is closely related to ours is CTPT (Xu et al., 2023), which focuses on cross-task few-shot settings, while we are the first to explore the model's ability in zero-shot predicting for unseen emotions. In the zero-shot setting, CTPT primarily improves the recognition of similar emotions across tasks but performs poorly in recognizing unseen emotions.

[2]Available at https://github.com/KunPunCN/ProEmoTrans/

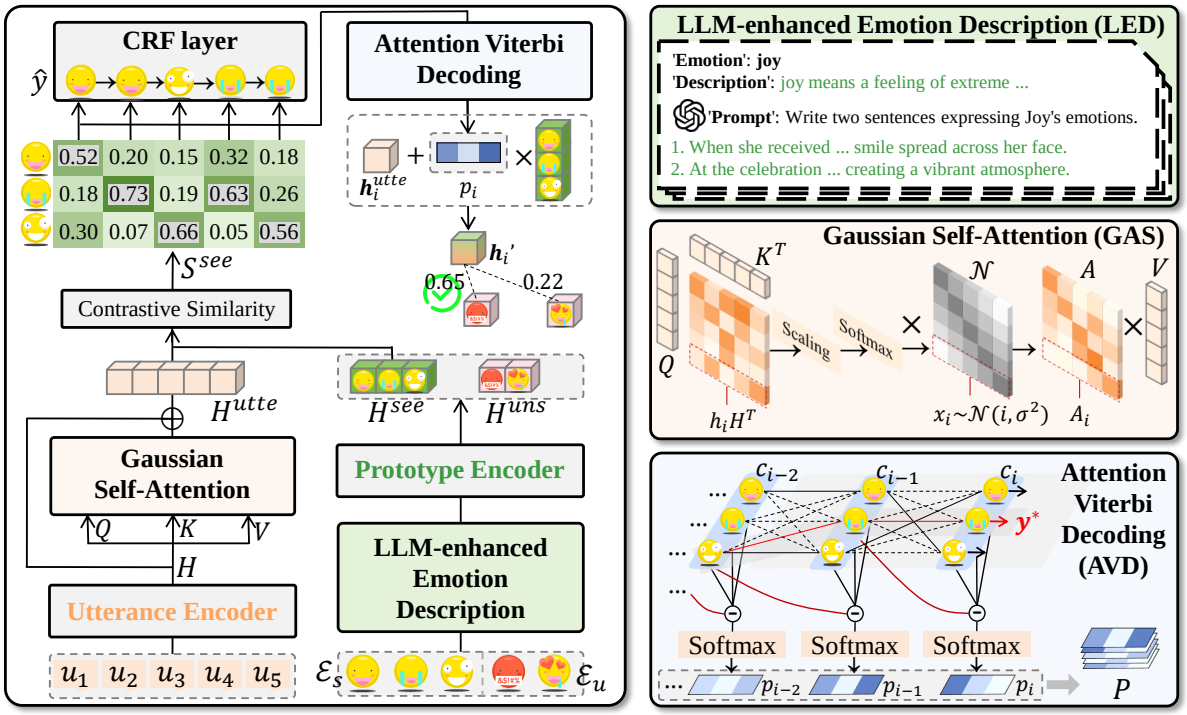


Figure 2: The architecture of our proposed EmoTrans.

Prototype alignment (Chen and Li, 2021; Zhao et al., 2023; Li et al., 2024) is a powerful method in ZSL. It first encodes sentence and label information into a hidden vector space, then aligns sentence embeddings with label prototype embeddings using semantic matching. Through this process, the model acquires the ability to generalize label knowledge. During the inference phase, the model encodes the unseen label information and makes predictions through nearest neighbor search. In other prototype-based zero-shot NLP research fields, such as zero-shot relation extraction, Zhao et al. (2023) proposes a fine-grained semantic matching method to reduce the negative impact of irrelevant features. Li et al. (2024) enhances label prototypes by introducing more side descriptions.

3 Methodology

3.1 Task Definition

Given a conversation $\mathcal{U} = \{(u_1, t_1), (u_2, t_2), ..., (u_N, t_N)\}$, where each utterance u_i has only one speaker t_i, N is the total number of utterances. Different utterances may belong to the same speaker, so it is possible to have $t_i = t_j (i \neq j)$. The training dataset $\mathcal{D}_s$ has a set of seen emotions $\mathcal{E}_s$, and the test dataset $\mathcal{D}_u$ has a set of unseen emotions $\mathcal{E}_u$. There is no overlap between $\mathcal{E}_s$ and $\mathcal{E}_u$. The objective of the UERC task is to learn from $\mathcal{D}_s$ and transfer the model to predict the unseen emotion label $e_i^{uns} \in \mathcal{E}_u$ of each utterance u_i.

3.2 Framework of ProEmoTrans

The overall architecture of our proposed ProEmoTrans is illustrated in Figure 2.

3.2.1 Emotion Prototype Encoding

Given a seen emotion word $e_i^{see} \in \mathcal{E}_s$, we can find its corresponding description[3] X_i^{desc} in the Wiktionary[4]. However, unlike direct descriptions, emotions in conversation are often expressed implicitly. This gap makes the prototype learned from emotional descriptions lack sufficient generalization, especially for more complex emotions (e.g., powerful).

To improve the quality of emotion prototypes, we propose the **LLM-enhanced Emotion Description (LED)** method. We first design a prompt template *Write two sentences expressing* [*MASK*]*'s emotions.* Afterward, by filling in the [*MASK*] position with the emotion word e_i^{see} and leveraging the LLM's prompt generation capabilities, we generate sentences X_i^{llm} that implicitly express that emotion. The enhanced description X_i^{see} is defined

[3]All the descriptions are listed in Appendix C.
[4]https://en.m.wiktionary.org

as the concatenation of e_i^{see}, X_i^{desc} and X_i^{llm}:

$$X_i^{see} = \{[CLS], e_i^{see}, X_i^{desc}, X_i^{llm}, [SEP]\}. \quad (1)$$

We feed it into the prototype encoder to obtain the final emotion prototype $\boldsymbol{h}_i^{see}$:

$$\boldsymbol{h}_i^{see} = Encoder_E(X_i^{see})[0], \quad (2)$$

where $\boldsymbol{h}_i^{see} \in \mathbb{R}^d$ is the first token (i.e., [*CLS*]) of the last hidden layer. Through the above process, we can encode the prototypes of each emotion word in the seen emotion set $\mathcal{E}_s$ and obtain $\boldsymbol{H}^{see} = (\boldsymbol{h}_1^{see}, \boldsymbol{h}_2^{see}, ..., \boldsymbol{h}_n^{see})$. Similarly, for the unseen emotion set $\mathcal{E}_u$, we have $\boldsymbol{H}^{uns} = (\boldsymbol{h}_1^{uns}, \boldsymbol{h}_2^{uns}, ..., \boldsymbol{h}_m^{uns})$, where n and m are the numbers of emotions in $\mathcal{E}_s$ and $\mathcal{E}_u$, respectively.

3.2.2 Utterance Encoding

Following previous works (Hu et al., 2021; Shen et al., 2021; Zhang et al., 2023), due to the conversation text being too lengthy, we use an utterance encoder to obtain the utterance representation $\boldsymbol{h}_i$:

$$\boldsymbol{h}_i = Encoder_U(u_i)[0]. \quad (3)$$

The representation of all utterances is denoted as $\boldsymbol{H} \in \mathbb{R}^{N \times d}$, where N is the number of utterances in $\mathcal{U}$. After that, we propose a non-parametric **Gaussian Self-Attention (GSA)** mechanism that effectively learns the inter-utterance relationships and alleviates overfitting to seen emotions.

Given the token $\boldsymbol{h}_i \in \boldsymbol{H}$, the Gaussian attention score $\boldsymbol{A}_i \in \mathbb{R}^N$ that attends to $\boldsymbol{H}$ is defined as:

$$\boldsymbol{A}_i = Softmax(\frac{\boldsymbol{h}_i \boldsymbol{H}^T}{d})\boldsymbol{\mathcal{N}}_i, \quad (4)$$

where $\boldsymbol{\mathcal{N}}_i \in \mathbb{R}^N$ are discrete values that follow the Gaussian distribution $\mathcal{N}(i, \sigma^2)$, and the variance σ is a hyperparameter. Using the Gaussian attention score, we aggregate highly relevant information from the entire conversation while reducing the impact of distant tokens. This inter-utterance relationship aggregation follows a non-parametric linear operation:

$$\boldsymbol{h}_i^{utte} = \boldsymbol{h}_i + \boldsymbol{A}_i \boldsymbol{H}, \quad (5)$$

where $\boldsymbol{h}_i^{utte} \in \mathbb{R}^d$ is the updated utterance representation. The final representation of all utterances is denoted as $\boldsymbol{H}^{utte} \in \mathbb{R}^{N \times d}$.

The GSA mechanism has two key properties: First, parameter-free. Previous supervised methods used parameterized modules (such as LSTM and GCN) to learn inter-utterance relationships. However, in unsupervised scenarios, parameterized modules led to overfitting on the training set, hindering generalization on unseen datasets (Appendix B.1). Second, distance-aware learning of inter-utterance relationships. Directly sampling discrete values from a one-dimensional Gaussian distribution based on the distance between utterances, with closer utterances receiving more attention.

3.2.3 Contrastive Similarity and Training

In the above sections, we obtained emotion prototypes and utterance representations. In this section, through nearest neighbor search, we can align utterances with their corresponding emotion labels. Inspired by infoNCE (Oord et al., 2018), we define a contrastive similarity to pull the utterance embeddings closer to their corresponding prototype embeddings while pushing apart the inconsistent ones. This similarity $\boldsymbol{S}^{see} \in \mathbb{R}^{N \times n}$ is defined as:

$$\boldsymbol{S}^{see} = Sim(\boldsymbol{H}^{utte}, \boldsymbol{H}^{see}), \quad (6)$$

$$s_{ij}^{see} = \frac{e^{cos(\boldsymbol{h}_i^{utte}, \boldsymbol{h}_j^{see})/\tau}}{\sum_{j=1}^{n} e^{cos(\boldsymbol{h}_i^{utte}, \boldsymbol{h}_j^{see})/\tau}}, \quad (7)$$

where Eq. (7) is the details of Eq. (6). $cos(\cdot)$ is a cosine similarity function and τ is a temperature hyperparameter. s_{ij}^{see} represents the probability of the i-th utterance expressing the j-th seen emotion.

Due to the transition dependencies between emotions, independent predictions are insufficient. Therefore, we subsequently feed $\boldsymbol{S}^{see}$ into a Conditional Random Field (CRF) (Lafferty et al., 2001). For a sequence of predictions: $\boldsymbol{y} = (y_1, y_2, ..., y_N)$, its CRF score can be defined as:

$$\mathcal{C}(\boldsymbol{y}) = \sum_{k=0}^{N} \boldsymbol{M}_{y_k, y_{k+1}} + \sum_{k=1}^{N} \boldsymbol{S}_{k, y_k}^{see}, \quad (8)$$

where $\boldsymbol{M} \in \mathbb{R}^{(n+2) \times n}$ is the transition matrix[5] of the CRF layer. y_0 and y_{N+1} are the additional *start* and *end* tags. The probability of the sequence $\boldsymbol{y}$ is a softmax over the scores of all possible sequences:

$$p(\boldsymbol{y}) = \frac{e^{\mathcal{C}(\boldsymbol{y})}}{\sum_{\tilde{y} \in \boldsymbol{Y}_{\mathcal{U}}} e^{\mathcal{C}(\tilde{\boldsymbol{y}})}}, \quad (9)$$

where $\boldsymbol{Y}_{\mathcal{U}}$ represents all possible predicted sequences. Our training goal is to minimize the loss: $\mathcal{L} = -log(p(\widehat{\boldsymbol{y}}))$, where $\widehat{\boldsymbol{y}}$ represents the true sequences.

[5] $n + 2$ is because there are start and end transitions here.

3.2.4 Inference

The original Viterbi decoding is limited to the seen emotions, and the valuable emotion transition dependencies learned by the CRF layer cannot be adapted to unseen emotions. To address this gap, we propose the **Attention Viterbi Decoding (AVD)** algorithm. We define the score of the i-th utterance expressing the j-th seen emotion as:

$$c_{ij} = \max_{\tilde{\boldsymbol{y}} \in \boldsymbol{Y}_{\boldsymbol{\mathcal{U}}_{[1:i]}}, \tilde{y}_k = j} \mathcal{C}(\tilde{\boldsymbol{y}}), \tag{10}$$

where $c_{0j} = \boldsymbol{M}_{0,j}$. $\boldsymbol{Y}_{\boldsymbol{\mathcal{U}}_{[1:i]}}$ represents all possible tag sequences from u_1 to u_i. The score c_{ij} represents the maximum CRF score of all possible sequences ending with $\tilde{y}_i = e_j^{see}$. Based on Eq. (8), we can derive that:

$$c_{ij} = \max_{1<=k<=n} (c_{(i-1)k} + \boldsymbol{M}_{k,j} + \boldsymbol{S}_{k,j}^{see}). \tag{11}$$

The time complexity of calculating a single c_{ij} is $\mathcal{O}(n)$. The overall time complexity for traversing all c_{ij} is $\mathcal{O}(Nn^2)$. During the traversal, we also record the path $\boldsymbol{y}^* = (y_1^*, ..., y_N^*)$ with the maximum CRF score, such that $c_{Ny_N^*} > c_{Nj}, \forall j \neq y_N^*$.

The final output of the AVD algorithm is a probability matrix $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{N \times n}$, where each $p_{ij} \in \boldsymbol{\mathcal{P}}$ is defined as follows:

$$p_{ij} = \frac{c_{ij} - c_{(i-1)y_{i-1}^*}}{\sum_{k=1}^{n}(c_{ik} - c_{(i-1)y_{i-1}^*})}, \tag{12}$$

where p_{ij} denotes the probability of the k-th utterance expressing the j-th seen emotion. Then, we can enhance the original utterance representation using the seen emotion prototypes:

$$\boldsymbol{h}_i' = \boldsymbol{h}_i^{utte} + \sum\nolimits_{j=1}^{n} p_{ij} \boldsymbol{h}_j^{see}, \tag{13}$$

where $\boldsymbol{h}_i'$ incorporates the seen emotion prototypes after considering similarity (from $\boldsymbol{S}^{see}$) and emotional dependencies (from $\boldsymbol{M}$).

For a given u_i, the predicted unseen emotion label is obtained through nearest neighbor search:

$$y_i^{uns} = \underset{1<=j<=m}{\arg\max}\ cos(\boldsymbol{h}_i', \boldsymbol{h}_j^{uns}). \tag{14}$$

4 Experiments Settings

4.1 Datasets

We evaluate our ProEmoTrans on three widely used datasets: **IEMOCAP** (Busso et al., 2008) is based on two actors performing a script. **EmoryNLP** (Zahiri and Choi, 2018) and **MELD** (Poria et al., 2019) contain scripts collected from the *Friends* TV series. We only use the text modality of these datasets and follow previous work in splitting the IEMOCAP dataset into training and validation sets. The dataset statistics are drawn in Table 1. We denote these datasets as $\boldsymbol{\mathcal{I}}$, $\boldsymbol{\mathcal{E}}$, and $\boldsymbol{\mathcal{M}}$, respectively. We iterate through different source datasets to train the model and use the validation and test sets of the other two datasets as the target unseen emotion datasets. For instance, to evaluate the model trained on $\boldsymbol{\mathcal{I}}$ for its performance on $\boldsymbol{\mathcal{M}}$ test set, we select $\boldsymbol{\mathcal{E}}$ as the validation set. The statistics of the unseen emotions under different source and target settings are shown in Appendix A.1.

Dataset	# Conversations			# Uterrances			# Emos.
	train	dev	test	train	dev	test	
IEMOCAP ($\boldsymbol{\mathcal{I}}$)	100	20	31	4810	1000	1623	6
EmoryNLP ($\boldsymbol{\mathcal{E}}$)	659	89	79	7551	954	984	7
MELD ($\boldsymbol{\mathcal{M}}$)	1038	114	280	9989	1109	2610	7

Table 1: Statistics of experimental datasets.

4.2 Implementation Details

We utilize Bert-base-uncased (Vaswani et al., 2017) as both the utterance and prototype encoder. We use ChatGPT-3.5 to generate enhanced emotion descriptions. In each training batch, we input the emotion descriptions and the utterances into the encoders simultaneously. We use the AdamW optimizer (Kingma and Ba, 2015) with a batch size of 4 and a learning rate of $2e-5$. The model is trained for 10 epochs with 100 warm-up steps. All experiments are conducted with an NVIDIA RTX 8000. The variance σ of the Gaussian distribution is set to 0.5, and the temperature τ in Eq. (7) is set to 0.02. We use the weighted-averaged F1 score as the evaluation metric, considering only unseen emotions. In each epoch, we evaluate the training model on the validation set and save the best one to test. All results are averaged across five runs with different random seeds.

4.3 Baselines

Due to limited research, we choose the following four types of baselines and make necessary modifications to their original architectures to achieve zero-shot prediction capability:

Feature-based models: **DialogueGCN** (Ghosal et al., 2019), **DialogueCRN** (Hu et al., 2021), and **DualGAT** (Zhang et al., 2023) design special GNN/RNN-based modules to extract better utter-

ance features and use a label-wise classification head to predict the label of each utterance. They use cross-entropy loss computed from the prediction logits and the labels. To enable zero-shot prediction capability, we replace the classification head with a prototype encoder, which enables the model to learn prototype vectors. Then we substitute the original cross-entropy loss with a contrastive loss based on prototype similarity (similar to Eq. 7).

Contrastive-based models: **SACL-LSTM** (Hu et al., 2023), **SCCL** (Yang et al., 2023), and **EACL** (Yu et al., 2024) focus on distinguishing semantically similar emotions using contrastive learning. Since these models natively use representation similarity for prediction, no modifications are needed.

Few-shot model: **CPTC** (Xu et al., 2023) leverages sharable cross-task knowledge from the source task to improve few-shot performance. By removing task-specific prompts, it can also perform zero-shot prediction. Unlike in their original work, we evaluate the model only on unseen emotions. To ensure fairness, all of these comparison models use BERT-base-uncased as their backbone.

LLMs: **Llama-3.1-8b** (Grattafiori et al., 2024), **Qwen-2.5-7b** (Yang et al., 2025), **GPT-4o** (Bubeck et al., 2023), and **DeepSeek-V3** (Liu et al., 2024) are used for zero-shot prediction. We design a unified prompt template:

Given a conversation: <INPUT>. Please analyze the emotion of each utterance in the conversation. The emotions are included in <LABEL SET>.

Models	$\mathcal{E} \rightarrow \mathcal{I}$			$\mathcal{M} \rightarrow \mathcal{I}$		
	wP.	*wR.*	*wF1.*	*wP.*	*wR.*	*wF1.*
DialogueGCN	6.83	4.55	5.84	5.61	3.48	4.71
DialogueCRN	7.48	6.48	6.51	7.08	5.16	6.44
DualGAT	9.08	5.58	7.49	7.11	5.14	6.12
CPTC	17.10	10.82	14.58	13.93	10.14	11.13
SACL-LSTM	33.05	24.50	20.55	36.39	19.25	19.90
SCCL	33.07	24.56	21.21	36.11	18.46	19.59
EACL	36.10	27.42	23.89	37.00	19.53	20.79
Llama-3.1-8b	38.17	20.30	23.63	44.33	30.61	24.79
Qwen-2.5-7b	39.34	21.58	24.20	45.15	31.37	25.66
GPT-4o	41.27	27.42	24.88	45.39	31.73	26.10
DeepSeek-V3	42.55	27.69	25.69	46.38	32.05	26.26
ProEmoTrans (Ours)	**47.80**	**32.95**	**37.27**	**47.11**	**30.90**	**32.36**
Models	$\mathcal{I} \rightarrow \mathcal{E}$			$\mathcal{M} \rightarrow \mathcal{E}$		
	wP.	*wR.*	*wF1.*	*wP.*	*wR.*	*wF1.*
DialogueGCN	7.12	2.54	2.94	6.19	1.34	1.74
DialogueCRN	4.32	3.27	3.29	5.52	1.23	2.09
DualGAT	9.53	3.92	4.35	3.71	1.26	1.96
CPTC	7.06	4.19	5.16	3.94	1.37	2.41
SACL-LSTM	15.52	16.49	15.34	14.25	8.85	10.07
SCCL	14.29	15.44	14.71	15.00	9.67	11.31
EACL	17.48	18.01	17.36	16.36	9.74	12.39
Llama-3.1-8b	31.32	22.18	24.10	20.11	16.83	16.42
Qwen-2.5-7b	31.08	22.47	24.05	21.17	16.71	17.09
GPT-4o	31.14	21.61	24.51	20.71	16.32	18.25
DeepSeek-V3	31.01	23.27	24.10	22.91	16.27	18.68
ProEmoTrans (Ours)	**31.36**	**27.67**	**28.34**	**24.98**	**19.07**	**20.73**
Models	$\mathcal{I} \rightarrow \mathcal{M}$			$\mathcal{E} \rightarrow \mathcal{M}$		
	wP.	*wR.*	*wF1.*	*wP.*	*wR.*	*wF1.*
DialogueGCN	5.76	3.12	4.44	5.42	1.99	2.67
DialogueCRN	7.46	4.00	5.13	6.93	3.15	3.95
DualGAT	8.20	4.12	5.07	6.23	2.08	2.94
CPTC	19.51	5.65	8.13	13.69	4.30	6.40
SACL-LSTM	31.60	19.29	25.60	29.55	22.28	25.48
SCCL	31.05	19.39	25.14	28.81	21.02	24.32
EACL	33.32	20.27	26.29	31.58	23.52	26.95
Llama-3.1-8b	31.08	22.47	24.05	32.81	25.36	27.73
Qwen-2.5-7b	30.50	43.80	35.12	34.72	26.96	29.76
GPT-4o	29.90	45.65	35.28	34.85	25.74	29.35
DeepSeek-V3	31.85	44.67	35.15	34.76	27.19	29.76
ProEmoTrans (Ours)	**35.74**	**45.32**	**38.59**	**36.30**	**36.02**	**35.64**

Table 2: The overall performance of all the compared baselines and our ProEmoTrans on benchmark datasets. Here *wP.*, *wR.*, and *wF1.* denote weighted-averaged precision, recall, and F1 score.

5 Results and Analysis

5.1 Main Results

The overall performance on the three datasets is reported in Table 2. We have the following observations: Our ProEmoTrans outperforms all other models by a significant margin. Compared to the best baseline DeepSeek-V3, ProEmoTrans achieved improvements in the weighted-averaged F1 score of 11.58%, 6.1%, 4.24%, 2.05%, 3.44%, and 5.88% across six different dataset settings. This demonstrates that our ProEmoTrans exhibits strong performance. The feature-based methods DialogueGCN, DialogueCRN, and DualGAT perform poorly due to their excessive parameter modules, which make them prone to overfitting on seen emotions. Few-shot model CPTC also shows inefficient recognition of unseen emotions. The contrastive-based methods SACL-LSTM, SCCL, and EACL focus on improving the distinguishability of different emotions. Learning differentiated emotional prototypes helps them perform better on the UERC task than other supervised methods. LLMs outperform other baselines with their rich prior knowledge. To investigate how our model improves performance compared to GPT-4o, we provide a more in-depth discussion in the fine-grained analysis (Section 5.5).

5.2 Ablation Study

We conduct ablation studies to investigate the effectiveness of the key components in our method. The results are shown in Table 3.

-w/o *LED* denotes removing the LED module

Models	$\mathcal{E} \rightarrow \mathcal{I}$	$\mathcal{M} \rightarrow \mathcal{I}$	$\mathcal{I} \rightarrow \mathcal{E}$	$\mathcal{M} \rightarrow \mathcal{E}$	$\mathcal{I} \rightarrow \mathcal{M}$	$\mathcal{E} \rightarrow \mathcal{M}$	Average
Proposed ProEmoTrans	**37.27**	**32.36**	**28.34**	**20.73**	**38.59**	**35.64**	**32.16**
- w/o *LED*	27.68	7.28	24.06	6.31	22.59	19.22	17.86 (14.30$^{\downarrow}$)
- w 1 *Desc.*	30.46	9.90	25.06	8.74	24.68	25.26	20.68 (11.48$^{\downarrow}$)
- w 3 *Desc.*	37.56	33.03	28.39	21.22	37.89	36.82	32.49 (**0.33**$^{\uparrow}$)
- w/o *GSA*	36.89	31.40	27.00	19.37	37.68	34.26	31.10 (1.06$^{\downarrow}$)
- w *SA*	36.27	30.78	26.47	19.82	37.01	33.45	30.63 (1.53$^{\downarrow}$)
- w/o *CRF*	31.22	19.20	18.29	16.59	33.82	32.27	25.23 (6.93$^{\downarrow}$)

Table 3: Ablation and comparison results for key components. Here 1 *Desp.* and 3 *Desp.* denote the number of generated descriptions in LED. *SA* denotes replacing GSA with the original self-attention mechanism.

and directly using dictionary definitions as its description. It is evident that removing the LED results in a significant 14.3% drop in the model's average wF1 score, highlighting the importance of descriptive information in enhancing emotion representation. In the original model, we use two descriptions (*2 Desc.*) to help the model fully capture the emotional semantics. To investigate the impact of the number of generated descriptions, we conduct experiments comparing the model's performance with different numbers of descriptions. As shown in Table 3, with one description (-w 1 *Desc.*), the average wF1 increases by 2.82% compared to *no Desp*. However, it still shows an 11.48% drop compared to the original 2 *Desc.*. With three descriptions (-w 3 *Desc.*), the average wF1 only slightly increases by 0.33%. This indicates that 2 *Desc.* are sufficient for the model to fully capture the semantic meaning.

-w/o *GSA* denotes removing the GSA mechanism and directly using $\boldsymbol{H}$ from Eq. (3) as the final utterance representations. This led to a decrease of 1.06% in the average wF1, demonstrating the positive role of the GSA mechanism in enhancing utterance representations. Since the GSA mechanism benefits from aggregating highly relevant information while reducing the negative impact of distant utterances, we further compare it with using the self-attention mechanism (*SA*) alone. The results show that the performance drops by 1.53%, and it even performs 0.47% worse than when no mechanism was used (-w/o *GSA*). This demonstrates that directly using *SA* for utterance representation learning has a detrimental effect, with the negative impact stemming from distant noise.

-w/o *CRF* denotes removing the CRF layer and the AVD algorithm, and during the inference phase, it directly uses h_i^{see} and h_j^{uns} for nearest neighbor search as specified in Eq. (14). The results show a decrease of 6.93% in average wF1, which demonstrates that the AVD algorithm, by leveraging the emotion transition dependencies learned by the CRF layer, plays a crucial role in enhancing the model's performance.

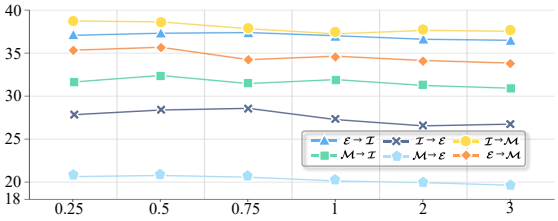


Figure 3: Effects of σ.

Model	Performance	Inference Costs
bert-base-uncased	32.16	6.21 /ms
roberta-base	33.02	6.34 /ms
bert-large-uncased	34.48	9.12 /ms
roberta-large	34.83	9.26 /ms

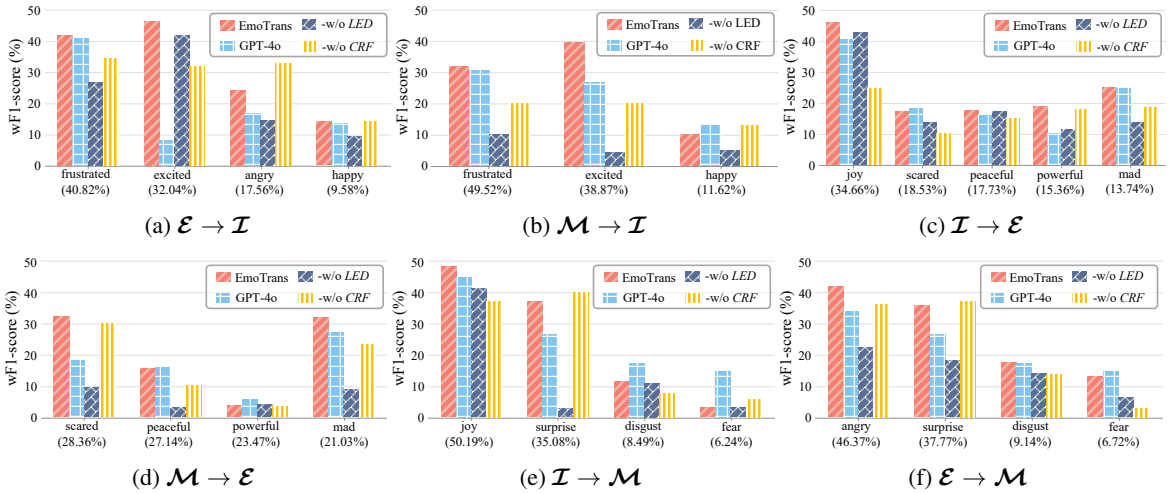
Table 4: Performance (wF1.) and computation cost (/ms) with different language models

5.3 Hyperparameter Sensitivity

The variance σ in the GSA mechanism controls the attention range. To study the impact of σ on performance, we conducted a sensitivity analysis, as shown in Figure 3. It can be observed that the best performance is achieved when σ is set to 0.5. As σ increases, the performance gradually decreases and converges. In fact, as σ grows, the Gaussian Self-Attention mechanism gradually degenerates into a standard self-attention mechanism.

5.4 Average Performance and Computation Cost with Different Language Models

To investigate the effect of using different pre-trained language models and the corresponding computation costs, we conduct experiments and record the average performance and inference costs in Table 4. Using roberta-base (Liu et al., 2019) improves the model's average performance by 0.86%. With the larger versions, Bert and Roberta improve the model's average performance by 2.32% and



(a) $\mathcal{E} \rightarrow \mathcal{I}$ (b) $\mathcal{M} \rightarrow \mathcal{I}$ (c) $\mathcal{I} \rightarrow \mathcal{E}$

(d) $\mathcal{M} \rightarrow \mathcal{E}$ (e) $\mathcal{I} \rightarrow \mathcal{M}$ (f) $\mathcal{E} \rightarrow \mathcal{M}$

Figure 4: Fine-grained analysis of different methods, with the proportion of unseen emotions also presented.

1.81%, respectively. However, the average inference time per sample increases by 2.91 ms and 2.92 ms, respectively.

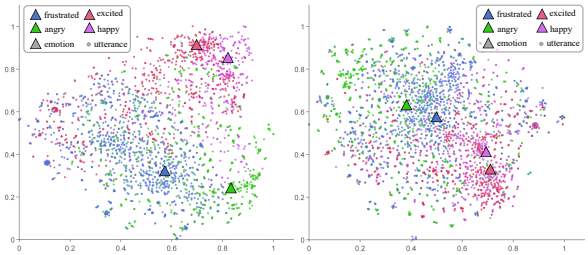
5.5 Fine-grained Analysis

As shown in Figure 4, we conduct an experiment to demonstrate the fine-grained performance of different methods. Comparing the performance of ProEmoTrans and GPT-4o, we can observe that ProEmoTrans performs better in most unseen emotions. However, as the emotion proportion decreases, ProEmoTrans shows a more noticeable decline in performance. We believe this is due to GPT-4o relying on prior knowledge, while ProEmoTrans depends on the quality of prototype-based representation learning, which makes it more sensitive to the distribution of categories.

Removing the LED (-w/o *LED*) causes a performance drop across all unseen emotions, to varying degrees, highlighting the LED's comprehensive contribution. Similarly, removing the CRF (-w/o *CRF*) also leads to a nearly overall performance decline, but in some cases, it improves performance. For example, in subplot (f), it leads to a 2.45% increase for *surprise*. This suggests that while the CRF layer optimizes global performance, it may not be ideal for certain local categories.

5.6 Visualization

To provide more interpretability, we visualize the embedding space of utterances and unseen emotions on $\mathcal{E} \rightarrow \mathcal{I}$ datasets using t-SNE (Van der Maaten and Hinton, 2008), as shown in Figure 5. First, we find that positive emotions (*excited* and *happy*) are farther apart from negative emotions (*frustrated* and *angry*), while emotions of the same



(a) Proposed ProEmoTrans (b) -w/o *CRF*

Figure 5: t-SNE visualization of utterance and emotion embeddings in $\mathcal{E} \rightarrow \mathcal{I}$ datasets.

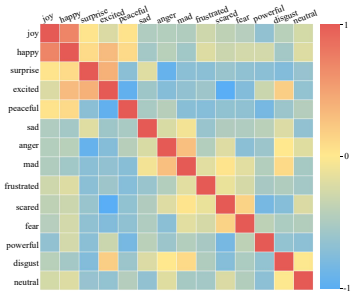


Figure 6: Heatmap of emotion prototype similarities.

polarity are closer to each other, which aligns with our intuition. Next, comparing subfigures (a) and (b), we can see that adding the CRF layer enhances the distinguishability of utterance and emotion embeddings, demonstrating the positive impact of the CRF layer and AVD algorithms in our method.

We also collect all the emotion prototype embeddings and compute their cosine similarities. The resulting heatmap is shown in Figure 6. It can be observed that, first, the cosine similarity is higher between similar emotions (e.g., *happy* and *joy*). Second, there is a more pronounced difference in similarity between positive and negative emotions.

Models	$\mathcal{E} \rightarrow \mathcal{I}$	$\mathcal{M} \rightarrow \mathcal{I}$	$\mathcal{I} \rightarrow \mathcal{E}$	$\mathcal{M} \rightarrow \mathcal{E}$	$\mathcal{I} \rightarrow \mathcal{M}$	$\mathcal{E} \rightarrow \mathcal{M}$	Average
Contrastive Similarity (Ours)	**37.27**	**32.36**	**28.34**	**20.73**	**38.59**	**35.64**	**32.16**
- w *Euclidean distances*	35.12	30.03	27.23	18.89	37.77	34.07	30.52
- w *Cosine similarity*	36.45	31.91	27.67	20.10	37.85	34.98	31.49
- w *Dot Product*	35.78	30.56	27.34	19.52	36.29	34.27	30.63

Table 5: The results of comparing contrastive similarity with other similarity metrics.

Models	$\mathcal{E}$	$\mathcal{I}$	$\mathcal{M}$
KET	13.12	16.46	8.97
TUCORE-GCN	13.11	15.27	25.96
EmotionFlow	14.65	16.99	29.34
SPCL	14.99	18.73	29.41
CTPT	<u>20.57</u>	<u>31.82</u>	<u>31.28</u>
ProEmoTrans (Ours)	**22.46**	**33.20**	**33.29**

Table 6: Performance of different ERC datasets under the few-shot settings (16-shot). All the baseline results are retrieved from Xu et al. (2023). We **bolded** the best result and <u>underline</u> the second best.

5.7 More Additional Experiments

5.8 Analysis on Contrastive Similarity

The contrastive similarity (Oord et al., 2018) can effectively measure the difference between two embeddings. To validate its effectiveness, we conducted experiments comparing it with other similarity metrics. The results are shown in Table 5. When using Euclidean distance, cosine similarity, and dot product, the model's performance decreased by 1.64%, 0.67%, and 1.53%, respectively, which proves the effectiveness of contrastive similarity.

5.8.1 Few-shot Performance

Our model can also be used for few-shot prediction without any modifications. To investigate the performance of our model in the few-shot setting, we conducted experiments as shown in Table 6. To ensure a fair comparison with the baselines, we follow the 16-shot setting and use weighted macro-F1 as the evaluation metric. The applied baselines include: **KEY** (Zhong et al., 2019) addresses ERC tasks by utilizing external knowledge bases. **TUCORE-GCN** (Lee and Choi, 2021) and **EmotionFlow** (Song et al., 2022b) are GCN-based and RNN-based ERC model, respectivly. **SPCL** (Song et al., 2022a) uses supervised contrastive learning to address the class imbalance problem in ERC. **CTPT** (Xu et al., 2023) is introduced in Section 4.3. According to the results, our ProEmoTrans outperforms the best baseline, CTPT, by 1.89%, 1.38%, and 2.01% on the three datasets, respectively. This demonstrates that our model also performs excellently in the few-shot setting.

6 Conclusion

In this paper, we propose a simple and effective method named ProEmoTrans for the newly proposed UERC task. First, we introduce an LLM-enhanced Emotion Description module to enhance emotion prototype learning. Next, a parameter-free Gaussian Self-Attention mechanism is designed to aggregate useful information from the conversation while filtering out noise. This mechanism can learn inter-utterance relations and prevent overfitting that could arise from parameter training. Finally, we propose an Attention Viterbi Decoding algorithm to transfer the useful seen emotion dependencies learned during training to unseen emotions. Extensive experiments on three datasets validate the effectiveness of our approach and the individual modules we designed. In future work, our goal is to further optimize prototype representations.

7 Limitations

Our LLM prompt templates rely on manual design, and their effectiveness has not been verified with more complex emotions. Developing automated prompt-tuning templates would be an interesting avenue for exploration. Additionally, our approach focuses solely on the text modality and does not incorporate multi-modal information, such as facial expressions, which could provide valuable additional information.

8 Acknowledgments

This research is supported by the National Key R&D Program of China (No. 2023YFC3303800), NSFC through grants 62322202, 62441612 and 62476163, Beijing Natural Science Foundation through grant L253021, Local Science and Technology Development Fund of Hebei Province Guided by the Central Government of China through grants 246Z0102G and 254Z9902G, the "Pionee" and "Leading Goose" R&D Program of Zhejiang through grant 2025C02044, Hebei Natural Science Foundation through grant F2024210008, and the Guangdong Basic and Applied Basic Research Foundation through grant 2023B1515120020.

References

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

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

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online. Association for Computational Linguistics.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909.

Paul Ekman. 1999. *Basic Emotions*, chapter 3. John Wiley and Sons, Ltd.

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, Online.

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, Hong Kong, China.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10835–10852, Toronto, Canada. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. 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 7042–7052, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13449–13467, Toronto, Canada. Association for Computational Linguistics.

Zehan Li, Fu Zhang, and Jingwei Cheng. 2024. AlignRE: An encoding and semantic alignment approach for zero-shot relation extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2957–2966, Bangkok, Thailand and virtual meeting.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

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.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4324–4330. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Robert Plutchik and Henry Kellerman. 2013. *Theories of emotion*, volume 1. Academic press.

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, Florence, Italy. Association for Computational Linguistics.

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, Online.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022a. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022b. Emotionflow: Capture the dialogue level emotion transitions. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 8542–8546. IEEE.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yige Xu, Zhiwei Zeng, and Zhiqi Shen. 2023. Efficient cross-task prompt tuning for few-shot conversational emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11654–11666, Singapore.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. 2025. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 14(4):3269–3280.

Zhenyu Yang, Xiaoyang Li, Yuhu Cheng, Tong Zhang, and Xuesong Wang. 2024. Emotion recognition in conversation based on a dynamic complementary graph convolutional network. *IEEE Transactions on Affective Computing*, 15(3):1567–1579.

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, Mexico City, Mexico. Association for Computational Linguistics.

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*.

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, Toronto, Canada.

Jun Zhao, WenYu Zhan, Xin Zhao, Qi Zhang, Tao Gui, Zhongyu Wei, Junzhe Wang, Minlong Peng, and Mingming Sun. 2023. RE-matching: A fine-grained semantic matching method for zero-shot relation extraction. In *Proceedings of the 61st Annual Meeting of ACL (Volume 1: Long Papers)*, pages 6680–6691, Toronto, Canada.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on EMNLP and the 9th IJCNLP (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

A More Details of Experiments Settings

A.1 Datasets

Source	Target		
	$\mathcal{I}$	$\mathcal{E}$	$\mathcal{M}$
$\mathcal{I}$	/	*po*, *pe*, *sc*, *jo*, *ma*	*su*, *di*, *fe*, *jo*
$\mathcal{E}$	*ex*, *fr*, *ha*, *an*	/	*su*, *di*, *fe*, *an*
$\mathcal{M}$	*ex*, *fr*, *ha*	*po*, *pe*, *sc*, *ma*	/

Table 7: Statistics of unseen emotions under different source and target settings. We use the first two letters to denote the emotions in Table 1, for example, *po* stands for *powerful*.

The statistics of the unseen emotions under different source and target settings are shown in Table 7. For example, if we chose $\mathcal{I}$ as source dataset and $\mathcal{E}$ as target dataset, the unseen emotions are *powerful*, *peaceful*, *scared*, *joy*, and *mad*.

A.2 Baselines

The details of baselines are as follows:

- **DialogueGCN** (Ghosal et al., 2019) uses a GCN to model the inter-utterance dependency.
- **DialogueCRN** (Hu et al., 2021) is one of the best RNN-based ERC models. They design multiple rounds of reasoning modules to extract and integrate emotional cues.
- **DualGAT** (Zhang et al., 2023) introduces a Dual Graph Attention Network to capture complex dependencies of discourse structure and speaker-aware context.
- **SACL-LSTM** (Hu et al., 2023) proposes a supervised adversarial contrastive learning method for learning class-spread structured representations.
- **SCCL** (Yang et al., 2023) proposes a supervised cluster-level contrastive learning method to incorporate measurable emotion prototypes.
- **EACL** (Yu et al., 2024) proposes an emotion-anchored contrastive learning framework, which generates more distinguishable utterance representations for similar emotions.
- **CPTC** (Xu et al., 2023) leverages sharable cross-task knowledge from the source task to improve few-shot performance.

We made the necessary modifications for each baseline to enable zero-shot prediction.

B More Additional Experiments

B.1 Results with Parameterized Modules

In supervised settings, previous methods have designed various parameterized modules to help learn better utterance representations. In the zero-shot setting, to validate their effectiveness, we conduct comparative experiments by replacing the Gaussian Self-Attention module in our model with LSTM, GCN, and GAT. The experimental results are shown in Table 8. It can be observed that the performance is quite weak, which proves that overfitting due to the parameter module severely hinders the generalization performance.

B.2 Utterance-level Performance

We conducted a comparative experiment on zero-shot ERC at the utterance level, with results shown in Table 9, where **-w utterance-level** refers to applying LLM baselines to prompt each individual utterance. Our experiments uncovered some intriguing findings: On the longer dialogue dataset (IEMOCAP, avg. length 52), utterance-level classification significantly outperformed the original conversation-level approach. We believe that excessively long conversations hinder LLM's emotional analysis capability by overwhelming context processing. On the other two datasets (avg. lengths 12 and 9), utterance-level performance was slightly lower than conversation-level. We attribute this to the loss of contextual information, which poses challenges for utterances with ambiguous emotional cues or those that are very brief. For example, the utterance "That only took me an hour." was misclassified as joy at the utterance level, but correctly classified as sad at the conversation level when the broader topic (divorce) was considered. Crucially, our method consistently maintains an advantage across different conversation lengths, despite the observed variations in zero-shot LLM classification performance.

C Details of LED Generated Descriptions

To eliminate biases introduced by the quality of generated descriptions, we regenerate new descriptions in each of the five random runs. The emotion descriptions generated using the LED module in one of the five runs are shown in Table 10.

Models	$\mathcal{E} \rightarrow \mathcal{I}$	$\mathcal{M} \rightarrow \mathcal{I}$	$\mathcal{I} \rightarrow \mathcal{E}$	$\mathcal{M} \rightarrow \mathcal{E}$	$\mathcal{I} \rightarrow \mathcal{M}$	$\mathcal{E} \rightarrow \mathcal{M}$	Average
Proposed ProEmoTrans	**37.27**	**32.36**	**28.34**	**20.73**	**38.59**	**35.64**	**32.16**
-w *LSTM*	9.37	8.82	7.74	6.04	7.70	6.78	7.74
-w *GCN*	8.42	8.10	6.47	5.19	7.63	5.97	6.96
-w *GAT*	7.57	8.93	6.19	5.65	7.11	6.08	6.92

Table 8: Comparative experiments by replacing the GSA module with LSTM, GCN, and GAT.

Models	$\mathcal{E} \rightarrow \mathcal{I}$	$\mathcal{M} \rightarrow \mathcal{I}$	$\mathcal{I} \rightarrow \mathcal{E}$	$\mathcal{M} \rightarrow \mathcal{E}$	$\mathcal{I} \rightarrow \mathcal{M}$	$\mathcal{E} \rightarrow \mathcal{M}$	Average
Proposed ProEmoTrans	**37.27**	**32.36**	**28.34**	**20.73**	**38.59**	**35.64**	**32.16**
DeepSeek-V3	25.69	26.26	24.10	18.68	35.15	29.76	26.61 ($5.55^{\downarrow}$)
-w utterance-level	30.06	31.52	23.27	19.04	34.86	29.45	28.03 ($4.13^{\downarrow}$)
GPT-4o	24.88	26.10	24.51	18.25	35.28	29.35	26.40 ($5.76^{\downarrow}$)
-w utterance-level	30.24	31.25	24.22	18.48	35.07	29.51	28.13 ($4.03^{\downarrow}$)

Table 9: Performance comparison on Zero-Shot ERC at the utterance level.

Emotion	Description	LED Generated Description
joy	a feeling of extreme happiness or cheerfulness, especially related to the acquisition or expectation of something good.	1. When she received the good news, a bright smile spread across her face. 2. At the celebration, laughter filled the room, creating a vibrant atmosphere.
happy	the emotion of being happy; elation.	1. She smiled as she read the message, feeling content with the little surprise. 2. As the music played, a warm sense of fulfillment washed over her, making everything else fade away.
surprise	the feeling that something unexpected has happened.	1. The unexpected turn of events left him speechless, eyes wide in disbelief. 2. She could hardly believe her ears when she heard the astonishing news.
excited	having great enthusiasm, passion, and energy.	1. Her heart raced as she opened the envelope containing the results. 2. He couldn't sit still, eagerly anticipating the start of the event.
peaceful	motionless and calm.	1. The gentle sound of the waves lapping against the shore filled her with calm. 2. Sitting under the shade of the old tree, he felt completely at ease.
sad	emotionally negative and feeling sorrow.	1. He stared out the window, his heart heavy with a lingering sense of loss. 2. As she walked through the empty hall, a wave of nostalgia washed over her.
anger	displaying or feeling anger.	1. She clenched her fists and glared at him when she heard the unfair criticism. 2. His face turned red as he confronted the person who had betrayed his trust.
mad	feeling annoyed.	1. He slammed the door shut, frustration boiling beneath the surface. 2. Her eyes flashed with irritation as she listened to the unfair remarks.
frustrated	the feeling of annoyance at impossibility from resistance or inability to achieve something.	1. She let out a sigh and ran her fingers through her hair, feeling exasperated with the situation. 2. He stared at the puzzle pieces scattered on the table, unable to find a solution.
scared	feeling afraid and frightened.	1. A cold sweat broke out on his forehead as he heard footsteps behind him in the dark. 2. She held her breath, feeling a knot tighten in her stomach during the thunderstorm.
fear	a strong, unpleasant emotion or feeling caused by actual or perceived danger or threat.	1. In the dark alley, a sudden noise made his heart race with unease. 2. She felt a chill run down her spine as shadows flickered around her.
powerful	having, or capable of exerting, power or influence.	1. Standing at the edge of the cliff, she felt an overwhelming sense of strength and determination. 2. The speaker's voice resonated through the hall, commanding everyone's attention.
disgust	to cause an intense dislike for something.	1. I couldn't believe it when my teammate ignored my advice during the game. 2. It drove me crazy when the internet kept disconnecting while I was working.
neutral	neither positive nor negative.	1. He sat quietly, showing no particular reaction to the events around him. 2. The room was filled with a quiet stillness as everyone focused on their tasks.

Table 10: Details of LED generated descriptions