CMTD: Cognitive Modeling with Traits and Distortions for Multimodal Emotion Recognition in Conversations

Minh-Tien Nguyen[1], **Huu-Loi Le**[2], **Manh-Cuong Phan**[1], **Hajime Hotta**[3]
[1] Hung Yen University of Technology and Education, Hung Yen, Vietnam.
tiennm@utehy.edu.vn; cuongpm@spkt.edu.vn
[2] AI Academy Vietnam, Vietnam.
loilh@aiacademy.edu.vn
[3] Hajime Institute, Singapore.
hotta@hajime.institute

Abstract

This paper introduces a new multi-agent framework, CMTD (Cognitive Modeling with Traits and Distortions), for multimodal emotion recognition in conversations (MERC). Instead of relying on shallow analysis of emotions, CMTD reconstructs a cognitive model by taking advantage of stable personality traits, dynamic cognitive distortions, visual and acoustic features of interlocutors to enhance the emotional intelligence of LLMs. CMTD includes trait, distortion detection, vision, and speech agents that provide psychological and multimodal indicators for the fusion agent to make the final prediction. Experimental results on MELD and IEMOCAP show that traits temper negativity bias from distortions, and cognitive modeling with psychological, visual, and acoustic information can improve the performance of MERC. CMTD is flexible and easy to adapt to advanced emotional AI systems.[1]

1 Introduction

Emotions play a crucial role in psychology, fostering human engagement in communication (Zaki, 2019; Hosseini and Caragea, 2021; Shen et al., 2023, 2024b). Due to real-world applications, emotion recognition has been shifted from text-only to multimodalities (multimodal emotion recognition in conversations - MERC) (Ma et al., 2023; Van et al., 2025; Dutta and Ganapathy, 2025). This is because in conversations, human emotions are complicated and require the deep comprehension of an input from multiple channels (Alani et al., 2023). Emotion-powered AI systems are useful in various actual scenarios, such as mental health support (Lai et al., 2023; Shen et al., 2024a; AlMakinah et al., 2024), customer service (Fung et al., 2016), or dialog systems (Lin et al., 2019b; Ma et al., 2020).

Psychological studies have confirmed interconnections between human thoughts and emotions (Wright et al., 2017; Beck, 2020; Li et al., 2025). For example, Cognitive behavioral therapy (CBT) has shown that emotions have an association with thoughts when people perceive an event (Beck, 2020) (happiness tends to form positive thoughts associated with positive emotions). The core component of this association are distortions (thinking patterns) that tend to bias short-term emotions (Ford et al., 2018; Curtiss et al., 2021). The composition of emotions is also formed at different levels such as static and context-independent (traits) (Goetz et al., 2015) and dynamic and context-dependent states (Goetz et al., 2015; Zheng et al., 2023).

MERC performance has recently been improved (Feng and Fan, 2025; Van et al., 2025; Ai et al., 2025; Ma et al., 2023; Huang et al., 2024b; Fu et al., 2025a; Wu et al., 2025a; Yang et al., 2025). However, complex thinking patterns at different perception levels in the human brain may challenge the shallow analysis of MERC. Perception of these patterns may require a cognitive model that can perceive multimodalities in different emotion levels (static or dynamic). Let's take Figure 1 as an example. Both text-only and DoT (Chen et al., 2023) methods could not predict the correct neutral emotion of Ross (the third utterance) due to the lack of: cognitive modeling (Text-only) and the understanding of multimodal information (DoT). In contrast, we consider personal traits formed in a long time have a correlation with static states of emotions and cognitive distortions formed in short-term situations and thoughts can be associated with dynamic states of emotions. By integrating personal traits and distortions with multimodal data, our model can correctly predict emotions. For example, the *conscientiousness* trait (linked to stronger self-regulation and reduced impulsive aggression (Javaras et al., 2012)) and the Mind Reading distortion provide additional indicators of textual, visual, and acoustic features (Figure 2) that allows our model to predict the emotion of the

[1]Github link: https://github.com/Shaun-le/CMTD.git

Findings of the Association for Computational Linguistics: ACL 2026, pages 839–854
July 2-7, 2026

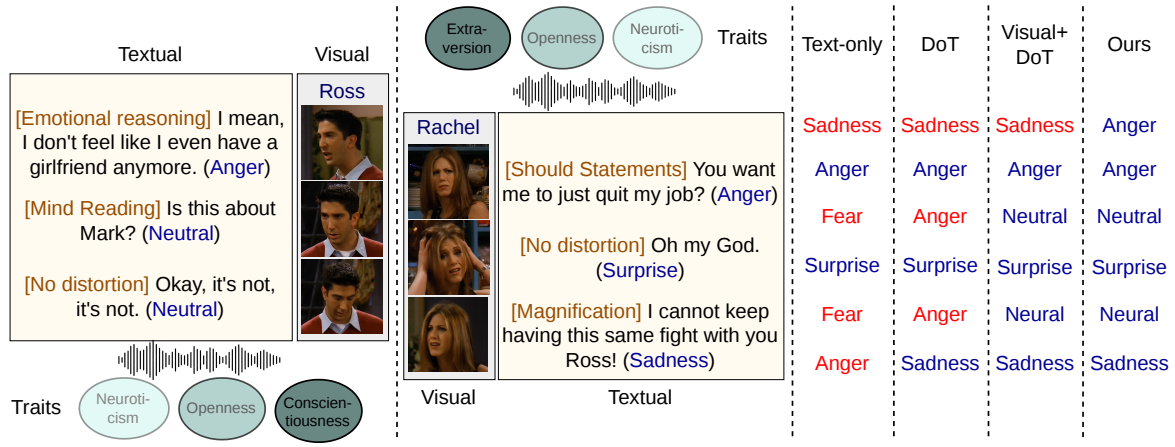

Figure 1: An example from MELD. Distortions are predicted by DoT (Chen et al., 2023) and traits are derived from Shen et al. (2025). Gold and correct predicted labels are in blue. Colors of traits represent their association for each speaker based on big five (Goldberg, 1992) (darker is stronger). Complete outputs of methods are shown in Figure 5.

second Ross's utterance as *neutral*. The fifth utterance with no distortion is misclassified by DoT due to a carry-over bias from the preceding distorted context. Again, conscientiousness acts as a static factor, combined with visual and acoustic features, guiding our model to predict a *neutral* emotion.

This observation encourages a question: **"Does the integration of personal traits and distortions with multimodal data (text, vision, and speech) yield reliable MERC gains in zero-shot settings?"** To answer this question, we introduce a new multi-agent framework that takes into account psychology (personal traits and distortion detection) combined with textual, visual, and acoustic information for MERC. Its design combines a set of agents by prompting LLMs. A personal trait agent analyzes the traits of interlocutors based on the big five concept (Goldberg, 1992). A distortion detection agent includes a set of small agents built up by Diagnosis-of-Thought (DoT) (Chen et al., 2023). A visual agent comprehends the facial expressions of interlocutors. A speech processing agent processes the speech of interlocutors. A fusion agent synthesizes psychological, textual, visual, and acoustic indicators for the final prediction. In summary, this paper makes three main contributions:

- It integrates psychology (traits and distortions), text, vision, and speech to reconstruct the cognitive model of interlocutors. It allows our method to combine static and dynamic emotions for deeper emotional understanding.
- It introduces a new multi-agent reasoning framework with distortion and trait moderation. The framework is flexible and easily adapted to advanced emotional AI systems.
- It provides a comprehensive zero-shot evaluation on MELD and IEMOCAP. The ablation study shows the behavior of the framework.

2 Related Work

2.1 Emotion Recognition in Conversations

Text-based ERC There are three main approaches of text-based ERC. The first approach designs various learning techniques such as mixture of experts (Lin et al., 2019a), emotions at coarse to fine levels (Li et al., 2020), multimodal fusion (Chudasama et al., 2022; Ai et al., 2025), or using external knowledge (Li et al., 2022). The second fine-tunes LLMs using fine-tuning (Lei et al., 2023; Zhang et al., 2023; Liu et al., 2024). The final creates agentic frameworks (Mozikov et al., 2024).

Multimodal ERC The early work of MERC used GRU (Majumder et al., 2019) or LSTM (Poria et al., 2017) to represent multimodal information. More recent work used the multimodal Transformer architecture (Ma et al., 2023) for speech or vision fusion using graph-based methods (Scarselli et al., 2008; Mai et al., 2023; Feng and Fan, 2025; Huang et al., 2024a; Van et al., 2025; Shou et al., 2025; Ai et al., 2025; Wu et al., 2025a), or cross-modal information fusion with Transformer (Zhang and Li, 2023; Ma et al., 2023; Huang et al., 2024b; Zhu et al., 2024). MERC has also been shifted to a generation task (Fu et al., 2025a; Cheng et al., 2024; Dutta and Ganapathy, 2025; Fu et al., 2025b; Tanioka et al., 2024; Yang et al., 2025).

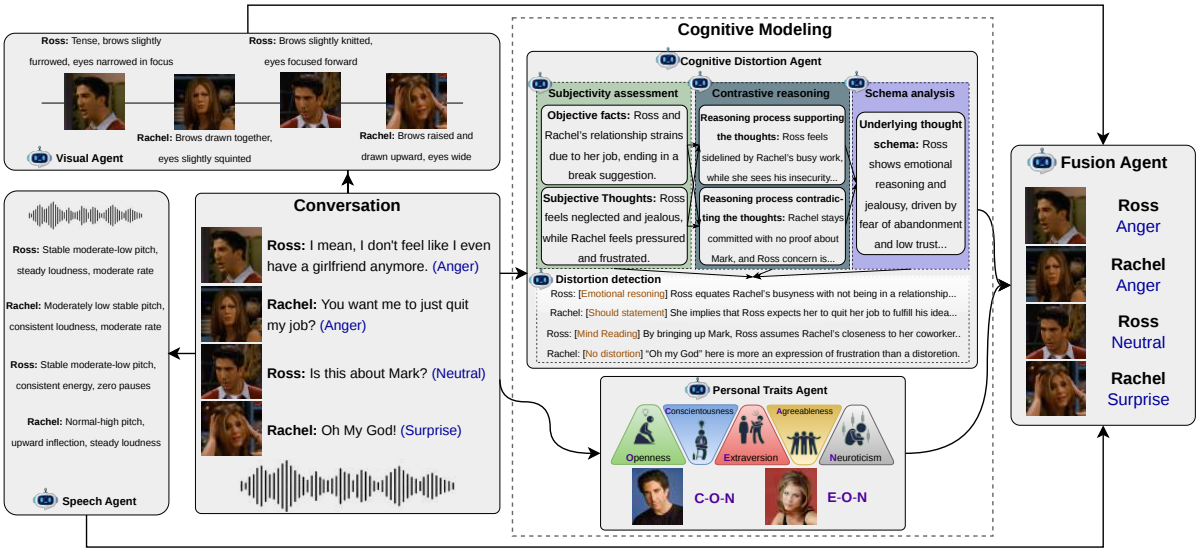

Figure 2: The architecture of the framework. The personal trait agent analyzes the traits of interlocutors. The distortion detection agent detects distortions of interlocutors. The visual agent comprehends facial expressions. The speech agent generates description of four acoustic features: frequency, energy, speaking rate, and voice quality. The fusion agent synthesizes indicators from trait, distortion, visual and speech agents for the final prediction.

2.2 Psychology-powered AI

Personal traits A personal trait (personality or characteristic) is a fundamental construct in psychology that reflects individual's behavior, thinking, and emotional patterns (Li et al., 2025). Psychological research (Rosenberg, 1998) distinguishes trait (static and context-independent) (Goetz et al., 2015) and state (dynamic and context-dependent) (Goetz et al., 2015; Zheng et al., 2023) emotions. Personal traits have recently been used to enhance the emotional intelligence of AI models (Wang et al., 2024b; Xue et al., 2024; Fu et al., 2025a; Shen et al., 2025; Wu et al., 2025b).

Cognitive distortion detection Distortion detection is the first important step that allows the therapist to understand the cognitive model of patients for their therapy in CBT (Wright et al., 2017; Beck, 2020) and ACT (acceptance and commitment therapy) (Harris, 2006; Hayes et al., 2011). This detection has recently been used to empower AI models such as psychotherapy (Chen et al., 2023), practical training in CBT treatment (Wang et al., 2024a), LLMs' reasoning (Lim et al., 2024), or cognition-aware coaching (Hotta et al., 2025).

Our method focuses on vision and speech integration and takes advantage of traits (static emotions) (Wang et al., 2024b; Li et al., 2025) and distortions (dynamic emotions using DoT (Chen et al., 2023)). However, we push DoT to a deeper level of emotional reasoning by: (i) adding personal traits as a new component, (ii) designing four agents for distortion detection rather than using a single long prompt, and (iii) adding visual and acoustic features. We also share the idea with InsiteOut in building multi-agent frameworks (Mozikov et al., 2024); however, our method reconstructs cognitive models of interlocutors using traits and distortions rather than using a shallow analysis of emotions.

3 Cognition-aware MERC

3.1 Problem Statement

Let $C = \{u_i = (x_i, v_i, s_i) | i = 1, ..., N\}$ be a conversation, where each utterance u_i consists of textual content x_i, a temporally aligned visual frame v_i, and the speaker's identity with the speech segment s_i. The task of MERC is to predict the entire sequence of labels $Y^N = (y_1, y_2, ..., y_N), y_i \in \mathcal{Y}$ (see Figure 1), where $\mathcal{Y}$ is a predefined set of emotional categories, and N is the total number of utterances in the conversation, using the combination of multimodal information from x_i, v_i, and s_i.

3.2 Overview of the Framework

Figure 2 introduces CMTD, the proposed psychology-aware framework, which is extensible for any existing LLMs. Given a conversation, the framework comprehends the conversation with cognitive modeling (**Section** 3.3.2), visual comprehension (**Section** 3.4), speech processing (Section 3.5), and information fusion (**Section** 3.6).

3.3 Cognitive Modeling

We define trait and distortion detection agents to model static and dynamic states of emotions.

3.3.1 Personal Trait Agent

Psychological studies have shown the correlation between traits and emotions (Wright et al., 2017; Beck, 2020; Wu et al., 2025b). For example, people with neuroticism tend to have negative emotions (e.g. sadness, anxiety) (Lahey, 2009a) (Figure 1). Inspired by this, we consider personal traits to have an association with static states of emotions. We follow the big five personality test[2] (Goldberg, 1992) that defines five traits: *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism* (Appendix A.1). These traits have been validated as stable dispositions that systematically influence affective responses of interlocutors.

To implement the trait agent, we design a method that elicits Big Five trait scores from LLMs (Zhu et al., 2025). Traits are inferred from speaker histories using zero-shot prompting with LLMs. More precisely, for the MELD dataset, characters come from a TV series with consistent personas across episodes. We therefore adopt the persona definitions from Shen et al. (2025), and use them as contextual anchors when prompting LLMs for trait inference. For IEMOCAP, where each dialogue involves just two interlocutors without pre-defined personas, we apply the zero-shot trait prompt for each character based on their turn-level speaking history. After LLMs produce scores, we select the top-3 traits with highest scores to represent their personality in that episode (Figure 1).

3.3.2 Cognitive Distortion Agent

We consider distortions associate with dynamic states of emotions because they are formed by the short-term perception of situations and thoughts. Inspired by DoT (Chen et al., 2023), we design a set of agents for distortion detection in four steps.

Subjectivity assessment The initial step in distortion detection is to evaluate the degree of subjectivity in the speaker's utterances. Since conversations often intertwine objective facts with subjective thoughts, this agent makes this distinction by identifying facts versus subjective thoughts. This agent was implemented by prompting LLMs to answer the following questions: what is the situation, facts, objectives, and thoughts that are subjective. By doing so, it establishes objective events as a trustworthy foundation while recognizing the subjective layer that frames personal perspectives.

Contrastive reasoning This agent applies two complementary reasoning paths to subjective thoughts grounded in objective facts. The first path provides justification that supports the expressed thoughts, while the second highlights reasoning that challenges or contradicts the thoughts. In practice, this agent assesses the reason why the thoughts are true or false and finds the reasoning processes that support and do not support these thoughts. By contrasting these opposing interpretations derived from the same evidence, the framework can uncover divergent perspectives and reveal more precisely the underlying thought schemas of speakers.

Schema analysis This agent answers why speakers form particular reasoning processes by two questions: why people come up with such reasoning process supporting the thought and what is the underlying cognition mode of it? In psychology, a schema refers to cognitive structures in which knowledge, beliefs, and expectations are integrated to shape an individual's perceptive model. By uncovering these schemas, the framework can reconstruct hidden cognitive patterns, thereby identifying the thought structures and potential distortions.

Distortion detection This agent simulates the diagnostic process in CBT, where a practitioner first constructs a cognitive model based on the situation, then identifies distorted thinking patterns before suggesting interventions (Beck, 2020). The detection agent recognizes distortions and generates their explanations at the utterance level based on information from the three above agents. It uses a prompt that includes three tasks: (i) analyzing the thought patterns of the participants and diagnosing of thought analyses; (ii) identifying if there is cognitive distortion in the conversation; (iii) recognizing the types of distortions. The taxonomy of distortions includes ten categories: personalization, mind reading, overgeneralization, all-or-nothing thinking, emotional reasoning, labeling, magnification, mental filter, should statements, and fortune-telling (Beck, 2020; Shreevastava and Foltz, 2021).

3.4 Visual Comprehension

The diverse spectrum of emotions may challenge text-only methods (Figures 1 and 5). When using

[2]https://www.kaggle.com/datasets/tunguz/big-five-personality-test

visual information, our model can recognize Ross's tense posture, slightly furrowed brows, narrowed eyes, and open mouth with lips stretched in mid-speech, leading to the correct prediction of anger.

To harness these nonverbal cues, we develop a Visual Agent that performs in parallel with the cognitive modeling agents. For each utterance, a representative facial frame is extracted to capture the speaker's stable expression. This agent then inspects detailed facial features, including gaze direction, eyebrow movement, and mouth configuration, which have well-established links to affective states (Ekman and Friesen, 1978). The structured descriptions of visual cues are aligned with each utterance and are then combined with the speaker's personal traits and the outputs from distortion detection.

3.5 Speech Processing

Speech carries rich paralinguistic information that cannot be fully captured by textual content alone. In human communication, vocal signals such as tone, pitch, speaking rate, and intensity often reveal underlying emotional states, even when the spoken words appear neutral (Schuller et al., 2018; Poria et al., 2017). For instance, elevated pitch and increased loudness may indicate anger or excitement, while slower speech with lower energy may reflect sadness or fatigue (see Figure 2) (Busso et al., 2008). To leverage these signals, we design a Speech Agent that operates alongside the cognitive and visual agents. Instead of relying on low-level handcrafted features or standalone classifiers, we adopt a prompt-based acoustic reasoning approach using LLMs (prompt in Table 6). Given an audio segment corresponding to each utterance, the Speech Agent analyzes the signal through structured instructions that guide the model to extract interpretable acoustic descriptions. The Speech Agent is prompted to infer four key signals.

Fundamental frequency (F0) This signal captures pitch levels, variability, and contour dynamics, which are associated with emotional arousal and tension (Schuller et al., 2018).

Loudness/energy This signal reflects vocal intensity and its variation, often linked to emotions such as anger, excitement, or fatigue (Kacur et al., 2021; Chowdhury et al., 2025).

Speaking rate and pauses This signal describes temporal delivery patterns, including speed and hesitation, which correlate with anxiety, confidence, or sadness (Narayanan et al., 2009; Yang et al., 2024).

Voice quality This signal characterizes perceptual qualities such as breathiness, strain, roughness, or clarity, which provide cues about stress, calmness, or vulnerability (Akccay and Oguz, 2020; Bhangale and Kothandaraman, 2023).

These speech-derived descriptions are then aligned with textual content, visual cues, and cognitive indicators (traits and distortions) and passed to the Fusion Agent. In particular, acoustic signals help resolve ambiguity when textual content represents a neutral emotion or visual cues are inconclusive. By incorporating prompt-based speech understanding, the framework captures low-level acoustic patterns, cognitive indicators, and high-level affective interpretations in a unified manner.

3.6 Fusion Agent

The Fusion Agent is the central integrator that unifies the complementary indicators of four perception layers: (i) static emotions from traits, (ii) dynamic emotions associated with distortions arising from subjective cognition, (iii) visual cues, and (iv) acoustic signals. Traits act as context-independent states of emotions that stabilize the fusion process. When distortions introduce short-term biases or visual cues appear ambiguous, trait-informed moderation provides a psychological anchor. In addition, acoustic signals provide a kind of physical indicators that have a strong correlation with emotions when interlocutors perceive a situation and speak out their thoughts. Combined with psychological and visual features, this prevents the model from overreacting to transient signals and encourages consistency across dialogue turns. Such integration simulates how humans interpret emotions not only by attending to momentary expressions or reasoning patterns, but also grounding them in knowledge of enduring dispositions. In practice, psychological, visual, and acoustic indicators are put in a prompt for LLMs to make the final prediction.

4 Experimental Settings

Datasets The evaluation uses two datasets. `MELD` (Poria et al., 2019) is a multimodal dataset that includes conversations collected from TV-series. Each utterance has emotion and sentiment labels and contains audio, visual, and textual modalities. `IEMOCAP` (Busso et al., 2008) includes dyadic conversations between actors, labeled with six emotions, and provides synchronized multimodal data.

The dataset was organized into five sessions, and in our experiments, we used Session 5 as the test set. Table 1 shows the statistics of these datasets.

Table 1: Statistics of MELD and IEMOCAP datasets.

Dataset	#Convs		#Utters		#Classes	Genre
	Train	Test	Train	Test		
MELD	1153	280	11098	2610	7	TV Sitcom
IEMOCAP	120	31	5810	1623	6	Dialogue

Baselines We employed strong LLMs such as **GPT-4o-mini**, **GPT-4.1-nano**, **GPT-4.1-mini**, **GPT-4.1**, **GPT-o4-mini**, **GPT-5**, **GPT-5-nano**, **GPT-5-mini**, **Qwen2.5-7B** (Qwen et al., 2025), and **Mistral-7B** (Jiang et al., 2023), which are recent and promising results in zero-shot settings.

Several baseline configurations were also observed. **Text-only** relies solely on the textual input of the conversation. **DoT** incorporates the deep analysis of cognitive distortions to simulate reasoning at the thought level, using the prompts shared by (Chen et al., 2023). **Text+Visual**, a single agent, combines facial and textual information from utterances for prediction. **Visal+DoT**, a single agent, integrates visual features into DoT prompts. We also reimplemented **InsideOut** (Mozikov et al., 2024), a framework that employs LLM agents for ERC with five different emotional roles.

Settings Experiments were carried out in the zero-shot setting. This is because this setting is practical for the deployment of AI systems in actual cases (Chen et al., 2023; Mozikov et al., 2024). All the prompts are shown in Table 6.

5 Results and Discussion

5.1 Performance Comparison

Improvements over base LLMs Table 2 reports the improvement of the proposed framework with strong base LLMs. The scores indicate that the proposed framework obtains the best average results. For each method, it is also the best in almost all cases. Improvements come from two aspects. First, the framework models static and dynamic states of emotions by taking into account personal traits and distortions. It allows the framework to reconstruct cognitive models of interlocutors in a conversation for better perceiving complex thinking patterns. Second, our CMTD combines psychology, vision, speech, and text in a multi-agent architecture that improves the reasoning capability of each agent for the final emotion prediction. Psychology (traits and distortions) helps to reconstruct cognitive models empowered by the comprehension of textual, visual, and acoustic indicators. Hence, it enables deeper perception of complex thinking patterns. Promising results confirm our hypothesis and answer the research question in Section 1.

The Visual+DoT and DoT methods follow our framework. For Visual+DoT, competitive results come from the combination of visual features and deeper reasoning with DoT. It is similar to DoT, in which the method is based on distortion detection to better understand complex emotions. The performance of the Text+Visual method is inconsistent, in which adding visual features reduces the scores on MELD but improves the scores on IEMOCAP. It suggests a more sophisticated combination of textual and visual features, rather than a simple combination in a single prompt. Interestingly, InsideOut achieves lowest performance even using a set of agents for reasoning. A possible reason is that InsideOut uses a shallow analysis of five basic emotions (anger, disgust, fear, happiness, and sadness). Therefore, similar emotions (joyful and happy) may challenge the InsideOut method.

Challenge to SOTA methods CMTD was challenged to SOTA methods to create a reference due to different settings. **CoMPM** extracts pre-trained memory tracking of speakers and considers it as external knowledge for MERC (Lee and Lee, 2022). **AdaIGN** models intra-speaker and inter-speaker context dependencies using an adaptive interaction graph network (Tu et al., 2024). **DE-GCN** captures dialogue and event relations for MERC by a graph neural network (GNN) (Ai et al., 2024). **GS-MCC** addresses the challenges of GNN by introducing a graph-spectrum-based collaborative learning framework (Ai et al., 2025).

Table 3 shows that the proposed method obtains competitive accuracy with SOTA methods. For MELD, its performance is behind GS-MCC around 9.00%. This is because GS-MCC is a complicated graph-spectrum collaborative learning model using fine-tuned data. It enables the adaptation of GS-MCC to the downstream MERC task. In contrast, our method is LLM-agnostic tested in a zero-shot setting. The gap on IEMOCAP is smaller, in which our method produces better scores than CoMPM and DE-GCN. Compared to GS-MCC, the gap reduces to around 3.14%. It confirms the efficiency of our method that models static and dynamic states of emotions with traits and distortions.

Table 2: Results on MELD and IEMOCAP datasets. ACC and F1 denote Accuracy and Weighted F1 (Tu et al., 2024; Ai et al., 2024; Mozikov et al., 2024; Ai et al., 2025). The best scores are in **bold**, and second best is underlined.

Base Model	Text-only		DoT		InsideOut		Text+Visual		Visual+DoT		CMTD	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MELD												
gpt-4o-mini	46.30	39.71	47.70	39.80	38.47	39.71	40.57	40.80	49.00	51.05	**60.28**	**58.95**
gpt-4.1-nano	42.45	39.34	44.38	43.86	33.80	32.91	47.93	49.68	45.02	46.99	**56.19**	**54.43**
gpt-4.1-mini	54.56	57.45	57.66	59.03	42.91	43.14	45.53	35.59	54.78	56.74	**59.78**	**60.05**
gpt-4.1	51.61	54.89	54.56	57.20	46.22	46.71	49.23	50.00	52.03	55.45	**64.15**	**63.53**
o4-mini	59.58	60.87	58.70	59.78	55.48	55.81	60.91	60.14	61.69	62.41	**64.23**	**64.17**
gpt-5	57.37	59.69	59.47	61.41	56.56	58.74	52.21	54.13	57.76	60.19	**60.80**	**61.89**
gpt-5-nano	59.96	60.62	60.89	**60.79**	55.97	56.87	52.34	53.27	56.46	56.40	**60.89**	58.29
gpt-5-mini	55.47	57.58	56.44	58.24	53.75	55.81	49.85	51.30	54.71	56.04	**60.18**	**61.01**
Qwen2.5-7B	39.16	39.48	43.48	34.06	31.48	31.59	39.32	37.78	45.74	38.40	**54.13**	**49.85**
Mistral-7B	44.15	40.67	43.58	37.82	40.29	36.30	37.12	34.06	47.73	42.31	**52.46**	**48.17**
Average	51.06	51.03	52.69	51.20	45.49	45.76	47.61	48.11	53.36	53.20	**59.31**	**58.03**
IEMOCAP												
gpt-4o-mini	51.39	49.36	54.80	53.31	46.70	42.58	53.20	52.64	54.90	54.25	**58.21**	**56.29**
gpt-4.1-nano	42.00	40.92	44.78	42.18	36.03	34.00	49.04	47.72	**51.60**	49.79	50.38	**51.82**
gpt-4.1-mini	49.47	47.60	54.90	51.55	47.87	47.47	52.88	51.72	53.20	52.08	**52.22**	**52.15**
gpt-4.1	50.75	49.15	51.28	49.16	47.97	47.36	51.81	51.24	59.81	60.67	**63.33**	**63.02**
o4-mini	47.97	46.69	49.89	47.74	48.61	46.95	51.28	51.08	54.05	53.37	**59.91**	**59.98**
gpt-5	61.09	61.23	62.15	61.53	61.75	61.49	62.33	61.63	62.47	61.93	**63.01**	**62.21**
gpt-5-nano	52.77	51.53	57.75	56.16	45.07	44.14	**58.10**	**56.93**	57.68	55.96	57.36	55.96
gpt-5-mini	62.69	62.70	65.27	**64.85**	45.41	45.00	63.75	63.31	63.97	63.96	**65.86**	64.29
Qwen2.5-7B	47.87	45.68	44.35	37.88	41.79	36.86	51.71	50.75	**53.62**	**51.06**	48.39	47.73
Mistral-7B	45.10	42.91	40.62	38.42	24.31	22.13	41.01	44.91	44.67	42.49	**49.46**	**49.86**
Average	51.11	49.76	53.24	51.49	47.35	45.19	53.51	53.19	55.60	54.56	**56.81**	**56.33**

Table 3: Accuracy of SOTA methods. Or CMTD uses the best models (o4-mini on MELD and GPT-5-mini on IEMOCAP). Results are from Ai et al. (2025).

Method	MELD	IEMOCAP
CoMPM (Lee and Lee, 2022)	67.30	63.00
AdaIGN (Tu et al., 2024)	70.70	66.80
DE-GCN (Ai et al., 2024)	68.80	65.50
GS-MCC (Ai et al., 2025)	**73.90**	**69.00**
CMTD	64.23	65.86

5.2 Ablation Study

5.2.1 Psychological grounding

This section assesses trait and distortion detection in CMTD using LLMs and human validation.

Trait detection For trait detection on MELD, we used persona definition by human from Shen et al. (2025). For IEMOCAP, we used LLMs and humans for this observation. For LLMs, we created gold labels (score 1-5) based on Big-Five using GPT-4o-mini, GPT-4o-mini, and Gemina-2.5-Flash. If at least two detectors give the same or very close score (difference ≤ 0.3), we used the average of those agreeing scores. Otherwise, the median was used. Second, we run GPT-5 to predict the traits. Finally, we compared the Mean Absolute Error (MAE) on the utterance level. MAE on IEMOCAP is 0.3788. Tiny MAE of automatic validation shows good accuracy of personal trait detection. We also asked three experts to give scores (1-5) on 10 conversations of IEMOCAP. The MAE between the outputs of GPT-5 and human labels is 0.2995, showing very good performance of the detection.

Distortion detection We followed the procedure of trait detection using LLMs and humans. For LLMs, we used voting among GPT-4o-mini, GPT-4o-mini, and Gemina-2.5-Flash to create distortion labels. We then compared the outputs of GPT-5 with the voting labels. The accuracy on MELD is 0.5561 and on IEMOCAP is 0.7603. We also asked 3 experts to create labels of small scale utterances (64 for MELD and 67 for IEMOCAP). The Fleiss' Kappa is 0.6605 showing good agreement. We

then compared the outputs of GPT-5 with human labels. Accuracy on MELD is 0.7812 and 0.8125 on IEMOCAP, showing good performance.

5.2.2 Contribution of components

The contribution of components was observed. The observation was done using the leave-one-out test with the average of accuracy and F1 scores.

Table 4: The contribution of components for MERC with average results of accuracy and F1 scores. Agents mean this setting use multiple agents for MERC. The full setting uses multiple agents with traits in Figure 2.

Data		Settings					
—	Text	✓	✓	✓	✓	✓	✓
	DoT		✓	✓	✓	✓	✓
	Visual			✓	✓	✓	✓
	Speech				✓	✓	✓
	Agents					✓	✓
	Full						✓
MELD	ACC	51.06	52.69	53.36	57.54	58.80	**59.31**
	F1	51.03	51.20	53.20	56.50	57.02	**58.03**
IEMOCAP	ACC	51.11	53.24	55.60	55.88	56.12	**56.81**
	F1	49.76	51.49	54.56	54.96	55.04	**56.33**

Table 4 shows that the model using all components produces the best scores. Removal of traits reduces performance because traits provide static states of emotions, which support textual and visual features. In contrast, converting from multiple agents to a single agent significantly reduces performance. This is because using a long single prompt may challenge the reasoning of LLMs. When acoustic features are removed, the performance reduces quite significantly on MELD. It shows that speech provides important signals for emotion recognition. When visual cues are removed, the performance reduction in IEMOCAP is larger than that in MELD, suggesting that the visual features of IEMOCAP may be more important than those of MELD. The text-only method produces the lowest performance. It suggests that capturing complex emotions requires deep perception rather than a shallow analysis on the surface textual level.

5.2.3 Traits, vision, speech, and distortions

We observed that not all utterances in a conversation have distortions (Figure 1). Hence, the framework is empowered by traits, vision, and speech. When distortions do not exist, traits, vision, and speech provide personality, facial expressions, and acoustic signals associated with emotions.

In Figure 3, there are 2165 utterances in MELD that DoT predicts as no distortions. By combining traits, visual cues, and speech, CMTD can correctly predict 1500 utterances, which are higher than the Text-only method about 14.82%. On IEMOCAP, the gap is smaller, in which CMTD performs better the text-only method about 7.72%. It supports the promising results of CMTD in Table 2.

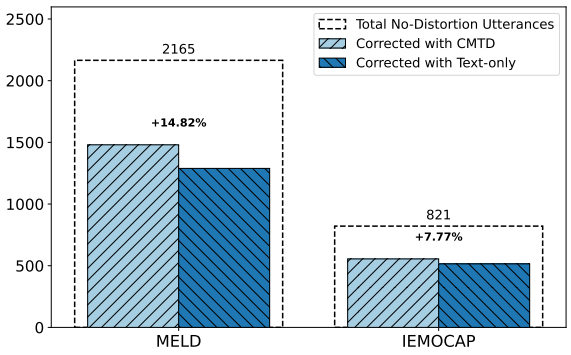

Figure 3: Distortions and other methods. Total no distortions were predicted by DoT.

5.2.4 Running time and cost

We investigated running time and cost, which are critical aspects for deploying AI systems. Table 5 shows this investigation with 100 samples from MELD. Text only, DoT, and Text+Visual are time and cost effective. It is understandable that these methods use single prompts. The Visual+DoT method is quite costly because it combines visual cues and cognitive modeling for reasoning. InsideOut and our CMTD are the most costly because they combine multiple agents for the final prediction. However, our CMTD only requires around 35 seconds with 0.000988 USD per sample (GPT-4o-mini), which is practical for actual cases. The time and cost of GPT-4.1 and GPT-5-mini are more expensive due to the quality of these models. We acknowledge this is slower than shallow baselines methods. However, in the domain of mental health and counseling, accuracy and deep empathy are often prioritized over millisecond latency. This latency is acceptable for several applications such as asynchronous support, typing delays in chatbots, or offline analysis of therapy sessions.

5.3 Error Analysis

Correct and wrong predictions Figure 4 shows the observation of predicted outputs between textual and motimodal models. In Figure 4(a), CMTD produces the highest number of correct utterances, followed, by Visual+DoT and Textual+Visual methods. It comes from the combination of cognitive modeling with multimodal indicators for better rea-

Table 5: The running time and cost of 100 samples from MELD. The time ratio was computed over the baseline.

Method	gpt-4o-mini			gpt-5-mini			gpt-4.1		
	Time(s)	Time ratio	Cost(USD)	Time(s)	Time ratio	Cost(USD)	Time(s)	Time ratio	Cost(USD)
Text only	200.32	—	0.005	984.76	—	0.011	155.76	—	0.067
DoT	198.81	0.992	0.016	3494.93	3.549	0.047	164.57	1.057	0.211
InsideOut	1883.51	9.403	0.053	7355.18	7.469	0.119	1574.87	10.111	0.899
Text+Visual	207.19	1.034	0.015	1085.78	1.103	0.027	184.04	1.182	0.199
Visual+DoT	1165.06	5.816	0.062	3919.48	3.980	0.179	1774.84	11.395	1.079
CMTD	3545.97	17.836	0.098	6516.74	1.865	0.255	4167.29	25.322	1.641

soning. In contrast in Figure 4(b), CMTD outputs less wrong predictions compared to others. It supports the best results of CMTD in Table 2.

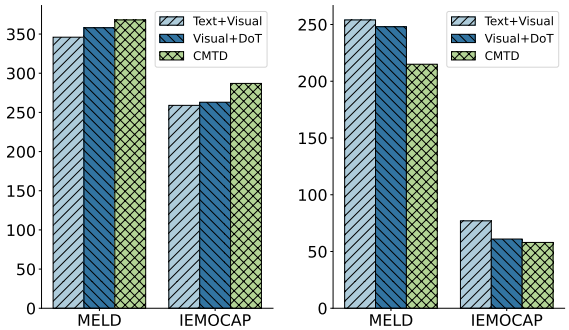

Figure 4: Observation between text-only and other visual models: (a) utterances incorrectly predicted by the Text-only method (y-axis) but corrected by visual methods; (b) utterances correctly predicted by the Text-only method (y-axis) but get wrong by visual methods.

Output observation In the first utterance of Figure 5, Ross expresses frustration where text-only and distortion-based models consistently misclassify the emotion as sadness due to pessimistic lexical cues. Our framework instead predicts anger, since the distortion agent identifies emotional reasoning (i.e. the speaker infers a negative state purely from a felt emotion, which is a known distortion that can amplify anger attribution), the visual agent detects tense facial expressions consistent with irritation, and acoustic description emphasizes minimal pauses, clear, smooth voice. Meanwhile, the trait profile (particularly high conscientiousness) acts as a regulatory anchor: empirical studies show that higher conscientiousness is associated with more effective recovery from negative emotional stimuli, i.e. individuals do not linger in or escalate negative affect (thus reducing bias toward sadness). (Barańczuk, 2019; Javaras et al., 2012).

In Ross's utterance "Okay, it's not, it's not", there is no distortion and the lexical content tends to show little emotional salience. Baselines that rely on textual features tend to misclassify it as fear or sadness, since the repetition and hesitation pattern resemble anxiety. CMTD correctly maintains the neutral label by integrating stable personality anchoring with visual cues: the speaker's face shows relaxed muscles, minimal brow movement, no mouth tension, all indicative of composure, and clear and smooth voice. These nonverbal indicators, fused with the trait-informed stability, prevent the system from drifting toward a negative emotion.

In the final utterance, distortion is a magnification pattern (exaggeration of conflict). Visual and acoustic signals may show signs of fatigue or distress, but textual and visual alone may lead to anger. Only with the distortion-sensitive DoT does the model apprehend that the speaker is venting sadness rather than outright hostility. Our CMTD, which incorporates DoT, thus correctly predicts sadness, whereas simpler fusion models fail by overemphasizing surface anger cues. In contrast, non-DoT methods could not produce correct labels.

6 Conclusion

This paper introduces a new multi-agent framework, CMTD, for multimodal emotion recognition in conversations. The framework takes into account traits and distortions to model static and dynamic states of emotions of interlocutors in a conversation. It allows the framework to reconstruct the cognitive model to assess emotional states in a deeper level. The framework is designed with the collaboration of agents to capture traits, distortions, visual and acoustic features. The final fusion agent synthesizes cognitive and multimodal indicators for the final prediction. Experimental results on MELD and IEMOCAP show that traits and distortions combined with multimodal features can enhance the emotional intelligence of LLMs. The framework is flexible and easy to adapt for advanced emotional AI systems. Future work will investigate the collaboration of agents in a more sophisticated way.

Limitations

Although achieving promising results, the proposed framework has the following limitations. First, all experiments were conducted on MELD and IEMOAP, two acted datasets collected from TV sitcoms and dialogues. In practice, there are some differences between acted and natural emotional speech conversations. It suggests that the proposed method should be tested on more general and natural conversation datasets. Second, the framework relies on a middle-frame strategy that extracts a representative image to capture visual cues of a speaker. While it is cost-effective and reduces latency, extracting a single static frame may discard temporal facial dynamics and micro-expressions that are often essential for recognizing fleeting emotions, e.g., disgust or surprise. Therefore, a video-based processing method should be considered with consideration of performance and latency. In addition, the latency of our framework is quite high compared to baselines. It is understandable that multi-agent frameworks usually require a longer time for reasoning. It can be acceptable in several applications such as asynchronous support (e.g., email responses, forum support), "typing..." delays in chatbots, or offline analysis of therapy sessions. Finally, the collaboration of agents is quite straightforward in which the final judge agent synthesizes information from other agents to make the final decision. It suggests that more sophisticated collaboration, e.g., multiple agents for the final judge, should be considered to improve the performance.

Ethics Statement

The authors confirm that this study does not have ethical issues. We note that distortion detection or the framework is not a substitute for screening or diagnosis of mental health or healthcare treatment. It only uses distortion detection as a step in emotional reasoning. The framework can also make wrong predictions of distortions. However, prediction is only used for emotion recognition in a computational linguistic task. We emphasize that distortion detection should be confirmed by psychologists in actual medical or mental healthcare applications. Thanks to shared prompts from InsiteOut and DoT authors, we can successfully run these methods. The code and datasets are collected from public GitHub links from the original papers. Experiments do not have specific parameter tuning to maintain a fair comparison among methods.

Acknowledgments

This work is part of the Cognition-aware AI project by Hajime Institute. This research is also supported by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2025.60.

References

Wei Ai, Yuntao Shou, Tao Meng, and Keqin Li. 2024. Der-gcn: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):4908–4921.

Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. 2025. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11418–11426.

Mehmet Baris Akccay and Kayhan Oguz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.

Muhammad Wasiq Alani, Jae-Hwan Kim, Hyoung-Gook Kim, and Jae-Yeol Woo. 2023. Hybrid multi-attention network for audio–visual emotion recognition through multimodal feature fusion. *Applied Sciences*, 13(7):4169.

Rawan AlMakinah, Andrea Norcini-Pala, Lindsey Disney, and M Abdullah Canbaz. 2024. Enhancing mental health support through human-ai collaboration: Toward secure and empathetic ai-enabled chatbots. *arXiv preprint arXiv:2410.02783*.

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

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Kishor Bhangale and Mohanaprasad Kothandaraman. 2023. Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. *Electronics*, 12(4):839.

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

Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association*

for Computational Linguistics: EMNLP 2023, pages 4295–4304.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.

Jaher Hassan Chowdhury, Sheela Ramanna, and Ketan Kotecha. 2025. Speech emotion recognition with lightweight deep neural ensemble model using handcrafted features. *Scientific Reports*, 15:11824.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661.

Joshua E Curtiss, Daniella S Levine, Ilana Ander, and Amanda W Baker. 2021. Cognitive-behavioral treatments for anxiety and stress-related disorders. *Focus*, 19(2):184–189.

Soumya Dutta and Sriram Ganapathy. 2025. Llm supervised pre-training for multimodal emotion recognition in conversations. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Paul Ekman and Wallace V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA.

Junwei Feng and Xueyan Fan. 2025. Cross-modal context fusion and adaptive graph convolutional network for multimodal conversational emotion recognition. *arXiv preprint arXiv:2501.15063*.

Julian D Ford, Damion J Grasso, Joan Levine, and Howard Tennen. 2018. Emotion regulation enhancement of cognitive behavior therapy for college student problem drinkers: A pilot randomized controlled trial. *Journal of Child & Adolescent Substance Abuse*, 27(1):47–58.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. 2025a. 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.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Yulin Wu, and Bingquan Liu. 2025b. Bemerc: Behavior-aware mllm-based framework for multimodal emotion recognition in conversation. *arXiv preprint arXiv:2503.23990*.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan. 2016. Zara the supergirl: An empathetic personality recognition system. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 87–91.

Thomas Goetz, Eva S Becker, Madeleine Bieg, Melanie M Keller, Anne C Frenzel, and Nathan C Hall. 2015. The glass half empty: How emotional exhaustion affects the state-trait discrepancy in self-reports of teaching emotions. *PloS one*, 10(9):e0137441.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Russell Harris. 2006. Embracing your demons: An overview of acceptance and commitment therapy. *Psychotherapy in Australia*, 12(4):70–6.

Steven C Hayes, Kirk D Strosahl, and Kelly G Wilson. 2011. *Acceptance and commitment therapy: The process and practice of mindful change*. Guilford press.

Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13018–13026.

Hajime Hotta, Huu-Loi Le, Manh-Cuong Phan, and Minh-Tien Nguyen. 2025. Metamo: Empowering large language models with psychological distortion detection for cognition-aware coaching. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 862–872.

Jian Huang, Yuanyuan Pu, Dongming Zhou, Jinde Cao, Jinjing Gu, Zhengpeng Zhao, and Dan Xu. 2024a. Dynamic hypergraph convolutional network for multimodal sentiment analysis. *Neurocomputing*, 565:126992.

Zilong Huang, Man-Wai Mak, and Kong Aik Lee. 2024b. Mm-nodeformer: Node transformer multimodal fusion for emotion recognition in conversation. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 4069–4073.

K. N. Javaras, M.J. Zonnevylle-Bender, J. Park, T. Uher, G. Turecki, J. Ferreira, C. M. Bulik, E. Koffel, M. Burmeister, and G. Turecki. 2012. Conscientiousness predicts greater recovery from negative emotion. *Personality and Individual Differences*, 52(5):587–592.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Juraj Kacur et al. 2021. On the speech properties and feature extraction methods in speech emotion recognition. *Sensors*, 21(5):1888.

Benjamin B. Lahey. 2009a. Public health significance of neuroticism. *American Psychologist*, 64(4):241–256.

Benjamin B. Lahey. 2009b. Public health significance of neuroticism. *American Psychologist*, 64(4):241–256.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

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

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

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001.

Zheng Li, Sujian Li, Dawei Zhu, Qilong Ma, and Weimin Xiong. 2025. Eerpd: Leveraging emotion and emotion regulation for improving personality detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7721–7734.

Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. Erd: A framework for improving llm reasoning for cognitive distortion classification. *arXiv preprint arXiv:2403.14255*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019a. Moel: Mixture of empathetic listeners. 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 121–132.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2019b. Caire: An empathetic neural chatbot. *arXiv preprint arXiv:1907.12108*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2023. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26:776–788.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Anh-Tuan Mai, Hai-Dang Kieu, Duc-Trong Le, et al. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Robert McCrae and Paul Costa. 1997. *Conceptions and Correlates of Openness to Experience*, pages 825–847.

Mikhail Mozikov, Nikita Severin, Maria Glushanina, Mikhail Baklashkin, Andrey Savchenko, and Ilya Makarov. 2024. Insideout: Unifying emotional llms to foster empathy. In *ECAI 2024*, pages 4499–4502. IOS Press.

Shrikanth S. Narayanan et al. 2009. Analysis of pausing behavior in spontaneous speech using real-time mri. *Journal of the Acoustical Society of America*.

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.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Erika L Rosenberg. 1998. Levels of analysis and the organization of affect. *Review of general psychology*, 2(3):247–270.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Björn Schuller et al. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*.

Jocelyn Shen, Daniella DiPaola, Safinah Ali, Maarten Sap, Hae Won Park, Cynthia Breazeal, et al. 2024a. Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: Comparative study. *JMIR Mental Health*, 11(1):e62679.

Jocelyn Shen, Joel Mire, Hae Won Park, Cynthia Breazeal, and Maarten Sap. 2024b. Heart-felt narratives: Tracing empathy and narrative style in personal stories with llms. *arXiv preprint arXiv:2405.17633*.

Jocelyn Shen, Maarten Sap, Pedro Colon-Hernandez, Hae Park, and Cynthia Breazeal. 2023. Modeling empathic similarity in personal narratives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6252.

Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. 2025. Coe: A clue of emotion framework for emotion recognition in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23548–23563.

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. 2025. Dynamic graph neural ode network for multi-modal emotion recognition in conversation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 256–268.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158.

Hiroki Tanioka, Tetsushi Ueta, and Masahiko Sano. 2024. Toward a dialogue system using a large language model to recognize user emotions with a camera. *arXiv preprint arXiv:2408.07982*.

Geng Tu, Tian Xie, Bin Liang, Hongpeng Wang, and Ruifeng Xu. 2024. Adaptive graph learning for multimodal conversational emotion detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19089–19097.

Cuong Tran Van, Thanh VT Tran, Van Nguyen, and Truong Son Hy. 2025. Effective context modeling framework for emotion recognition in conversations. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024a. Patient-: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797.

Yufeng Wang, Chao Chen, Zhou Yang, Shuhui Wang, and Xiangwen Liao. 2024b. Ctsm: Combining trait and state emotions for empathetic response model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4214–4225.

Joshua Wilt and {William R} Revelle. 2016. *Extraversion*. Oxford Univeristy Press.

Jesse H Wright, Gregory K Brown, Michael E Thase, and Monica Ramirez Basco. 2017. *Learning cognitive-behavior therapy: An illustrated guide*. American Psychiatric Pub.

Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025a. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. *arXiv preprint arXiv:2505.20511*.

Jiaqiang Wu, Xuandong Huang, Zhouan Zhu, and Shangfei Wang. 2025b. From traits to empathy: Personality-aware multimodal empathetic response generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8925–8938.

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.

Yang Yang, Xunde Dong, and Yupeng Qiang. 2025. Mse-adapter: A lightweight plugin endowing llms with the capability to perform multimodal sentiment analysis and emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25642–25650.

Zijun Yang, Zhen Li, Shi Zhou, Lifeng Zhang, and Seiichi Serikawa. 2024. Speech emotion recognition based on multi-feature speed rate and lstm. *Neurocomputing*, 601:128177.

Jamil Zaki. 2019. *The war for kindness: Building empathy in a fractured world*. Crown.

Xiaoheng Zhang and Yang Li. 2023. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13099–13110.

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

Juan Zheng, Susanne Lajoie, and Shan Li. 2023. Emotions in self-regulated learning: A critical literature review and meta-analysis. *Frontiers in Psychology*, 14:1137010.

Jianfeng Zhu, Ruoming Jin, and Karin G Coifman. 2025. Investigating large language models in inferring personality traits from user conversations. *arXiv preprint arXiv:2501.07532*.

Xiaofei Zhu, Jiawei Cheng, Zhou Yang, Zhuo Chen, Qingyang Wang, and Jianfeng Yao. 2024. Cmath: Cross-modality augmented transformer with hierarchical variational distillation for multimodal emotion recognition in conversation. *arXiv preprint arXiv:2411.10060*.

A Appendix

A.1 Trait Information

This section shows detailed information of traits mentioned in Section 3.3.1 **Neuroticism** is usually associated with heightened negative affect and vulnerability to sadness and anxiety (Lahey, 2009b). In contrast, **conscientiousness** is linked to enhanced self-regulation and faster recovery from negative emotion, thereby reducing impulsive anger (Javaras et al., 2012). **Extraversion** is correlated with positive affect and greater emotional expressivity, increasing the likelihood of joy or surprise in social interaction (Wilt and Revelle, 2016). Finally, **openness** reflects sensitivity to novelty and complex stimuli, which can manifest as surprise or mixed emotional states during unexpected conversational turns (McCrae and Costa, 1997).

A.2 Output observation

Figure 5 shows the output observation of all MERC methods in Table 2. Detailed discussion is shown in Section 5.3 (Output observation).

A.3 Prompts

Table 6 shows the prompts using in all MERC methods in Table 2. The DoT method is a version of Visual+DoT by removing the visual part.

Audio description

Ross 1: Stable moderate-low pitch, moderate steady energy, moderate rate, minimal pauses, clear, smooth voice.

Rachel 1: Stable moderately low pitch, consistent energy, moderate rate, deliberate pauses, clear, smooth voice.

Ross 2: Stable moderate-low pitch, consistent energy, moderate deliberate rate, zero pauses, clear, smooth voice.

Rachel 2: Normal-high pitch, upward inflection, consistent energy, very fast rate, zero pauses, clear voice.

Ross 3: Stable normal pitch, consistent energy, moderate deliberate rate, strategic pauses, clear, smooth voice.

Rachel 3: High fluctuating pitch, high forceful energy, fast clipped rate, infrequent pauses, strained, tight voice.

Figure 5: A conversation between Ross and Rachel from the MELD dataset. Corresponding gold labels and correct prediction are in blue. Distortions are in brown. Wrong predictions are in red.

Table 6: The prompts using in the baselines and proposed framework.

Method	Prompt
Visual agent	You will be given: (1) A spoken sentence, and (2) An image captured at the moment it was spoken. Your task: Analyse the speaker's facial expression in the image and provide a brief, precise description of their emotional or expressive state based solely on their facial cues, without interpreting the sentence content.
Text-only	Based on the provided conversation, identify the primary emotion expressed for each utterance in the conversation. The conversation is as follows: {conversation}
Text+Visual	Based on the provided conversation and the facial expression of each speaker when they speak, your task is identify the primary emotion expressed for each utterance in the conversation. The conversation is as follows: {conversation} Facial expression of each speaker in conversation: {facial expression}
Visual+DoT	Given a conversation, your task is to: 1) Finish a few diagnostic thought questions to analyse the thought patterns of the participants. Then, based on the diagnosis of thought analysis, 2) identify if there is cognitive distortion in the conversation; 3) Recognise the specific types of the cognitive distortion. The conversation is as follows: {conversation} Facial expression of each speaker: {facial expression} Here is the diagnose of thought questions: a) What is the situation? Find out the facts that are objective; what are the participants thinking or imagining? Find out the thoughts or opinions that are subjective. b) What makes the participants think the thought is true or not true? Find out the reasoning processes that support and do not support these thoughts. c) Why do the participants come up with such reasoning processes supporting the thought? What's the underlying cognition mode of it? Based on the diagnosis of thought, For each speech identify if there is cognitive distortion, specify the type of distortion. Here we consider the following common distortions: (followed by the descriptions and examples of all ten prompts included in the dataset metadata (Shreevastava and Foltz, 2021))
CMTD	
Personal traits	You are an expert psychologist specializing in personality analysis. Based on the Big Five Personality Traits model, you will evaluate the personality of two individuals the speaker and the listener based on their conversation in a given situation. Each response reflects different dimensions of personality traits. For each of the Big Five traits, consider the following facets: Conscientiousness: order, dutifulness, achievement striving, self-discipline, deliberation. Agreeableness: trust, straightforwardness, altruism, compliance, modesty, tendermindedness. Neuroticism: anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability. Openness: fantasy, aesthetics, values. Extraversion: warmth, gregariousness, assertiveness, excitement-seeking. Instructions: Read the given situation and conversation carefully. Evaluate both the speaker and the listener based on their dialogue. For each person, assign a score between 1 (Very Low) and 5 (Very High) for each of the Big Five traits. Scores may be rounded to the nearest tenth (e.g., 3.5). 1: Very Low - The response shows little to no alignment with the trait's facets. 2: Low - The response shows weak alignment with the trait's facets. 3: Moderate - The response shows some alignment but not strongly. 4: High - The response strongly aligns with the trait's facets. 5: Very High - The response shows exceptional alignment with the trait's facets.
Subjectivity assessment	What is the situation? Find out the facts that are objective; What is the speaker thinking or imagining? Find out the thoughts or opinions that are subjective.
Contrastive reasoning	What makes the speaker think the thought is true or not true? Find out the reasoning processes that support and do not support these thoughts.
Schema analysis	Why does the speaker come up with such reasoning process supporting the thought? What's the underlying cognition mode of it?
Distortion detection	Given a conversation, and the diagnosis of thoughts of the speakers in the conversation, your task is to: 1) Finish a few diagnostic thought questions to analyse the thought patterns of the participants. Then, based on the diagnosis of thought analysis, 2) identify if there is cognitive distortion in the conversation; 3) Recognise the specific types of the cognitive distortion.
Audio description	You are an expert in acoustic signal analysis. I will provide you with an audio input. Your task is to extract and return the following speech-related acoustic features from the audio: 1. Fundamental Frequency (F0): Describe the pitch characteristics of the voice across the entire audio. Note whether the pitch sounds high or low, how stable or fluctuating it is, any significant rises or falls in pitch contour, and how these patterns may reflect emotional tension, calmness, enthusiasm, or other affective cues. 2. Loudness / Energy: Describe the perceived loudness and vocal energy. Indicate whether the voice sounds strong, soft, steady, or variable, and comment on shifts in intensity that may signal emotional states such as excitement, anger, fatigue, or resignation. 3. Speaking Rate and Pauses: Describe the overall pace of speech and the use of pauses. Note whether the delivery feels fast, slow, or controlled; whether pauses seem frequent, hesitant, or purposeful; and explain how these pacing features might relate to emotions such as anxiety, confidence, sadness, or thoughtfulness. 4. Voice Quality Features: Describe the overall voice quality without using technical measurements. Comment on whether the voice sounds breathy, strained, hoarse, rough, tight, smooth, or clear. Highlight fluctuations or irregularities and interpret how these qualities might reflect emotional states such as stress, frustration, calmness, or vulnerability.
Fusion	Below is a conversation between multiple participants, along with corresponding analysis: {Conversation} Facial expression of each speaker: {facial expression} The diagnosis of thought: {diagnosis of thought} The cognitive distortion of each speaker: {distortion} The personality traits:{traits} The analysis of speech for each speaker is follows: {audio} Based on above analysis, your task is identify the primary emotion of utterance.