EAAN-ERC: Expert Adaptive Agreement Network for Emotion Recognition in Conversation

Anonymous ACL submission

Abstract

Existing studies for emotion recognition in conversation (ERC) focus on modeling conversational context, however, they overlook the influence of diverse human evaluator panels on the emotional annotations of datasets. We observed in an existing ERC dataset that different evaluator panels for assessed utterances in conversations impact the final emotional evaluation results due to the subjective nature of each evaluator's perception and interpretation of emotion. To address this issue, we propose a novel Expert Adaptive Agreement Network for Emotion Recognition in Conversation (EAAN-ERC), a method designed to imitate the evaluation and annotating process of emotions by diverse evaluator panels. Specifically, we first mimic experienced evaluators by setting up multiple expert models. Subsequently, we emulate diverse evaluator panels by adaptively mixing expert models matched with specified evaluator panels. Furthermore, we simulate the evaluator panels' emotional evaluation by computing emotional probability and confidence for the assessed utterance. Ultimately, we mimic the agreement of an evaluator panel by integrating emotional probability with confidence. Extensive experiments on the widely used ERC dataset IEMOCAP, which to the best of our knowledge is the only ERC dataset that makes the evaluator panel information publicly available, have reflected exceptional results, establishing new standards in weighted average accuracy and F1-score. These promising results demonstrate the efficacy of our EAAN-ERC.

1 Introduction

Emotion Recognition in Conversation (ERC) is a widely researched task in Natural Language Processing (NLP). Its primary objective is to identify the emotional state of a speaker throughout a conversation. The ERC task is significant in various applications, such as emotional support (Liu et al., 2021; Tu et al., 2022), customer service (Li et al., 2019; Lou et al., 2023; Qiu et al., 2020), and more. Identifying emotions in a conversation is challenging due to contextual dependencies. Current methods, such as recurrence-based approaches (Poria et al., 2017; Majumder et al., 2019; Ghosal et al., 2020; Hu et al., 2021, 2023) focus on participant speaking order but struggle with distant utterances. To overcome this, graph-based methods (Ghosal et al., 2019a; Shen et al., 2021b; Zhang et al., 2023a) utilize participants' information and location relationships, enabling query utterances to extract insights from both nearby and distant utterances.

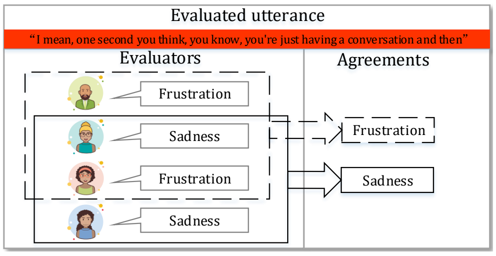

Figure 1: A case illustrating the impact of different sets of evaluators on emotional evaluation results.

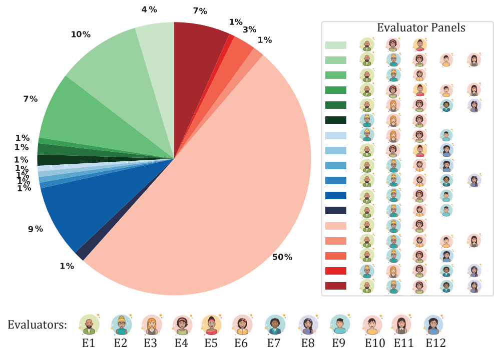

Figure 2: Percentage of conversational utterances evaluated by different evaluator panels in the IEMOCAP (Busso et al., 2008) dataset.

However, existing studies focus on modeling conversational context while overlooking the impact of diverse human evaluator panels on emotional annotations within conversational utterances. In the existing ERC corpus, the process of emotion annotation primarily includes enlisting human evaluators to create diverse evaluator panels. These panels subjectively assess the emotions conveyed in conversational utterances and then utilize a majority vote to assign emotional labels. Considering the annotation process, we claim that ***diverse evaluator panels can influence emotion evaluation results*** due to the subjective nature of each evaluator's perception and interpretation of emotion. As illustrated in Figure 1, when the evaluator panel {E1, E2, E4} assesses a given utterance, the agreement yields a frustration result, whereas, with evaluators {E2, E4, E12}, the final evaluation result is Sadness. Additionally, we have noted the presence of different evaluator panels in public ERC datasets assigned to evaluate distinct utterances. As illustrated in Figure 2, there are 17 distinct evaluator panels (marked in different colors) for assessing utterances in conversations. For instance, the evaluator panel {E1, E2, E4} evaluated around 50% of all conversational utterances in the ERC dataset IEMOCAP (Busso et al., 2008). This underscores the significance of leveraging evaluator panel information for enhancing emotion inference in ERC datasets with diverse evaluator panels.

In this paper, we introduce a novel approach called Expert Adaptive Agreement Network for Emotion Recognition in Conversation (EAAN-ERC). This method is carefully crafted to emulate the process of emotion evaluation and annotation conducted by diverse evaluator panels. Specifically, initially, we formulate an ERC Expert Pool Initialization process to mimic evaluators by setting up multiple expert models. These expert models acquire the emotional evaluation knowledge of evaluators and serve as foundational components, enabling the representation of diverse evaluator panels for emotion assessment. In particular, we introduce a proxy expert model to stand in for evaluators absent from the training set or those who infrequently participate in evaluations in the training set. The advantage of this lies in our model's ability to handle unseen evaluators outside the training set or those whose emotion evaluation knowledge is challenging to learn due to their limited evaluation samples. Following that, we establish an ERC Expert Assignments module to emulate diverse evaluator panels by adaptively mixing expert models matched with specified evaluator panels. This enables us to utilize models to represent designated evaluator panels in existing ERC datasets, ensuring that our method considers the influence of different evaluator panels on emotion evaluation. Moreover, we construct an ERC Expert Evaluations module to simulate the emotional evaluation of each evaluator within various evaluator panels, which involves computing emotional probability and confidence for the assessed utterance using the corresponding expert model. This enables an evaluator panel-specific emotion inference. In the end, we establish an ERC Expert Agreements module to mimic the agreed process of an evaluator panel by integrating emotional probability with confidence. Incorporating each evaluator's confidence in this integration can, to a certain extent, enhance the assessment of the agreed emotion.

To the best of our knowledge, IEMOCAP (Busso et al., 2008) is the only ERC dataset that makes the evaluator panels information associated with each utterance publicly available, facilitating the validation of our proposal's effectiveness. Consequently, we conduct extensive experiments on the widely used ERC dataset IEMOCAP. The experimental results show that our EAAN-ERC model performs better than the state-of-the-art models in both weighted average accuracy and F1-score, demonstrating its effectiveness. Overall, the main contributions of this paper are summarized as follows:

- We present a novel approach named Expert Adaptive Agreement Network for Emotion Recognition in Conversation (EAAN-ERC). This method is designed to imitate the emotion evaluation and annotation process carried out by diverse evaluator panels.

- Specifically, we design four components, that is, ERC Expert Pool Initialization, ERC Expert Assignments, ERC Expert Evaluations, and ERC Expert Agreements, to imitate enlisted evaluators, diverse evaluator panels, emotional evaluations, and emotional agreements, respectively.

- To the best of our knowledge, different from existing studies that model from the perspective of conversational context for ERC, we are the first to model from the perspective of the evaluator panels for more accurate emotion inference in conversations.

- We conduct extensive experiments on the widely-used ERC dataset IEMOCAP. Experimental results demonstrate that EAAN-ERC outperforms the existing state-of-the-art models in terms of weighted average accuracy and F1-score. This demonstrates the effectiveness of our EAAN-ERC in the context of ERC.

2 Related Work

2.1 Emotion Recognition in Conversation

Existing research on Emotion Recognition in Conversations (ERC) primarily focuses on deducing emotional categories by constructing models of conversational context using recurrence or graph propagation structures.

In recurrence-based approaches, bc-LSTM (Poria et al., 2017) captures context-level features from surrounding utterances based on Long Short Term Memories (LSTMs) (Hochreiter and Schmidhuber, 1997; Graves, 2014). DialogRNN (Majumder et al., 2019) utilizes three GRUs to sequentially monitor the speaker's state, contextual information, and emotion throughout a conversation. COSMIC (Ghosal et al., 2020) employs GRUs to leverage various aspects of commonsense knowledge and learn interactions between participants. DialogueCRN (Hu et al., 2021) integrates reasoning modules over multiple turns, employing LSTMs to extract and integrate emotional cues from a cognitive perspective. CauAIN (Zhao et al., 2022) uses causal clues to model speaker dependencies. SACL-LSTM (Hu et al., 2023) proposes a contextual adversarial training strategy to learn more diverse features from context.

In terms of graph-based methods, DialogueGCN (Ghosal et al., 2019a) uses a directed graph to model conversational context, representing utterances as nodes and capturing speaker dependencies and positions as edges. This approach effectively addresses challenges in context propagation, enabling a comprehensive understanding of the interplay between speakers. DAG-ERC (Shen et al., 2021b) constructs a directed acyclic graph from the conversation, considering speaker identity and positional relationships to propagate remote and local information. SGED+DAG (Bao et al., 2022) explores speaker interactions with a one-layer DAG. Transformer (Vaswani et al., 2017), while not explicitly a graph-based method, can be considered as such due to the fully connected graph-like nature of its self-attention mechanism (Shen et al., 2021c). DialogXL (Shen et al., 2021a) enhances XLNet (Yang et al., 2019) by incorporating improved memory and dialog-aware self-attention. **TODKAT (Zhu et al., 2021)** integrates commonsense knowledge and a task for detecting topics based on Transformer. CoG-BART (Li et al., 2022) leverages a response generation task to enhance BART(Lewis et al., 2020a)'s ability. SPCL (Song et al., 2022) proposes supervised prototypical contrastive learning loss for imbalanced classification and difficulty-measure function for curriculum learning to handle extreme samples. MPLP (Zhang et al., 2023c) mimics the thinking process of a human being based on BART. HAAN-ERC (Zhang et al., 2023b) employs a hierarchical approach within the Transformer architecture to model various influences, effectively inferring the emotional category of speakers. DualGAT (Zhang et al., 2023a) introduces a novel Dual Graph Attention network to address the oversight of discourse structure in conversation by simultaneously incorporating complementary elements of discourse structure and speaker-aware context.

Unlike the above methods that model from the perspective of conversational context for ERC, this paper models from the perspective of the evaluator panels for more accurate emotion inference in conversations.

2.2 Label Disagreement Modeling

There exist several studies to model disagreed labels for emotion recognition based on individual utterances (ERI) (Chou et al., 2022; Wu et al., 2022; Han et al., 2017; Dang et al., 2017; Atcheson et al., 2019; Wu et al., 2023a; Sridhar and Busso, 2020; Ando et al., 2019, 2018; Fayek et al., 2016) as well as ERC (Wu et al., 2023b). For ERI, Chou et al. (Chou et al., 2022) propose to leverage the relation between emotions to enhance disagreed label learning. Wu et al. (Wu et al., 2022) propose resolving the issue of inconsistent annotations in hard emotion labels for classification using Bayesian statistics. Han et al. (Han et al., 2017) propose a 'soft-prediction' framework to shape a humanoid emotion prediction. Dang et al. (Dang et al., 2017) propose a paradigm that incorporates the uncertainty information of speech frames by explicitly accounting for multi-rater variability in the system. Atcheson et al. (Atcheson et al., 2019) combine Gaussian processes with neural networks, which take advantage of the flexible modeling power of LSTM networks along with the probabilistic handling of ambiguity offered by Gaussian processes for continuous emotion recognition from speech.

We et al. (Wu et al., 2023a) propose a Bayesian approach called deep evidential emotion regression (DEER) to estimate the uncertainty in emotion attributes from speech. Sridhar et al. (Sridhar and Busso, 2020) used regression models with emotion uncertainty to predict speech emotion. They utilized Monte Carlo dropout, which involves multiple feed-forward passes through a deep neural network using dropout regularization in both training and inference. Ando et al. (Ando et al., 2019) introduce estimating multi-label emotion existence (MLEE) as an auxiliary task to support dominant emotion recognition from speech. Ando et al. (Ando et al., 2018) utilize ambiguous emotional utterances with soft-target training to address the lack of training data compared to model complexity. Fayek et al. (Fayek et al., 2016) incorporate inter-annotator variability for speech emotion recognition. However, Different from the above studies which mainly focus on soft-prediction of emotion and uncertainty estimation for ERI, our approach emphasizes the ERC task and aims to imitate the evaluation and annotating process of emotions to address the disturbing subjective perception of evaluator panels.

To address the inherent ambiguity of emotions and the subjectivity of human perception of ERC, Wu et al. (Wu et al., 2023b) propose a distribution-based ERC approach, which introduces Bayesian training loss by conditioning each emotional state on an utterance-specific Dirichlet prior distribution, and conduct experiments on the IEMOCAP dataset, achieving good classification accuracy. Different from the distribution-based study, our evaluator identity information-based approach considers addressing the subjectivity of human perception for ERC from the perspective of the imitation of diverse human evaluator panels.

3 Methodology

3.1 Problem Definition

Considering a conversation $\{u_0, u_1, ..., u_T\}$ composed of a sequence of utterances, we define u_t as the t-th utterance in this conversation, where $t \in \{0, ..., T\}$. Each utterance u_t is uttered by the speaker $s(u_t) \in \mathcal{S}$, where $\mathcal{S}$ is the collection of all of the participants. Each utterance u_t is evaluated by an evaluator panel $e(u_t) \subseteq \mathcal{E}$, where $\mathcal{E} = \{E_1, E_2, ..., E_{|\mathcal{E}|}\}$ is the collection of all of the evaluators. We define $i \in \{1, 2, ..., |\mathcal{E}|\}$. $y_t^i \in \mathbb{R}^C$ is the emotional category label of utterance u_t evaluated by the evaluator $E_i \in e(u_t)$, where C is the number of emotion categories. $y_t \in \mathbb{R}^C$ is the emotional category label of utterance u_t agreed by $e(u_t)$.

Given an utterance u_t to be evaluated, along with its conversational context $context_t = \{s(u_0), u_0, ..., s(u_{t-1}), u_{t-1}, s(u_t)\}$ and evaluator panel $e(u_t)$, our goal is to design a model to predict the emotional category label y_t.

3.2 Our Model

In this section, we introduce our proposed EAAN-ERC model. The overall architecture of this model is illustrated in Figure 3, which comprises four components: ERC Expert Pool Initialization, ERC Expert Assignments, ERC Expert Evaluations, and ERC Expert Agreements.

3.2.1 ERC Expert Pool Initialization

we first need to create models to represent human evaluators. The process involves counting the evaluators present in the training set. Each evaluator is then represented by an expert model, denoted as Expert i, which will be used for emotion evaluation. The expert model is composed of three parts: an ERC backbone that learns emotion representation, an emotion classifier that computes emotion logits, and a confidence regressor that calculates the expert model's confidence in the emotion assessment of the utterance. We define conf. as the abbreviation of confidence. The architecture of each expert model is defined in Equation (1).

$$\begin{aligned} R_t &= \text{ERC Backbone}(u_t, context_t) \\ logits_t &= W_1 R_t + b_1 \\ conf._t &= W_2 R_t + b_2 \end{aligned} \tag{1}$$

where $R_t \in \mathbb{R}^{D_e}$, $W_1 \in \mathbb{R}^{C \times D_e}$, $b_1 \in \mathbb{R}^C$, $logits_t \in \mathbb{R}^C$, $W_2 \in \mathbb{R}^{1 \times D_e}$, $b_2 \in \mathbb{R}^1$, $conf._t \in \mathbb{R}^1$. D_e is the dimension of the emotion representation R_t.

In particular, we observed in ERC datasets that the number of utterances evaluated by some evaluators is very small, which may make it difficult for the corresponding expert model to learn the evaluator's emotional evaluation experience and represent it effectively. To overcome this issue, we sort the evaluators in descending order according to the number of utterances they evaluated and set a threshold M to filter the Top-M evaluators. we then create expert models to represent the Top-M evaluators. This will result in a pool of expert models called the ERC expert pool (EEP). Each

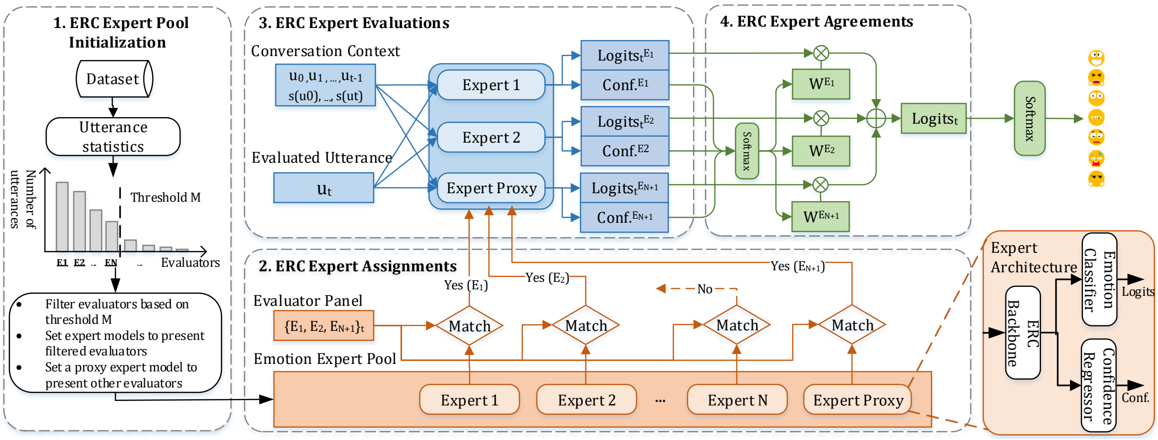

Figure 3: The overall architecture of our EAAN-ERC.

expert model Expert i is supervised by the emotion labels y^i evaluated by the corresponding evaluator E_i to learn the knowledge of this evaluator, which is defined in Equation (2). Formally, for $\forall$ Expert $i \in$ EEP:

$$\begin{aligned} logits_t^i, \ldots &= \text{Expert } i(u_t, context_t) \\ \mathcal{P}_t^i &= \text{Softmax}(logits_t^i) \\ \mathcal{L}^i &= -\sum_{\beta=1}^{B}\sum_{t=1}^{T(\beta)} \log \mathcal{P}_{\beta,t}^i[y_{\beta,t}^i] \end{aligned} \tag{2}$$

where Expert i is one of the experts in the expert pool, $\mathcal{P}_t^i \in \mathbb{R}^C$ denotes the probability distribution of emotional categories, B is the number of conversations, $T(\beta)$ is the number of utterances in the β-th conversation, $y_{\beta,t}^i$ is the ground truth label evaluated by E_i, and $\mathcal{L}^i$ is the training loss of Expert i.

In addition, we also set up a proxy expert model, namely Expert $Proxy$, to represent unseen evaluators outside the training set or those who infrequently participate in evaluations in the training set (lower than the threshold M). The training process is similar to other expert models. Different from other expert models, Expert $Proxy$ is supervised by the emotion labels y, which are agreed upon by different evaluator panels. This indicates that the proxy model trained using these diverse evaluator panels agreed emotion labels is somewhat evaluator-independent to a certain extent and is more suitable for handling emotion inference of unseen evaluators. Finally, the Expert $Proxy$ is also placed in the EEP for subsequent procedures. Through the above process, we complete the initialization of the EEP. The expert models within the EEP serve as foundational components, enabling the representation of various evaluator panels for emotion assessment.

3.2.2 ERC Expert Assignments

Upon completing the initialization of the EEP, we establish an ERC Expert Assignments module, to assign the corresponding experts to the current utterance undergoing evaluation. Specifically, we utilize the evaluator panel $e(u_t)$ corresponding to the utterance u_t to determine which ERC expert models are designated for the emotion evaluation of u_t. If $e(u_t)$ is an empty set, signifying that evaluators for assessing the utterance are unseen, then the Expert $Proxy$ is assigned to conduct the emotion evaluation. Conversely, for each Expert i in the EEP, if the evaluator E_i corresponding to the expert model Expert i is present in the evaluator panel $e(u_t)$, we assign the expert model Expert i to represent the evaluator E_i for utterance u_t. The Expert $Proxy$ is assigned to represent any evaluators that Expert i uncovered.

In this way, the assigned expert models construct an ERC Expert Set EES_t for the emotional evaluation of utterance u_t.

3.2.3 ERC Expert Evaluations

Once we have obtained the ERC Expert Set EES_t, we employ the expert models within it to assess the emotion conveyed by the utterance u_t, defined in Equation (3). Specifically, we feed the evaluated utterance u_t and its conversation context $context_t$ into each expert model within the EES_t, obtaining the corresponding emotion logits and confidence values. Formally, we define each Expert $j \in EES_t$:

$$logits_t^j, conf._t^j = \text{Expert } j(u_t, context_t) \tag{3}$$

The emotion logits are then used to compute emotion probabilities, while the confidence values determine the weight of each expert model on the utterance in the subsequent steps. The calculation of confidence values helps enhance, to a certain extent, the assessment of the agreed emotion.

Subsequently, the emotion logits and confidence values obtained earlier are input into respective sets named $LogitsSet_t$ and $Conf.Set_t$, defined in Equation (4).

$$
\begin{aligned}
LogitsSet_t &= \{logits_t^j|\text{Expert } j \in EES_t\} \\
Conf.Set_t &= \{conf._t^j|\text{Expert } j \in EES_t\}
\end{aligned} \tag{4}
$$

3.2.4 ERC Expert Agreements

Each expert model in the EES_t has deduced the corresponding emotional logits and confidence scores for the assessed utterance, which are utilized to facilitate an agreement among experts for deriving the collectively agreed-upon emotional evaluation result. Specifically, first, the confidence scores in $Conf.Set_t$ for the utterance u_t undergo conversion into confidence probabilities through the SoftMax function, serving as weights for the emotion logits inferred by each expert model in EES_t, defined in Equation (5).

$$WeightSet_t = \text{Softmax}(Conf.Set_t) \tag{5}$$

Then, the emotion logits inferred by all expert models in EES_t undergo weighting and averaging with corresponding weights to yield the agreed-upon emotional logits. Subsequently, the Softmax function is applied to calculate the probability distribution $\mathcal{P}_t^{agree}$ of agreed-upon emotions. This process is defined in Equation (6).

$$
\begin{aligned}
&logits_t^{agree} = \text{Sum}(\\
&\text{Concat}(LogitsSet_t) * \text{Concat}(WeightSet_t)) \\
&\mathcal{P}_t^{agree} = \text{Softmax}(logits_t^{agree})
\end{aligned} \tag{6}
$$

where $logits_t^{agree} \in \mathbb{R}^C$, $\mathcal{P}_t^{agree} \in \mathbb{R}^C$.

Finally, we utilize cross-entropy to calculate the error $\mathcal{L}^{agree}$ between the probability distribution $\mathcal{P}_t^{agree}$ of agreed-upon emotions and the corresponding ground truth, defined in Equation (7).

$$\mathcal{L}^{agree} = -\sum_{\beta=1}^{B}\sum_{t=1}^{T(\beta)} \log \mathcal{P}_{\beta,t}^{agree}[y_{\beta,t}] \tag{7}$$

3.2.5 Objective Function

In the final step, we sum the emotional losses of expert models in the EEP and then weight-average this sum with the loss of emotions after expert agreements, to derive the final loss $\mathcal{L}$ serving as the objective function. We employ an optimization algorithm based on backpropagation, such as Adam(Kingma and Ba, 2014), to update the model parameters, thereby optimizing the objective function. The objective function is defined in Equation (8).

$$\mathcal{L} = \mathcal{L}^{agree} + \alpha * \text{Sum}\{\mathcal{L}^i|\text{Expert } i \in \text{EEP}\} \tag{8}$$

where $\alpha > 0$ is the weight of ERC expert models' emotional losses.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We assess the effectiveness of our approach using the widely used ERC dataset IEMOCAP (Busso et al., 2008). The statistical findings for the dataset are presented in Table 1, focusing solely on the text modalities within them. The ERC dataset known as IEMOCAP comprises 151 two-way conversations held across five sessions, involving ten unique speakers. The testing phase is specifically allocated to the final session. Within the dataset, there exists a total of 7,433 utterances, each of which is assigned a label representing one of six emotions: happy, sad, neutral, angry, excited, and frustrated. There are a total of 12 human evaluators when annotating this dataset. 5 of them evaluated utterances across train and test datasets. Except for these 5 evaluators, 6 of them participated in the evaluation of utterances in train sets, and 1 of them participated in the evaluation of utterances in test sets. Due to the absence of a predefined validation set in the dataset, we adhere to the methodology employed in prior studies (Hazarika et al., 2018; Ghosal et al., 2019a; Hu et al., 2023) and randomly extract 10% of the training conversations in IEMOCAP as validation sets.

4.1.2 Baselines

To ensure a comprehensive evaluation of EAAN-ERC, we perform a comparative analysis, comparing our model against the following existing works:

Table 1: Statistics of the dataset.

	Train	Test	Total
# Utterances	5810	1623	7433
# Conversations	120	31	151
# Evaluators(∩+others)	11(5+6)	6(5+1)	12
# Classes		6	

Table 2: Comparison of our EAAN-ERC against various baselines.

Methods	w-F1.	w-Acc.
bc-LSTM*	62.84	63.08
DialogueRNN*	64.65	64.85
DialogueGCN*	62.11	62.49
TODKAT*	61.33	61.11
CauAIN*	65.01	65.08
CoG-BART*	64.87	65.02
SGED+DAG*	66.27	66.29
DAG-ERC*	66.53	66.54
SPCL*	66.93	66.71
COSMIC	66.22	66.25
DialogXL	65.88	65.78
HAAN-ERC	66.36	66.5
DialogueCRN	68.49	67.63
MPLP	64.89	64.92
DualGAT	65.41	65.57
SACL-LSTM	68.72	68.63
EAAN-ERC (Ours)	**69.75**	**69.83**

bc-LSTM (Poria et al., 2017) employs an utterance-level LSTM to capture contextual features. **DialogueRNN (Majumder et al., 2019)** uses three GRUs to track the speaker's state, context, and emotion during a conversation. **DialogueGCN (Ghosal et al., 2019b)** uses a directed graph to represent conversational context. **TODKAT (Zhu et al., 2021)** integrates commonsense knowledge and a task for detecting topics. **CauAIN (Zhao et al., 2022)** uses causal clues in commonsense knowledge to enrich the modeling of speaker dependencies. **CoG-BART (Li et al., 2022)** uses a response generation task to enhance BART(Lewis et al., 2020b)'s ability. **SGED+DAG (Bao et al., 2022)** is a speaker-guided framework with a one-layer DAG that can explore complex speaker interactions. **DAG-ERC (Shen et al., 2021b)** builds a directed acyclic graph from the conversation to capture its underlying structure. **SPCL (Song et al., 2022)** designs a supervised prototypical contrastive learning loss to tackle imbalanced classification and employs a difficulty-measure function for curriculum learning to handle extreme samples. **COSMIC (Ghosal et al., 2020)** utilizes GRUs to learn interactions between participants and different aspects of commonsense knowledge. **DialogXL (Shen et al., 2021a)** improves XLNet by incorporating better memory and dialog-aware self-attention. **HAAN-ERC(Zhang et al., 2023b):** leverages dialogue context information to model intra-speaker, inter-speaker, intra-modal, and inter-modal influences based on the Transformer. **DialogueCRN (Hu et al., 2021)** utilizes LSTMs to extract emotional cues and reason over multiple turns. **MPLP (Zhang et al., 2023c)** mimics the thinking process of a human being. **DualGAT (Zhang et al., 2023a)** combines discourse structure and speaker-aware context. **SACL-LSTM (Hu et al., 2023)** design a contextual adversarial training strategy to learn more diverse features from context.

4.1.3 Settings

We adopt the end-to-end manner to train EAAN-ERC. The batch size is set to 2. We use Adam (Kingma and Ba, 2014) optimizer to train our model. We set the learning rate as $1e-4$ and the number of epochs as 100. The ERC backbone is the same as (Hu et al., 2023). we run five random seeds and report the average result of the test sets. The key hyper-parameter α is tried in the set {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0} (see Appendix B.1). The codes are implemented in PyTorch[1].

4.2 Model Comparison

We conducted a comparative study to evaluate the effectiveness of our EAAN-ERC on the IEMOCAP dataset. We used Weighted F1-score (w-F1.) and Weighted Accuracy (w-Acc.) as evaluation metrics. The results of our experiment can be found in Table 2. * means the results are from (Hu et al., 2023). In each group, the better-performing method that passed the significant hypothesis test (p-value less than 0.05) is marked in bold. In Table 2, we can observe that our proposed method, EAAN-ERC, outperforms the current state-of-the-art baselines on all metrics. This indicates that, in contrast to existing baselines that do not account for the influence of different evaluator panels, EAAN-ERC effectively addresses the impact of diverse evaluator panels on emotion evaluation. Our method implements the idea of imitating the evaluation and annotation process of emotions by diverse evaluator panels, which helps enhance the performance of emotion inference in the ERC dataset with di-

[1]Our original codes will be released on GitHub upon acceptance.

verse evaluator panels. Overall, these significant comparison results demonstrate the efficacy of our proposed method.

4.3 Ablation Study

In this ablation experiment, we aim to verify the importance of Expert $Proxy$, Expert-specific loss, Expert Assignment mechanism, and Expert Confidence. To verify the importance of these components, we remove them one at a time to evaluate their impacts in terms of w-Acc. and w-F1. on the IEMOCAP dataset. The ablation experiment results are shown in Table 3. We can see that when each of the above components is removed, the model's scores on the w-F1. and w-Acc. metrics are reduced to varying degrees. In particular, the effects on the model's performance from large to small are Expert $Proxy$, Expert-specific Loss, Expert Assignment, and Expert Confidence. When the Expert $Proxy$ is removed, EAAN-ERC cannot assign the proxy model to represent the unseen evaluator, which biases the final evaluation results. When Expert-specific Loss is removed, although the models in the expert pool can represent the evaluators, they cannot learn the evaluation experience of the corresponding evaluators, thus the performance of the EAAN-ERC decreases when performing evaluator-specific emotional evaluation. When Expert Assignment is removed, all expert models in the expert pool are assigned to participate in emotion evaluation, which also brings a certain bias to the final evaluation results. We observe that the Expert Assignment has a relatively small impact on model performance. This may be due to the more comprehensive emotional representation extracted by more expert models, despite being perturbed by irrelevant evaluators, which causes a relatively small reduction in model performance. Finally, when Expert Confidence is removed, the performance of the model decreases minimally, which means that calculating confidence will strengthen to a certain extent the overall assessment of agree-upon emotions. In summary, through this ablation experiment, we verified how important these components are to the model performance.

Table 3: Ablation study of four components in our EAAN-ERC on the IEMOCAP dataset.

	w-F1.	w-Acc.
EAAN-ERC	**69.75**	**69.83**
w/o Expert Proxy	68.32	68.11
w/o Expert-specific Loss	68.89	69.32
w/o Expert Assignment	69.45	69.25
w/o Expert Confidence	69.53	69.39

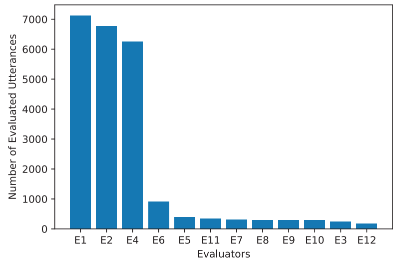

Figure 4: The number of evaluated utterances by evaluators.

4.4 Impact of the threshold M

We then analyze the impact of threshold M on model performance. In the IEMOCAP dataset, the number of utterances evaluated by each evaluator (sorted from largest to smallest) is shown in Figure 4. The impact of threshold M on model performance is shown in Figure 5. For example, when M is 3, the first three evaluators (E1, E2, and E4) in Figure 4 are filtered, which will be set up corresponding expert models to represent and participate in the subsequent emotion evaluation. From Figure 5, we can find that when M increases from 1 to 3, the performance of the model improves. This is affected by the number of evaluators. When M is set to 4, the performance of the model drops sharply. This is because the number of utterances evaluated by evaluator E6 is too small, causing the corresponding expert model to be unable to learn its evaluation experience, resulting in significant evaluation errors. When the evaluators increase from 4 to 12, the performance of the model has a significant rising stage in the early stage. The underlying reason is that as the number of expert models increases, the extracted emotion representation is more comprehensive, which to a certain extent makes up for the errors caused by expert models in learning evaluators' experiences. In the later stage, the performance of the model decreases again. The potential reason is that the gain brought by the number of expert models is less than the disturbance caused by the expert models in learning evaluators' experiments. Overall, the model performs best when M is set to 3. Through this section, we know how the threshold M affects the model's performance.

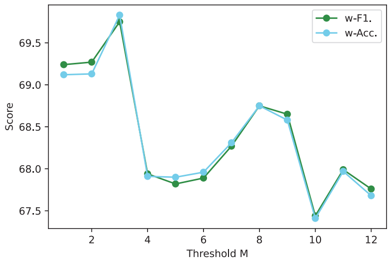

Figure 5: Impact of the threshold M on model performance.

4.5 Case Study

To further explain how our EAAN-ERC works, we visualize a case from IEMOCAP as shown in Figure 6. This case illustrates how our EAAN-ERC imitates human evaluators to evaluate an utterance. We observed from this case that our EAAN-ERC predicted the same label as each evaluator. For instance, EAAN-ERC simulates E1, E2, and E4 to predict emotion labels as ”Sad”, ”Neutral”, and ”Neutral”, respectively. Then through the weighted aggregation of the predicted emotion distribution, we obtained a final prediction result ”Neutral” consistent with the label after the evaluators agreed. In particular, for the weights, since the evaluator E1 made an evaluation contrary to the agreed emotion, it is given the smallest weight by EAAN-ERC when aggregation. On the contrary, the emotion probability corresponding to the evaluator E4 is given the greatest weight when aggregating, and that of E2 is in the middle. In this way, our model EAAN-ERC can simulate the evaluation process of human evaluators to obtain evaluator-specific emotion evaluation results.

5 Conclusion

For more accurate ERC, we propose a new method called Expert Adaptive Agreement Network for Emotion Recognition in Conversation (EAAN-ERC) for evaluator panel-specific emotion identification. Our method imitates the process of evaluating and annotating emotions by diverse evaluator panels. Specifically, we use multiple expert models to mimic experienced evaluators and adaptively mix them to emulate diverse evaluator panels. We also calculate the emotional probability and confidence of each assessed utterance to simulate the evaluator panels' emotional evaluation. Finally, we integrate the emotional probability with confidence to mimic the agreement of an evaluator panel. Extensive experiments on the widely used ERC dataset IEMOCAP demonstrate the effectiveness of our EAAN-ERC.

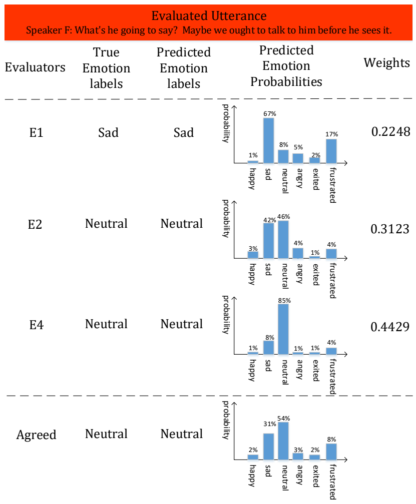

Figure 6: A case demonstrating how our EAAN-ERC imitates human evaluators' annotation process.

Limitations

In this paper, for evaluator panel-specific emotion identification, we propose a new approach called Expert Adaptive Agreement Network for Emotion Recognition in Conversation (EAAN-ERC), which imitates the process of evaluating and annotating emotions by diverse evaluator panels. Despite that our proposed method can effectively improve the performance of ERC on the IEMOCAP dataset, it is suitable for situations where there are a small number of human evaluators (such as there are 12 evaluators in IEMOCAP), and cannot be directly applied to scenarios where there are many human evaluators (such as crowdsourcing). To address this issue, clustering a large number of human evaluators is feasible so that EAAN-ERC can be applied to the above scenario. Annotation work on datasets with very large human evaluators needs to be done and made public in the future, and how to effectively cluster human evaluators also needs to be further explored in the future. These limitations will be left for future research.

The method in this article utilizes the hard label of the emotion evaluated by the evaluator to supervise the training of the corresponding expert model, which may cause the randomness in the emotion evaluation from the same evaluator to be ignored. It is a potential solution to establish soft labels for each evaluator's emotional evaluation during the dataset annotation process to introduce randomness and guide the expert model to learn the emotional evaluation distribution. Moreover, there is also some randomness in the assignment of evaluators. The approach in this paper refers to the proxy expert model to represent rare evaluators to address this issue. However, when there is no evaluator identity represented by a designated expert in the test sample, the method in this article will degenerate into an ERC backbone model in which evaluator information is not utilized. Better ways to represent rare evaluators need to be further explored in the future. Evaluator subjective similarity calculation may be a solution, which enables rare evaluators to be represented by existing expert models corresponding to common evaluators with high subjective similarity. How to design the evaluator's subjective similarity calculation method is also one of the issues that need to be solved in the future.

References

Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, and Yushi Aono. 2018. Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4964–4968. IEEE.

Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, and Yushi Aono. 2019. Speech emotion recognition based on multi-label emotion existence model. In *INTERSPEECH*, pages 2818–2822.

Mia Atcheson, Vidhyasaharan Sethu, and Julien Epps. 2019. Using gaussian processes with lstm neural networks to predict continuous-time, dimensional emotion in ambiguous speech. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 718–724. IEEE.

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Speaker-guided encoder-decoder framework for emotion recognition in conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4051–4057. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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(4):335–359.

Huang-Cheng Chou, Chi-Chun Lee, and Carlos Busso. 2022. Exploiting co-occurrence frequency of emotions in perceptual evaluations to train a speech emotion classifier. *Interspeech 2022*.

Ting Dang, Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah. 2017. An investigation of emotion prediction uncertainty using gaussian mixture regression. In *INTERSPEECH*, pages 1248–1252.

Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *2016 international joint conference on neural networks (IJCNN)*, pages 566–570. IEEE.

Deepanway Ghosal, Navonil Majumder, Alexander F. 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, Online Event, 16-20 November 2020*, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019a. 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. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019b. 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. Association for Computational Linguistics.

Alex Graves. 2014. Generating sequences with recurrent neural networks.

Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 890–897.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10835–10852. 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.

D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *Computer Science*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bryan Li, Dimitrios Dimitriadis, and Andreas Stolcke. 2019. Acoustic and lexical sentiment analysis for customer service calls. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5876–5880.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11002–11010.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. 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 3469–3483, Online. Association for Computational Linguistics.

Wenzhe Lou, Wenzhong Yang, and Fuyuan Wei. 2023. Dialogcin: Contextual inference networks for emotional dialogue generation. *Applied Sciences*, 13(15).

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Lisong Qiu, Yingwai Shiu, Pingping Lin, Ruihua Song, Yue Liu, Dongyan Zhao, and Rui Yan. 2020. What if bots feel moods? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1161–1170.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. 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.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021c. 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. Association for Computational Linguistics.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Pro-*

ceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kusha Sridhar and Carlos Busso. 2020. Modeling uncertainty in predicting emotional attributes from spontaneous speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8384–8388. IEEE.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wen Wu, Chao Zhang, and Philip Woodland. 2023a. Estimating the uncertainty in emotion attributes using deep evidential regression. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15681–15695, Toronto, Canada. Association for Computational Linguistics.

Wen Wu, Chao Zhang, and Philip C Woodland. 2023b. Distribution-based emotion recognition in conversation. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 860–867. IEEE.

Wen Wu, Chao Zhang, Xixin Wu, and Philip C Woodland. 2022. Estimating the uncertainty in emotion class labels with utterance-specific dirichlet priors. *IEEE Transactions on Affective Computing*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023a. 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. Association for Computational Linguistics.

Tao Zhang, Zhenhua Tan, and Xiaoer Wu. 2023b. Haanerc: hierarchical adaptive attention network for multimodal emotion recognition in conversation. *Neural Computing and Applications*, pages 1–14.

Ting Zhang, Zhuang Chen, Ming Zhong, and Tieyun Qian. 2023c. Mimicking the thinking process for emotion recognition in conversation with prompts and paraphrasing. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6299–6307. ijcai.org.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4524–4530. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. 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 1571–1582, Online. Association for Computational Linguistics.

A Key Hyperparameter α

The impact of the key hyperparameter α is shown in Figure 7. We can see that as the hyperparameter increases, the performance of the model generally gradually increases and stabilizes later, and achieves the best performance when α is equal to 0.8. This shows that the expert model's learning of each evaluator's evaluation experience will positively affect the emotion inference on IEMOCAP.

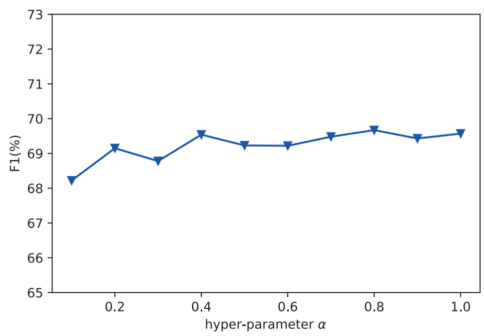

Figure 7: The impact of the key hyperparameter α on the performance of EAAN-ERC.

B Comparison of our EAAN-ERC against Distribution-based ERC

To tackle the inherent ambiguity of emotions and the subjective nature of human perception in ERC, Wu et al. (Wu et al., 2023b) suggest a distribution-based ERC method. This approach involves incorporating Bayesian training loss by linking each emotional state to a specific Dirichlet prior distribution based on the utterance. In contrast to the distribution-based approach, our method focuses on leveraging evaluator identity information to address the subjectivity of human perception in ERC. Our approach aims to emulate diverse human evaluator panels, offering a distinctive perspective on addressing this challenge.

B.1 Experimental Setting

To further demonstrate the effectiveness of our approach, we compared our EAAN-ERC with the approach in (Wu et al., 2023b) (we called it "Distribution-based ERC"). We follow Distribution-based ERC's 4-way emotion evaluation experimental setup, where leave-one-session-out 5-fold cross-validation (5-fold CV) was performed and the average results are reported. Also same as Distribution-based ERC, weighted Accuracy (w-Acc.) and unweighted Accuracy (u-Acc.) are used as evaluation metrics for 4 categories.

In this experiment, the label ”Frustrated” is set to -1 to exclude it from training and testing. All labels of ”Excited” are changed to ”Happy”. The batch size is set to 2. We use the Adam(Kingma and Ba, 2014) optimizer to train the model. The learning rate is 1e-4, and epochs are set to 100. Early stopping is performed when the valid set performance does not improve for 20 consecutive epochs. The experiments were conducted on A100 and the code was implemented in PyTorch. Relevant code and checkpoints will be made public on Github after acceptance.

The experimental results are shown in Table 4. We can see from this table that our method significantly outperforms Distribution-based ERC by 10.05% and 10.46% on w-Acc and u-Acc respectively. This shows that our evaluator identity information-based EAAN-ERC effectively addresses the subjectivity of human perception of ERC from the perspective of the imitation of diverse human evaluator panels. This promising experimental result further demonstrates the effectiveness of our method.

C Model Complexity and Computational Efficiency

The parameter size of our EAAN-ERC is 12M. In our experiments on A100 (training consumes about 9G memory), each epoch training consumes about 8.8 seconds, and it takes about 15 minutes to complete a training task (i.e. 100 epochs in our experiments). We subjectively consider that the training resources and time consumption are acceptable. Each epoch inference consumes about 5.8s. There are 151 dialogues in total, the average length of each dialogue is about 50 turns, and the time required to infer the conversation context of 50 turns is about 38ms. We subjectively consider that the inference speed is also acceptable in real-life applications.

Table 4: Comparison of our EAAN-ERC with the Distribution-based ERC(Wu et al., 2023b). Leave-one-session-out 5-fold cross-validation (5-fold CV) was performed and the average results are reported.

	w-Acc.	u-Acc.
Distribution-based ERC	77.83	78.12
EAAN-ERC	**87.88**	**88.58**