Who You Are, What You Say: Intra- and Inter-Context Personality for Emotion Recognition in Conversation

Tazeek Bin Abdur Rakib[*], **Lay-Ki Soon, Wern Han Lim**
Monash University, Malaysia
{tazeek.binabdurrakib, soon.layki, lim.wern.han}@monash.edu

Abstract

Emotion recognition in conversation (ERC) requires understanding both contextual dependencies and speaker-specific cues. Existing approaches often treat conversation context as a single representation or encode speaker identity shallowly, limiting their ability to capture fine-grained emotional dynamics. We propose PERC, a personality-aware ERC framework that (1) segregates conversational context into intra- and inter-speaker components, (2) models static or dynamic personality traits to represent stable and evolving speaker dispositions, and (3) performs contrastive cross-alignment between intra–intra and inter–inter representations to enforce contextual and personality consistency. Experiments on three ERC benchmarks show that PERC achieves new state-of-the-art performance, improving weighted F1 by up to 2.74% over non-LLM methods and 0.98% over recent LLM-based methods. Our results demonstrate the effectiveness of integrating context segregation, personality modeling, and contrastive alignment for emotion reasoning in dialogue.

1 Introduction

Identifying the emotions of different speakers in a conversational setting plays a crucial role in developing empathetic and human-centered chatbots (Majumder et al., 2020; Rakib et al., 2025). Unlike sentiment analysis, which generally categorizes expressions as positive, negative, or neutral, emotion recognition captures a more fine-grained understanding of a speaker's affective state (Ekman, 1971, 1992). Moreover, Emotion Recognition in Conversation (ERC) is shaped not only by a speaker's internal feelings but also by the surrounding conversational context and the personalities of other participants. Incorporating speaker personality information can enhance emotion understanding and improve user engagement in emotion-driven applications such as psychotherapy (Althoff et al., 2016; Pérez-Rosas et al., 2017) and customer support (Lin et al., 2019; Rashkin et al., 2018).

Current state-of-the-art (SOTA) approaches in ERC primarily focus on modeling speaker dynamics within dialogues. These methods often capture the intra- and inter-relationships between speakers and utterances using graph neural networks (GNNs) (Wu et al., 2020), where speakers and utterances are represented as nodes and edges, respectively. Intra-speaker features connect a speaker to their own historical utterances, while inter-speaker features capture the relationships between the target speaker and other participants in the dialogue. Alternatively, some approaches encode speaker information into pre-trained language models through attention-based mechanisms (Jiao et al., 2020) or memory modules (Majumder et al., 2019). Furthermore, these models are often enhanced by integrating external knowledge sources, such as commonsense reasoning, to better capture the psychological and contextual aspects of speakers (Li et al., 2021). However, very limited work has explicitly explored personality-specific representations in ERC. To the best of our knowledge, only ERC-DP (Wang et al., 2024) incorporates personality information, predicting a speaker's personality traits using five independent classifiers and subsequently fusing these traits with the speaker's past, present, and future utterances to predict the target emotion. This rarity highlights a significant research gap: personality traits are a key factor influencing emotional expression among speakers, yet remain largely underexplored in current ERC studies.

Despite these advances, the utilization of speaker features in ERC remains limited and often dependent on external knowledge bases. First, existing approaches do not fully leverage speaker information, treating it either superficially in GNNs via intra- and inter-speaker edges or merely as auxil-

[*] Corresponding author

Findings of the Association for Computational Linguistics: EACL 2026, pages 3293–3308
March 24-29, 2026 ©2026 Association for Computational Linguistics

iary input in other models. Second, reliance on commonsense reasoning introduces dependency on external knowledge sources, which may require re-training as such knowledge bases can become outdated and fail to capture the diversity of conversational scenarios. Third, context representations are often diluted, as they aggregate all historical utterances indiscriminately, introducing noise due to the involvement of multiple speakers. Lastly, the ERC-DP model incorporates future utterances to predict the emotion of a current utterance, resulting in hindsight bias and limiting applicability in real-time systems. While speaker and conversational context are leveraged, intra- and inter-speaker features can be further enriched with their corresponding contextual representations, enabling more fine-grained features that have the potential to significantly improve emotion prediction.

Motivated by the gaps and opportunities in exploring personality features for speakers, we introduce PERC (Personality-Enhanced ERC). We first segregate the dialogue context into intra- and inter-speaker components. Next, we perform two forms of feature extraction for the respective intra- and inter-components: personality trait feature extraction for capturing speaker personalities, and dialogue context extraction for modeling the segregated conversational context. The personality feature extraction considers both static (stable) and dynamic (evolving) personalities, using a pre-trained personality extractor model. Finally, we perform contrastive cross-alignment of the intra-features (intra-context and intra-personality) and inter-features (inter-context and inter-personality). The purpose of this alignment is grounded in psychology: a speaker's personality should be closely aligned with the context in which they speak (Morency, 2013).

In summary, the contributions of this paper are as follows:

- Incorporate personality traits for ERC, comparing both static and dynamic personality representations. This involves extracting personality features via a pre-trained model and updating speaker information according to static or dynamic personality.

- Introduce segmentation of conversational context into intra- and inter-speaker components. This improves the modeling of conversational dynamics and enriches the corresponding features, departing from the conventional approach of using the overall context as a single feature representation.

- Perform contrastive cross-alignment between intra-intra and inter-inter pairings, which pulls positive pairs together and pushes negative pairs apart, enhancing the model's decision-making boundaries.

- Evaluate our approach on three benchmark ERC datasets. Results show that our method outperforms existing state-of-the-art approaches, demonstrating the effectiveness of integrating personality traits and fine-grained contextual features.

2 Related Work

2.1 Emotion Recognition in Conversation

Most approaches for ERC can be broadly categorized into several methodological families. **Sequential-based models** such as DialogueRNN (Majumder et al., 2019) and HiGRU (Jiao et al., 2019) capture contextual dependencies across utterances using recurrent structures like Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs) (Hu et al., 2023). These models may also adopt hierarchical memories to preserve long-term dependencies (Jiao et al., 2020) or employ dual-stream mechanisms to attend to both local and global contexts (Li et al., 2024b).

Graph-based models represents conversations as relational structures. Graph Neural Networks (GNNs) (Wu et al., 2020) are commonly used to construct inter- and intra-speaker edges (Ghosal et al., 2019; Ishiwatari et al., 2020; Bao et al., 2022) and to model causal information flow through directed acyclic graphs (DAGs) (Shen et al., 2021b). Beyond structural modeling, psychological and commonsense knowledge can be integrated through external knowledge bases such as ATOMIC (Sap et al., 2019), enabling richer inference of interpersonal and causal dynamics (Li et al., 2021; Hu et al., 2021).

Pre-trained Language Models (PLMs) such as XLNet and RoBERTa have also been leveraged to enhance contextual understanding in ERC (Yang, 2019; Li et al., 2020; Kim and Vossen, 2021). The pre-trained memory of PLMs can serve as implicit contextual knowledge (Lee and Lee, 2021), or be further enriched with external commonsense events (Ghosal et al., 2020; Zhu et al., 2021). Moreover, PLMs have been combined with prompt-tuning strategies to mitigate data imbalance and improve generalization (Gao et al., 2024).

Finally, recent efforts have explored **training strategies** that enhance ERC robustness, such as curriculum learning (Yang et al., 2022; Li et al., 2024a) for dynamic difficulty scheduling, and contrastive learning (Li et al., 2022; Kang and Cho, 2024) to disentangle emotion representations and promote discriminative features.

2.2 Speaker Modeling

In terms of utilizing speaker features for ERC, several approaches have been proposed. For example, speaker features are often stored as individual GRU states (Majumder et al., 2019), or encoded through speaker-aware or dependency-based encoders to enhance context differentiation. Some methods further employ auxiliary tasks such as speaker identification (Li et al., 2020) or incorporate speaker tokens (Kim and Vossen, 2021) into the input sequence to make the model speaker-sensitive.

Graph-based approaches (Shen et al., 2021b), on the other hand, explicitly model inter- and intra-speaker relationships by connecting utterances through corresponding speaker nodes (Ghosal et al., 2019; Shen et al., 2021b). These graphs can be further enriched with psychological or commonsense-based interactions (Li et al., 2021), where speaker features are inferred from mental state reasoning models such as COMET (Bosselut et al., 2019). Such approaches enable implicit reasoning about speaker intentions and reactions.

However, several issues remain. Most methods treat speaker features as categorical entities, limiting the ability to capture deeper and dynamic interactions between speakers. Knowledge-based reasoning approaches are constrained by the coverage and timeliness of their external resources, which may become outdated or incomplete. Lastly, intra- and inter-speaker representations are often shallow, reflecting only surface-level utterance dependencies without modeling the underlying personality traits or affective tendencies of the speakers.

2.3 Personality Traits & Emotions

Personality traits represent enduring patterns of thinking, feeling, and behaving that influence how individuals express and regulate emotions in social interactions (Costa Jr and McCrae, 1992; Barańczuk, 2019). Among existing psychological models, the Big Five (Gosling et al., 2003) framework remains the most widely adopted due to its empirical robustness, cross-cultural generalizability, and continuous representation of personality dimensions—unlike categorical typologies such as MBTI (Furnham, 1996).

Although personality traits are considered stable and static, recent studies show that personality states can vary with situational and contextual factors (Fleeson, 2007; Beckmann and Wood, 2017). This underscores the need to model personality as context-sensitive rather than fixed, especially in conversational settings where behavioral cues dynamically evolve (Jung et al., 2014). Capturing personality as it manifests in context has been shown to improve behavioral and emotional inference (Lewis, 1999). Psychological evidence also highlights strong correlations between personality and affective tendencies: for instance, neuroticism aligns with negative emotions, while extraversion relates to positive affect (Larsen and Ketelaar, 1991; Gross, 1998).

These findings suggest that personality provides a reliable prior for emotional prediction. In our work, we introduce personality-aware representations (Baumert et al., 2017) that guide both intra- and inter-context modeling. Personality vectors act as anchors that stabilize each speaker's emotional evolution (intra-context) and regulate cross-speaker affective influence (inter-context). Through cross-alignment between these spaces, our framework jointly enforces self-consistency and relational coherence, thus enabling emotion recognition that is both person-specific and interaction-aware.

3 Methodology

3.1 Problem Overview

A conversation C can be represented as a sequence of N utterances, $C = (u_1, u_2, \ldots, u_N)$. The goal of ERC is to predict an emotion label for each utterance in the conversation sequence. To be precise, ERC seeks to learn a mapping from $(u_1, u_2, \ldots, u_N)$ to $Y = (y_1, y_2, \ldots, y_N)$, where each $y_i \in \mathcal{Y}$ and $\mathcal{Y}$ is the predefined set of emotion categories based on the given dataset.

3.2 Architecture Breakdown

The PERC model takes a queried utterance as input and predicts its respective emotion label through several main components as shown in Figure 1.

First, the intra- and inter-context segregation (Section 3.3) separates the utterances spoken by the queried speaker from those spoken by others.

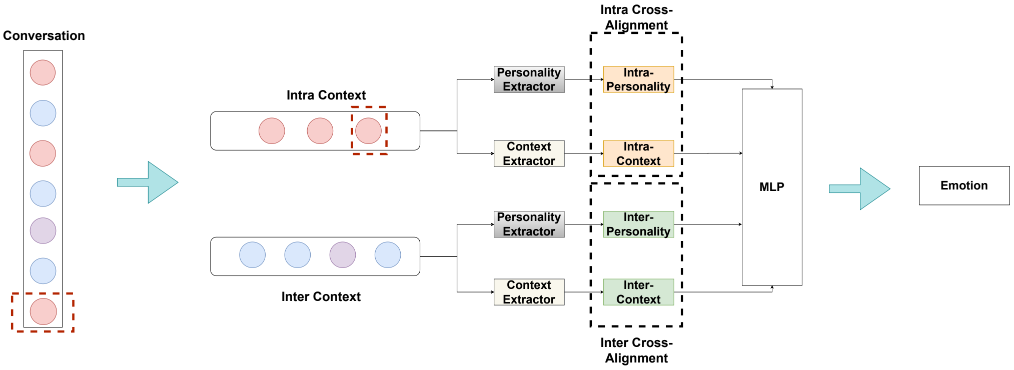

Figure 1: PERC pipeline. The red dotted box highlights the queried utterance on the emotion to predict. The black dotted box highlights the cross-alignment of the personality and context pairs.

Then, the personality feature extraction is done for the intra- and inter-personality representations from a pretrained personality model (Section 3.4). Next, the dialogue context encoding captures semantic features of the conversation using a fine-tuned RoBERTa encoder (Section 3.5). Contrastive cross-alignment is then performed that aligns personality and context features across intra- and inter-views to encourage consistency (Section 3.6). Finally, classification is performed, with losses computed for both classification and alignment quality (Section 3.7). A running example is shown in Appendix C.

3.3 Intra- and Inter-Context Segregation

In a conversation with two or more speakers, the utterances available at turn t for the queried speaker s can be categorized as:

- **Intra-context:** utterances spoken by s up to t.
- **Inter-context:** utterances spoken by *all other speakers* up to t.

Current ERC models often mix all utterances together, which can obscure speaker-specific patterns and reduce performance (Majumder et al., 2019; Ghosal et al., 2020; Shen et al., 2021a). By segregating the context, we preserve the nuances of the queried speaker's behavior and the influence of surrounding speakers, enabling the model to capture self-directed and externally influenced conversational dynamics more effectively (Nasello et al., 2023; Barrett, 2022; Wynn et al., 2024).

To limit context length, we use a sliding window of size n up to the current utterance u_t. The intra-context for speaker s is defined as:

$$u_{\text{intra}}^{(s)} = \left(u_{t-n}^{(s)}, u_{t-n+1}^{(s)}, \dots, u_t^{(s)}\right), \qquad (1)$$

The inter-context aggregates utterances from all other speakers ($\bar{s} \neq s$) up to turn t:

$$u_{\text{inter}} = \left(u_j^{\bar{s}} \mid j \in [t-n, t],\ \bar{s} \neq s\right), \qquad (2)$$

where $u_j^{\bar{s}}$ is the j-th utterance from any non-queried speaker. u_{inter} and u_{intra} are then used for personality encoding and conversational context encoding. An example of the segmentation is shown in Table 4.

3.4 Personality Feature Extraction

To capture personality characteristics from the intra- and inter-context segments, we employ a dedicated personality model. This model is trained separately because standard ERC datasets do not provide explicit personality annotations (Poria et al., 2018; Busso et al., 2008).

3.4.1 Personality Model

To obtain personality features for downstream ERC, we train a personality recognition model using the Essays dataset (Pennebaker and King, 1999), which has been shown to provide meaningful personality features for ERC (Wang et al., 2024). This dataset is based on the Big Five personality model and is annotated by psychology students, with each sample assigned a binary label (0 or 1) indicating the presence or absence of each trait. Accordingly, personality recognition is formulated as a multi-label classification task.

We improve on the BERT-MLP (Wang et al., 2024) design in two ways: (1) we fine-tune the BERT encoder (Devlin et al., 2018) on the Essays dataset to better adapt to personality detection, and

(2) we replace multiple MLPs with a single BiLSTM layer that jointly classifies all five traits. This setup enhances modeling of sequential dependencies and reduces the overhead of training multiple classifiers. The resulting personality extractor is trained only once on the Essays dataset and remains frozen during ERC training. The same extractor is applied across all three ERC datasets without any further fine-tuning. The BERT-BiLSTM implementation is detailed in Appendix A.

3.4.2 Personality Feature Extraction

We employ the newly trained BERT-BiLSTM model to extract personality features for both intra- and inter-speaker segments. This yields two types of embeddings: the **intra-personality** vector z^p_{intra}, representing the queried speaker and their own utterances, and the **inter-personality** vector z^p_{inter}, representing personality cues derived from the surrounding conversational context.

Each segment is first tokenized using the BERT tokenizer and encoded with BERT to obtain the [CLS] embedding:

$$H_{intra} = \text{BERT}(u_{intra})[CLS] \tag{3}$$

The embedding H_{intra} is then passed through a BiLSTM, and the final hidden state is taken as the personality feature vector:

$$z^p_{intra} = \text{BiLSTM}(H_{intra}) \tag{4}$$

The same process is applied to obtain z^p_{inter}.

We further explore two hypotheses about how personality manifests in conversation: **static** and **dynamic**.

In the **static** setting, a speaker's personality is assumed to remain constant throughout the dialogue, independent of context (Fleeson, 2007). This is only applicable to intra-personality representations, since inter-personality aggregates cues across multiple speakers.

In contrast, the **dynamic** setting assumes that personality may evolve as the conversation progresses and topics shift (Beckmann and Wood, 2017). Here, personality features are re-extracted at every turn, allowing PERC to adapt to contextual changes. We provide details on the update mechanism for z^p_{intra} and z^p_{inter} in Appendix B.

3.5 Conversation Context Encoding

We extract conversational context representations for both intra- and inter-speaker segments introduced in Section 3.3. Explicitly encoding these segments with a fine-tuned model allows the model to capture distinct signals from the queried speaker and surrounding speakers, avoiding dilution of feature states.

Each utterance sequence is tokenized and passed through a RoBERTa encoder (Liu et al., 2019) fine-tuned on the ERC dataset. We use the final CLS hidden state as the context representation:

$$z^c_{\text{intra}} = \text{RoBERTa}\big(u_{\text{intra}}\big)[CLS], \tag{5}$$

$$z^c_{\text{inter}} = \text{RoBERTa}\big(u_{\text{inter}}\big)[CLS]. \tag{6}$$

Using the CLS representation captures global sequence-level semantics, which is more suitable for modeling conversational context than averaging token embeddings, as the overall meaning of the context is critical. An example of this is shown in Table 4.

3.6 Contrastive Cross-Alignment

After obtaining the personality and context representations for both intra- and inter-speaker segments, we perform **contrastive cross-alignment** to integrate these two components. The underlying hypothesis is that personality and conversational context are inherently related in how speakers express themselves (intra) and interact with others (inter). Aligning these representations encourages PERC to learn consistent and discriminative features for emotion recognition (Leikas et al., 2012).

We adopt the InfoNCE objective (Oord et al., 2018) for contrastive learning. Specifically, the intra-personality feature z^p_{intra} and intra-context feature z^c_{intra} form a positive pair, while inter-personality z^p_{inter} and inter-context z^c_{inter} act as negative samples. The reverse pairing (inter as positive, intra as negative) is also applied to balance both views.

This encourages alignment between personality and context features from the same view while pushing apart mismatched pairs. The total alignment loss $\mathcal{L}_{\text{align}}$ is defined as:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}, \tag{7}$$

where $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{inter}}$ denote the intra- and inter-view contrastive losses, respectively (details in Appendix D).

Through cross-alignment, PERC learns to associate coherent personality–context signals within each conversational view and to separate incongruent ones, leading to more robust emotional representations and improved classification performance.

3.7 Classification

For prediction, we concatenate the four representations from intra- and inter-pairs into a single feature vector:

$$z = \left[z^{p}_{\text{intra}} \oplus z^{p}_{\text{inter}} \oplus z^{c}_{\text{intra}} \oplus z^{c}_{\text{inter}}\right]. \tag{8}$$

This combined representation z is passed through a single-layer MLP with ReLU activation, followed by a linear transformation to produce the emotion logits:

$$\hat{y}_t = W(ReLU(\text{MLP}(z)) + b, \tag{9}$$

where $\hat{y}_t$ are the unnormalized scores over emotion classes at timestep t. The predicted probability distribution is obtained via the softmax function and the final predicted emotion label is obtained:

$$p_t = arg\max_y(softmax(\hat{y}_t)), \tag{10}$$

Loss Function. We employ a joint objective consisting of a focal loss for classification and the contrastive cross-alignment loss defined in Section 3.6.

The focal loss (Lin et al., 2017) for a given instance with ground-truth label y is formulated as:

$$p_y = p_t[y] = \text{softmax}(\hat{y}_t)_y \tag{11}$$

$$\mathcal{L}_{\text{focal}} = -\alpha_y\,(1 - p_y)^{\gamma}\,\log(p_y), \tag{12}$$

where α_y is a class-balancing weight, and γ is the focusing parameter that reduces the loss contribution from easy examples.

Finally, the overall training objective combines the classification and contrastive losses:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda\,\mathcal{L}_{\text{align}}, \tag{13}$$

where λ is a hyperparameter controlling the trade-off between classification accuracy and representation alignment.

4 Experimental Setup

4.1 Datasets

As our framework is composed of two distinct components, we utilize the following datasets:

Personality The Essays dataset (Pennebaker and King, 1999) is used for personality prediction. It is composed of text samples, where each sample is annotated with the author's Big Five personality traits. The labels are binary, indicating whether the author exhibits (1) or does not exhibit (0) a given personality trait.

ERC Following (Wang et al., 2024), we use three benchmark ERC datasets: MELD (Poria et al., 2018), EmoryNLP (Zahiri and Choi, 2018), and IEMOCAP (Busso et al., 2008). MELD and EmoryNLP are based on the Friends TV series and involve multi-party dialogues (more than two speakers). In contrast, IEMOCAP contains dyadic interactions in actor-driven settings. Following (Ishiwatari et al., 2020; Shen et al., 2021a), we adopt the weighted-F1 score as the evaluation metric for all experiments in this paper.

The detailed statistics of the datasets, along with the hyperparameter settings, are provided in Appendix E.

4.2 SOTA Baselines

We compare our approach against a set of SOTA baselines, including both non-LLM and LLM-based methods. The non-LLM baselines comprise DialogueRNN (Majumder et al., 2019), DialogueGCN (Ghosal et al., 2019), COSMIC (Ghosal et al., 2020), HiTrans (Li et al., 2020), AGHMN (Jiao et al., 2020), DialogueCRN (Hu et al., 2021), DAG-ERC (Shen et al., 2021b), DialogXL (Shen et al., 2021a), EmoBERTa (Kim and Vossen, 2021), CoMPM (Lee and Lee, 2021), SPCL (Song et al., 2022), SACL (Hu et al., 2023), ERNetCL (Li et al., 2024a), SGED (Bao et al., 2022), SKAIG (Li et al., 2021), CauAIN (Zhao et al., 2022), CoG-BART (Li et al., 2022), BERT-ERC (Qin et al., 2023), ERC-DP (Wang et al., 2024), CEPT (Gao et al., 2024), DualRAN (Li et al., 2024b), ML-ERC (Kang and Cho, 2025), and DERC-PL (Li et al., 2025). The LLM-based baselines include InstructERC (Lei et al., 2023), BiosERC (Xue et al., 2024), and LaERC-S (Fu et al., 2025).

Full description of the SOTA baselines are given in Appendix F. All best-performing scores are reported from their respective papers.

5 Results & Analysis

Table 1 shows the performance of our approach compared to SOTA baselines. All experiments are repeated five times with different random seeds to ensure consistency and the average reported. **Pairwise t-tests between ERC-DP and our method yield p-values below 0.05, indicating that the observed performance gains are statistically significant.**

5.1 Main Results

For IEMOCAP, which consists of dyadic, acted conversations, our method achieves **73.14%** weighted F1, representing an improvement of 1.44% over BERT-ERC (non-LLM) and 0.74% over LaERC-S (LLM). While the acted nature of IEMOCAP makes emotions relatively easier to identify, these results indicate that contrastive dynamic modeling contributes additional gains by capturing fine-grained shifts between closely related emotional states. On the MELD and EmoryNLP datasets, our approach achieves **70.25%** and **42.11%**, respectively, surpassing the previous best non-LLM results of 67.51% and 40.94%, as well as LLM-based results of 69.27% and 42.08%. The improvements are more pronounced in multi-party conversational settings, where modeling speaker individuality and long-term emotional dependencies is particularly challenging.

This shows that explicitly separating personality and conversational context provides complementary cues that context-only or graph-based methods miss. This separation enables more precise disambiguation of subtle emotions, such as frustration versus anticipation. Across all three datasets, our model achieves new state-of-the-art results. The largest gains occur in multi-speaker and nuanced dialogue scenarios (MELD and EmoryNLP). These consistent improvements highlight the effectiveness of personality modeling and the robustness of our approach across diverse conversational settings. We also provide additional experimental results in Appendix H.

5.2 Ablation Studies

Impact of personality We compare the baseline without personality features against models incorporating personality as shown in Table 2. The baseline achieves 67.41% on MELD and 68.32% on IEMOCAP. Adding static personality provides minor improvements, but its one-off representations are limited in capturing evolving conversation dynamics. In contrast, dynamic personality consistently outperforms both the baseline and static variants, achieving 70.25% on MELD and 73.14% on IEMOCAP. These results indicate that modeling personality as an evolving feature, rather than a fixed trait, better captures speaker-specific cues and enhances emotion recognition in conversation.

Sliding window effect. We analyze the impact of different sliding window sizes, as shown in Table 2. For static personality, performance decreases as the window size n increases. This is because larger windows require more utterances before extracting the personality, leading to the addition of zero-padding, which introduces noise and limits the usefulness of static features. In contrast, dynamic personality benefits from larger windows, as it continuously updates personality features at every step, leveraging the extended conversational history to capture evolving traits and emotions. These gains are particularly notable on IEMOCAP, where dyadic, acted conversations exhibit sharper emotional shifts, while MELD shows smaller but consistent improvements in multi-speaker, noisier scenarios.

Context Segregation To evaluate the importance of context segregation, we conduct four experiments: using full context, only intra-context, only inter-context, and explicitly segregated intra- and inter-contexts, all without cross-alignment. As shown in Table 3, several trends emerge. Limiting the model to only intra-context significantly decreases performance (–3.54% on MELD, –2.55% on IEMOCAP), while using only inter-context also underperforms (–1.17% and –0.87%, respectively), highlighting the necessity of both self-directed (intra) and externally influenced (inter) utterances. Furthermore, explicitly segregating and jointly modeling intra- and inter-context improves results over full context alone (+0.82% on MELD; +2.59% on IEMOCAP), demonstrating that this segregation captures complementary dynamics lost in plain context representations, particularly in dyadic dialogues such as IEMOCAP.

Cross-alignment Impact To emphasize the value of cross-aligning personality and context features, we conduct ablations with and without this component. As shown in Table 3, incorporating cross-alignment improves performance by +1.12% on MELD and +1.13% on IEMOCAP. These results indicate that enforcing consistency between speaker-centric personality traits and context-centric dialogue information helps the model capture subtle emotional cues more effec-

Approach	MELD	EmoryNLP	IEMOCAP
Non-LLM Approaches			
HiGRU (Jiao et al., 2019)	56.81	34.48	58.54
DialogRNN-RoBERTa (Majumder et al., 2019)	63.20	37.75	63.92
DialogueGCN (Ghosal et al., 2019)	58.10	-	64.18
COSMIC (Ghosal et al., 2020)	65.21	38.11	65.28
HiTrans (Li et al., 2020)	61.94	36.75	64.50
AGHMN (Jiao et al., 2020)	58.10	-	63.50
DAG-ERC (Shen et al., 2021b)	63.65	39.02	68.03
DialogueCRN (Hu et al., 2021)	63.42	38.91	66.33
SKAIG (Li et al., 2021)	65.18	38.88	66.98
DialogXL (Shen et al., 2021a)	62.41	34.73	66.20
EmoBERTa (Kim and Vossen, 2021)	66.51	-	68.57
COMPM (Lee and Lee, 2021)	66.52	38.93	69.46
SPCL (Song et al., 2022)	67.25	**40.94**	69.74
SGED (Bao et al., 2022)	65.46	40.24	68.53
CauAIN (Zhao et al., 2022)	65.46	-	67.61
CoG-BART (Li et al., 2022)	64.81	39.04	66.18
BERT-ERC (Qin et al., 2023)	67.11	39.84	**71.70**
SACL (Hu et al., 2023)	66.45	39.65	69.22
EmotionIC (Liu et al., 2024)	66.40	40.01	69.61
ERC-DP (Wang et al., 2024)	67.34	40.10	69.64
CEPT (Gao et al., 2024)	**67.51**	-	70.53
DualRAN (Li et al., 2024b)	66.24	39.22	69.73
ERNetCL (Li et al., 2024a)	66.31	39.71	69.73
ML-ERC (Kang and Cho, 2025)	65.90	37.35	68.00
DERC-PL (Li et al., 2025)	59.32	32.78	61.05
LLM Approaches			
InstructERC (Lei et al., 2023)	69.15	41.37	71.39
BiosERC (Xue et al., 2024)	68.72	41.44	69.02
LaERC-S (Fu et al., 2025)	**69.27**	**42.08**	**72.40**
PERC	**70.25** (+2.74)	**42.11** (+1.17)	**73.14** (+1.44)

Table 1: Weighted F1 comparison with state-of-the-art ERC methods. Non-LLM and LLM baselines are separated, with the best scores in each group highlighted in **bold**. PERC achieves improvements over the best non-LLM methods (indicated in parentheses) while maintaining competitive performance relative to LLM-based approaches.

tively, yielding more discriminative and psychologically grounded emotion representations across diverse conversational settings.

5.3 Limitations

While our proposed framework achieves state-of-the-art performance, several limitations remain. The contrastive cross-alignment treats intra speaker-context pairs as positive and inter pairs as negative. In reality, conversational dynamics such as empathy or mimicry may also involve alignment between speakers, which this assumption does not capture. Also, the dynamic personality modeling introduces additional computational overhead, as features must be recomputed at every conversational turn, which may reduce scalability in long or multi-party conversations. In addition, our current approach pools speaker representations via concatenation, which may lose per-speaker granularity. A natural extension is to compute personality vectors for each speaker and fuse them using cross-attention, allowing the model to learn which speakers most influence a target utterance while providing interpretable attention weights. However, this would increase computational cost and model parameters, and we leave it as a promising direction for future work.

Approach	MELD	IEMOCAP
No Personality	67.41	68.32
Static (n=2)	68.23	69.51
Static (n=3)	67.39	67.91
Static (n=4)	66.45	66.43
Dynamic (n=2)	68.56	70.87
Dynamic (n=3)	69.81	72.63
Dynamic (n=4)	70.25	73.14

Table 2: Comparison of Static vs Dynamic personality with differing sliding window size

Approach	MELD	IEMOCAP
Full context	68.31	69.42
Only intra	64.77	66.87
Only inter	67.14	68.55
Inter and Intra (w/o c-a)	69.13	72.01
Inter and Intra (w c-a)	70.25	73.14

Table 3: Impact of segregating the context into granular parts. *c-a* refers to **Cross-Alignment**.

6 Conclusion

We introduced PERC, a personality-aware framework for ERC. PERC first segregates intra- and inter-speaker contexts to disentangle self- and cross-speaker emotional influences. It then infers static or dynamic personality traits to capture how individual dispositions evolve across dialogue turns. Finally, contrastive cross-alignment between intra–intra and inter–inter representations enforces contextual consistency and sharper decision boundaries. Experiments on three ERC benchmarks show that PERC outperforms state-of-the-art methods, confirming that dynamically aligned personality cues enrich both speaker modeling and emotional reasoning. Future work includes extending PERC to multilingual conversations and exploring multimodal inputs for emotion recognition.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Speaker-guided encoder-decoder framework for emotion recognition in conversation. *arXiv preprint arXiv:2206.03173*.

Urszula Barańczuk. 2019. The five factor model of personality and emotion regulation: A meta-analysis. *Personality and individual differences*, 139:217–227.

Lisa Feldman Barrett. 2022. Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist*, 77(8):894.

Anna Baumert, Manfred Schmitt, Marco Perugini, Wendy Johnson, Gabriela Blum, Peter Borkenau, Giulio Costantini, Jaap JA Denissen, William Fleeson, Ben Grafton, and 1 others. 2017. Integrating personality structure, personality process, and personality development. *European Journal of Personality*, 31(5):503–528.

Nadin Beckmann and Robert E Wood. 2017. Dynamic personality science. integrating between-person stability and within-person change.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

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.

Paul T Costa Jr and Robert R McCrae. 1992. Four ways five factors are basic. *Personality and individual differences*, 13(6):653–665.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Paul Ekman. 1992. Are there basic emotions?

William Fleeson. 2007. Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of personality*, 75(4):825–862.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. 2025. Laerc-s: Improving llm-based emotion recognition in conversation with speaker characteristics. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6748–6761.

Adrian Furnham. 1996. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. *Personality and individual differences*, 21(2):303–307.

Qingqing Gao, Jiuxin Cao, Biwei Cao, Xin Guan, and Bo Liu. 2024. Cept: a contrast-enhanced prompt-tuning framework for emotion recognition in conversation. In *Proceedings of the 2024 Joint International*

Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 2947–2957.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3):271–299.

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. *arXiv preprint arXiv:2306.01505*.

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

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8002–8009.

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. Association for Computational Linguistics.

Carl Gustav Jung, Gerhard Adler, and Richard Francis Carrington Hull. 2014. *The development of personality*. Routledge.

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

Yujin Kang and Yoon-Sik Cho. 2025. Beyond single emotion: Multi-label approach to conversational emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24321–24329.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Randy J Larsen and Timothy Ketelaar. 1991. Personality and susceptibility to positive and negative emotional states. *Journal of personality and social psychology*, 61(1):132.

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

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*.

Sointu Leikas, Jan-Erik Lönnqvist, and Markku Verkasalo. 2012. Persons, situations, and behaviors: Consistency and variability of different behaviors in four interpersonal situations. *Journal of Personality and Social Psychology*, 103(6):1007.

Michael Lewis. 1999. On the development of personality. *Handbook of personality: Theory and research*.

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

Jiang Li, Xiaoping Wang, and Zhigang Zeng. 2024b. A dual-stream recurrence-attention network with global–local awareness for emotion recognition in textual dialog. *Engineering Applications of Artificial Intelligence*, 128:107530.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 1204–1214.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200.

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

Ximing Li, Yuanchao Dai, Zhiyao Yang, Jinjin Chi, Wanfu Gao, and Lin Yuanbo Wu. 2025. Utterance-level emotion recognition in conversation with conversation-level supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24503–24511.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Yingjian Liu, Jiang Li, Xiaoping Wang, and Zhigang Zeng. 2024. Emotionic: emotional inertia and contagion-driven dependency modeling for emotion recognition in conversation. *Science China Information Sciences*, 67(8):182103.

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. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.

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*, pages 6818–6825.

Louis-Philippe Morency. 2013. The role of context in affective behavior understanding. *Social Emotions in Nature and Artifact*, 2:8–27.

Julian A Nasello, Jean-Marc Triffaux, and Michel Hansenne. 2023. Individual differences and personality traits across situations. *Current Issues in Personality Psychology*, 12(2):109.

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

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Xiangyu Qin, Zhiyu Wu, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, Li Wang, and Jinshi Cui. 2023. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13492–13500.

Tazeek Bin Abdur Rakib, Ambuj Mehrish, Lay-Ki Soon, Wern Han Lim, and Soujanya Poria. 2025. Dialogxpert: Driving intelligent and emotion-aware conversations through online value-based reinforcement learning with llm priors. *arXiv preprint arXiv:2505.17795*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Preprint*, arXiv:1811.00146.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*.

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Emotion recognition in conversation via dynamic personality. In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 5711–5722, Torino, Italia. ELRA and ICCL.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Camille J Wynn, Tyson S Barrett, and Stephanie A Borrie. 2024. Conversational speech behaviors are context dependent. *Journal of Speech, Language, and Hearing Research*, 67(5):1360–1369.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. 2024. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pages 277–292. Springer.

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

Zhilin Yang. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI Workshops*, volume 18, pages 44–52.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *IJCAI*, pages 4524–4530.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arXiv preprint arXiv:2106.01071*.

A Training Personality Extraction Model

Formally, given an input text x, we first encode it with the fine-tuned BERT to obtain the [CLS] vector:

$$h = \text{BERT}(x)[CLS]. \tag{14}$$

The vector h is then passed through a single-layer BiLSTM (Hochreiter and Schmidhuber, 1997), from which we take the final hidden state z:

$$z = \text{BiLSTM}(h). \tag{15}$$

We project z using a ReLU activation and dropout:

$$\hat{z} = \text{Dropout}(\text{ReLU}(z)). \tag{16}$$

Finally, the prediction layer outputs probabilities for each of the five traits:

$$\hat{y} = \sigma(W\hat{z} + b), \tag{17}$$

where σ denotes the sigmoid activation applied element-wise.

The model is trained with the binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{BCE}} = -\sum_{i=1}^{5} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{18}$$

where $y_i \in \{0, 1\}$ is the ground-truth label for trait i.

B Dynamic vs Static Personality

In the case of **static personality**, the personality traits are extracted only once from a sliding window of n utterances and remain fixed throughout the conversation. The static embedding for speaker s is defined as:

$$z^p_{\text{intra},i} = f(p_{i-n}, \dots, p_i),$$

where $f(\cdot)$ denotes the extraction function applied over the speaker's n most recent utterances. For all earlier timesteps before the window, we apply zero-padding to represent the absence of prior personality information:

$$z^p_{\text{intra},j} = \mathbf{0}, \quad j < i - n.$$

In contrast, **dynamic personality** re-estimates the traits at every conversational turn, allowing the embedding to evolve as the dialogue progresses:

$$\{z^p_{\text{intra},i-n}, \dots, z^p_{\text{intra},i}\}, \quad \forall i.$$

This continuous update enables the model to adapt to contextual and emotional shifts, reflecting changes in speaker expression over time.

C Running Example

Turn-by-turn segmentation. Table 4 presents a running example illustrating how inter- and intra-speaker segments are constructed for each utterance. For Utterance 1, only an intra-segment exists, as no prior context is available and the utterance is produced by speaker *Jade*. For Utterance 2, the inter-segment contains the previous utterance (1) from a different speaker, while the intra-segment contains *Chandler*'s own utterance. At Utterance 3, the inter-segment aggregates *Chandler*'s prior utterance (2) for contextual grounding, whereas the intra-segment aggregates *Jade*'s utterances (1, 3). For Utterance 4, the inter-segment includes all preceding utterances from other speakers (1–3), and the intra-segment contains *Ross*'s current utterance. Finally, for Utterance 5, the inter-segment aggregates utterances from other speakers (1, 3, 4), while the intra-segment aggregates *Chandler*'s utterances (2, 5).

Segment representations and emotion prediction. Intra-segments consist exclusively of utterances produced by the target speaker and are used to construct intra-personality and intra-context representations. Inter-segments consist of utterances from all other speakers and are used to model inter-personality and inter-context representations. Both segments are aggregated within a sliding window and encoded using a frozen personality extractor for personality representations and a fine-tuned RoBERTa encoder for contextual representations. The resulting personality and context embeddings are aligned via contrastive cross-alignment, and the fused representations are subsequently used to predict the emotion label of each utterance.

D Contrastive Cross-Alignment

First, we define the cosine similarity function as:

$$s(x, y) = \exp(\mathrm{sim}(x, y)/\tau). \tag{19}$$

where $\mathrm{sim}(\cdot)$ denotes cosine similarity and τ is the temperature parameter.

$$\mathcal{L}_{\text{intra}}^{p\to c} = -\log \frac{s(z_{\text{intra}}^{p}, z_{\text{intra}}^{c})}{s(z_{\text{intra}}^{p}, z_{\text{intra}}^{c}) + \sum_n s(z_{\text{intra}}^{p}, z_{\text{inter},n}^{c})} \tag{20}$$

$$\mathcal{L}_{\text{intra}}^{c\to p} = -\log \frac{s(z_{\text{intra}}^{c}, z_{\text{intra}}^{p})}{s(z_{\text{intra}}^{c}, z_{\text{intra}}^{p}) + \sum_n s(z_{\text{intra}}^{c}, z_{\text{inter},n}^{p})} \tag{21}$$

The final intra-view $\mathcal{L}_{\text{intra}}$ loss averages both directions:

$$\mathcal{L}_{\text{intra}} = \frac{1}{2}\left(\mathcal{L}_{\text{intra}}^{p\to c} + \mathcal{L}_{\text{intra}}^{c\to p}\right), \tag{22}$$

Likewise, a similar formulation is used for the inter-view loss $\mathcal{L}_{\text{inter}}$. This bidirectional formulation ensures that personality and context embeddings reinforce each other, fostering coherent intra- and inter-speaker representations for robust emotion classification.

E Dataset Breakdown

Personality The Essays (Pennebaker and King, 1999) is based on the stream of consciousness in a controlled environment. This dataset was created by students of the APA (American Psychological Association). In addition, every essay is labeled based on classes of the Big 5 personality traits. Hence, this dataset includes a label for each personality trait. More information is shown in Table 5.

ERC Different emotions are used across the datasets. It can be seen that EmoryNLP is composed of the following emotions: peaceful, powerful, sad, mad, scared, neutral. IEMOCAP is composed of the following emotions: happiness, excited, sadness, anger, frustrated, neutral. MELD is composed of the following emotions: happiness, surprise, sadness, disgust, fear, and neutral. The size of the dataset is shown in Table 6.

F SOTA Baselines

We compare our method against the following state-of-the-art baselines:

- **DialogueRNN** (Majumder et al., 2019): employs an RNN-based architecture to track the emotional states of speakers throughout a conversation.
- **DialogueGCN** (Ghosal et al., 2019): captures intra- and inter-speaker dependencies through a graph-based representation.
- **COSMIC** (Ghosal et al., 2020): integrates commonsense knowledge into conversational modeling.
- **HiTrans** (Li et al., 2020): combines local and global contextual modeling using a hierarchical transformer framework.

Utterance #	Utterance	Speaker	Inter-segment	Intra-segment	Emotion
1	Oh, Bob, he was nothing compared to you. I had to bite my lip to keep from screaming your name.	Jade	[ ]	[1]	Sadness
2	Well, that makes me feel so good.	Chandler	[1]	[2]	Joy
3	It was just so awkward and bumpy.	Jade	[2]	[1, 3]	Neutral
4	Bumpy?	Ross	[1, 2, 3]	[4]	Surprise
5	Well, maybe he had some kind of uh, new, cool style, that you're not familiar with.	Chandler	[1, 3, 4]	[2, 5]	Neutral

Table 4: Example dialogue illustrating inter- and intra-speaker segments with associated emotion labels.

Trait	Yes	No
OPN	1271	1196
CON	1254	1213
EXT	1275	1192
AGR	1309	1158
NEU	1234	1233

Table 5: Value counts for each of the personality traits.

- **AGHMN** (Jiao et al., 2020): adopts a hierarchical memory network with attention-based GRUs to model and summarize historical context.
- **DialogueCRN** (Hu et al., 2021): leverages contextual reasoning networks for multi-turn emotional inference.
- **DAG-ERC** (Shen et al., 2021b): employs a directed acyclic graph network to capture causal flow between utterances.
- **DialogXL** (Shen et al., 2021a): adapts XLNet with a dialog-aware self-attention mechanism for longer context modeling.
- **EmoBERTa** (Kim and Vossen, 2021): prepends speaker identities to each utterance using special separation tokens.
- **CoMPM** (Lee and Lee, 2021): integrates speaker-specific pretrained memory within a contextual memory pretraining framework.
- **SPCL** (Song et al., 2022): employs supervised prototypical contrastive learning for emotion classification.
- **SACL** (Hu et al., 2023): combines adversarial learning with supervised contrastive learning for robust emotion representations.
- **ErNETCL** (Li et al., 2024a): utilizes curriculum learning to capture temporal context in conversations.
- **SGED** (Bao et al., 2022): introduces a speaker-guided encoder-decoder framework to model intra- and inter-speaker dependencies.
- **SKAIG** (Li et al., 2021): incorporates psychological states of speakers within a graph-based structure.
- **CauAIN** (Zhao et al., 2022): integrates emotion cause detection to improve contextual understanding.
- **CoG-BART** (Li et al., 2022): adopts a supervised contrastive learning approach within an encoder-decoder architecture.
- **BERT-ERC** (Qin et al., 2023): fine-tunes BERT by incorporating contextual and structural dialogue information.
- **InstructERC** (Lei et al., 2023) reformulates ERC as a generative instruction-following task using LLMs, integrating retrieval-based dialogue supervision and auxiliary alignment tasks to model speaker roles and emotional tendencies.
- **ERC-DP** (Wang et al., 2024): introduces prompt-based personality features for emotion recognition.
- **BiosERC** (Xue et al., 2024) employs LLMs to extract speaker biographical and personality-related information from conversations as auxiliary knowledge to enhance utterance-level emotion classification.

Dataset	Training	Dev/Test
EmoryNLP (Zahiri and Choi, 2018)	9934	1344/1328
MELD (Poria et al., 2018)	9989	1109/2610
IEMOCAP (Busso et al., 2008)	5810	1623

Table 6: Breakdown of the datasets used. Sample counts are in terms of the number of utterances used.

- **CEPT** (Gao et al., 2024): applies prompt-tuning combined with masked language modeling and supervised contrastive learning.
- **DualRAN** (Li et al., 2024b): employs a hybrid RNN and attention-based model to capture both local and global contexts.
- **LaERC-S** (Fu et al., 2025) leverages LLMs with a two-stage reasoning framework to infer speaker characteristics and track speaker-level emotional dynamics for utterance-level emotion prediction.
- **ML-ERC** (Kang and Cho, 2025): presents a multi-label pluggable framework for ERC tasks.
- **DERC-PL** (Li et al., 2025): presents a weakly-supervised framework for tackling ERC.

G Hyperparameter Settings

Number of Speakers	MELD
2	72.35
3	71.04
>3	67.36

Table 7: Weighted F1 performance on MELD depending on the number of speakers in the conversation.

Model	MELD	IEMOCAP
Overall	70.25	73.14
First utterance	54.77 (-15.48)	59.16 (-13.98)

Table 8: Impact of context (overall vs first utterance only) on ERC performance. Scores in parentheses show the drop from using only the first utterance.

We divide our experimental setup into three components:

- **Personality Prediction:** We fine-tune the BERT model with a batch size of 16 using the Adam (Kingma and Ba, 2014) optimizer with a learning rate of 1×10^{-2}. Personality modeling is performed with 10-fold cross-validation. The BiLSTM layer uses a hidden dimension of 128.
- **Fine-tuning with RoBERTa:** The RoBERTa encoder is fine-tuned with a batch size of 16, using the Adam optimizer with a learning rate of 2×10^{-4}.
- **ERC Modeling:** We set the weighting factor λ to 0.7 for the contrastive loss and the temperature parameter τ to 0.10. A sliding window of size $n = 4$ is used for both personality and context modeling, and the classification layer employs a hidden dimension of 128. We use the Adam optimizer with a learning rate of 1×10^{-4}. For the focal loss, we set α to 0.5 and γ to 2.0.

H Additional Experimental Results

We present supplementary experiments that further analyze the components and behavior of PERC. These experiments complement the main results reported in Section 5 and provide deeper insights into the contributions of personality modeling, context structure, and conversational complexity. Specifically, we examine: (i) performance variation with the number of speakers, (ii) the impact of relying solely on the first utterance in a conversation, and (iii) the effect of intra- and inter-personality features

Tables 7–9 summarize these results. Table 7 presents MELD performance stratified by the number of speakers, highlighting how conversational complexity affects model accuracy. Table 8 compares performance when only the first utterance is available versus the full conversation, illustrating the importance of context and sequential information for ERC. Finally, Table 9 reports an ablation study of intra- and inter-personality features, showing how each component contributes to ERC performance.

Performance Based on Number of Participants. Results are reported only for MELD, as it contains conversations with two or more speakers (IEMOCAP has only two speakers per conversation). As

Configuration	**MELD**	**IEMOCAP**
Baseline – No personality segments	68.31	69.42
Baseline + Intra-Personality	68.56	69.74
Baseline + Inter-Personality	68.71	69.86
Baseline + Intra-Personality + Inter-Personality	68.92	70.34

Table 9: Effect of intra- and inter-personality features on ERC performance (weighted F1). We use the full context in the baseline for fair evaluation on the personality segments.

seen in Table 7, performance decreases as the number of participants increases. This is highly likely due to more complex interactions, varied personalities, and potential interruptions that affect inter-personality and inter-context modeling. In contrast, two-speaker conversations yield optimal performance, as the signals are simpler and easier for the model to capture.

Performance on Initial Utterance. Early-turn predictions are more challenging due to limited conversational context, as shown in Table 8. Intra-personality and intra-context information is restricted to a single utterance, while inter-personality and inter-context are absent for the first turn (zero-padding is used to represent missing prior information). Additionally, many first-turn utterances are neutral, increasing misclassification risk. Overall, these results demonstrate that first-turn classification is intrinsically harder, but subsequent turns benefit from richer context and personality information, allowing PERC to achieve strong performance across the full conversation.

Inter- and Intra-Personality Influence. The results in Table 9 shows that incorporating personality traits improves ERC performance, with the combination of intra- and inter-personality yielding the largest gains. The benefit is observed across both MELD and IEMOCAP, indicating generalizability. Notably, the inter-personality extractor provides a slightly larger improvement in isolation, suggesting that context from surrounding speakers plays a more influential role than a single speaker's personality when considered alone. Overall, these ablations demonstrate that our personality extractor, together with the intra-inter segmentation approach, meaningfully enhances ERC performance.