MindTracker: Unveiling Implicit Emotions in Long-Horizon Dialogues

Zhiqiang Gao[1] , **Jing Han**[1,4] , **Zhuochu Wang**[1] , **Shihao Gao**[1] , **Cheng Zhu**[1,3] , **Kehan Wang**[1] , **Huan Zhao**[1] , **Zixing Zhang**[1,2*]
[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]Shenzhen Research Institute, Hunan University, China
[3]Gannan University of Science and Technology, China
[4]Yuelushan Center for Industrial Innovation, China
{gaozhiqiang, jhan, zhuochuwang, shihaogao, zhucheng, wangkh, hzhao, zixingzhang}@hnu.edu.cn

Abstract

Affective computing has achieved notable success in recognizing explicit emotions from short, isolated dialogue segments. However, human emotions are often implicitly expressed, internally regulated, and dynamically evolve over extended interactions. Existing models struggle to disentangle internal emotional states from external expressions, and fail to capture the emotional inconsistency that emerges across long-horizon dialogues. To address this limitation, we introduce **Emotional Inconsistency Analysis** (EIA), a long-horizon task formulation for identifying and reasoning about discrepancies between implicit and explicit emotions in dialogue. To support this task, we construct the **MaskDialog** dataset carefully curated from TV-drama and large language models (LLMs) generated. We further propose two LLM-based baseline approaches, i.e., One-shot Self-consistent Inference and Cascaded Multi-step Inference, and conduct comprehensive analyses on dialogue construction strategies and inference behaviors. Experiments across six mainstream LLMs reveal that EIA remains highly challenging, particularly in modeling implicit emotional trajectories and cross-turn inconsistency. Overall, EIA reframes emotion understanding from short-term recognition to longitudinal, implicit emotion tracking, with implications for dialogue systems and human–computer interaction.

1 Introduction

There have been remarkable advances in affective computing in recent years. Research has progressed not only in fine-grained analysis [Lian *et al.*, 2025a; Zhang *et al.*, 2024] and multimodal fusion [Luo *et al.*, 2024], but also in addressing challenges related to cross-domain generalization and model explainability [Li *et al.*, 2024; Han *et al.*, 2025]. Current research in affective computing exhibits apparent limitations. A primary limitation is the focus on short-term, explicit emotions. Emotion, however, is a complex state encompassing two distinct layers: explicit and implicit. Explicit emotion refers to outwardly displayed affect, whereas implicit emotion denotes an individual's genuine, internal subjective experience. This separation originates in the brain's regulatory mechanisms, which continuously modulate outward expressions to meet situational demands, often leading to the concealment of genuine inner states. Consequently, methods that rely solely on transient external information fail to capture the deeper, often concealed, dynamics of authentic emotional experience [Schneeberger *et al.*, 2024].

Consider a character who smiles and says "I'm fine" after a traumatic event. Typical emotion models would likely classify this emotion as "joy", failing to recognize the facade of explicit positivity that masks underlying "pain". Consequently, such models are unable to see beyond this surface expression, much less infer the implicit emotion. This limitation becomes particularly problematic in long-horizon dialogues, where a character's genuine emotional state is continually shaped by accumulated experiences, evolving relationships, and internal psychological development.

In real social interactions, what people show is not always what they feel. As psychological and social theories suggest, individuals often regulate their displays (e. g., maintaining composure, concealing fear, signaling loyalty), which naturally gives rise to feigned explicit emotions that diverge from genuine implicit emotions [Goffman, 2023]. Thus, we contend that effective modeling of such cases requires must not only recognize surface expressions but also decipher the long-horizon cognitive history behind them. This necessitates a modeling framework that traces how implicit emotional states evolve through accumulated experiences.

Motivated by this gap, we propose **Emotional Inconsistency Analysis (EIA)**, a task requiring models to go beyond explicit emotion labels and reason about the gap between a character's implicit and explicit emotions. As shown in Figure 1, given a target character, current scene dialogue and prior context, EIA demands fine-grained emotional inconsistency analysis covering **Inconsistency Detection** (subtask-I), **Dual Emotion Recognition** (subtask-II), and **Trigger-Rationale Attribution** (subtask-III).

A major barrier to studying such a complex, multi-faceted task is the lack of suitable benchmarks. To address this, we constructed **MaskDialog**, a long-horizon dialogue dataset tailored for EIA. It comprises a curated set of instances, with

*Corresponding author

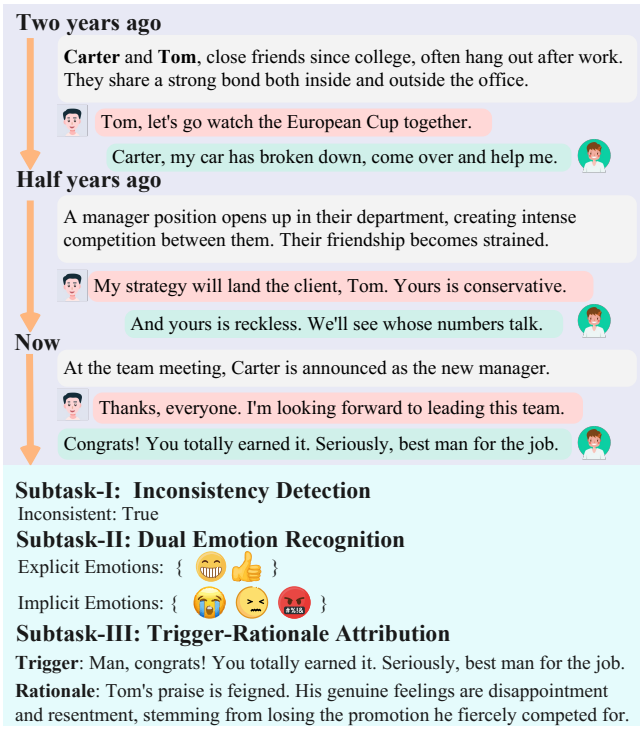

Figure 1: An example of **Emotional Inconsistency Analysis (EIA)**. Given a target character (*Tom*), preceding scenes, and the current scene, the task requires Inconsistency Detection, Dual Emotion Recognition, and Trigger-Rationale Attribution.

a balanced distribution of inconsistent and consistent cases. The instances were sourced from both real TV-drama scripts and LLM-generated scripts. To construct the dataset, we employed two LLM-assisted pre-annotation pipelines, **Script-to-Instance** and **LLM-to-Script**, both of which are followed by a unified human verification and quality-control procedure. To address EIA, we propose two LLM-based inference methods: **One-shot Self-consistent Inference** and **Cascaded Multi-step Inference**. We benchmarked both methods on **MaskDialog** across six mainstream LLMs and further conducted ablation studies to investigate the impact of long-horizon context construction and character-state modeling.

Our main contributions are summarized as follows:

- We formalize **EIA**, a novel long-horizon task for analyzing explicit and implicit emotional inconsistency in long-horizon dialogues. EIA comprises three subtasks: *Inconsistency Detection*, *Dual Emotion Recognition*, and *Trigger-Rationale Attribution*.
- We construct **MaskDialog**, a benchmark dataset for EIA comprising a curated set of instances with a balanced distribution of inconsistent and consistent cases, and provide two LLM-assisted pre-annotation pipelines: *Script-to-Instance* and *LLM-to-Script*.
- We develop two LLM-based baseline methods for EIA, **One-shot Self-consistent Inference** and **Cascaded Multi-step Inference**. Furthermore, we provide systematic analyses on the impact of long-horizon context construction strategies.

2 Related Work

In this section, we review the related research on emotion recognition and implicit emotion.

2.1 Emotion Recognition

Emotion recognition has been widely studied across text-, speech-, and vision-based settings and has gradually expanded from coarse categorical prediction to more context-aware and conversational emotion understanding [Gao *et al.*, 2025; Pereira *et al.*, 2024]. Recent surveys and reviews highlight the increasing importance of modeling context, speaker interaction, and pragmatic clues in dialogues [Plaza-del Arco *et al.*, 2024]. A representative research line is emotion recognition in conversation (ERC). Popular benchmarks include DailyDialog and multimodal datasets like MELD [Poria *et al.*, 2019; Li *et al.*, 2017], which provide turn-level emotion annotations for multi-turn conversations. Recent work further emphasizes contextual reasoning and psychological state modeling, yet dominant frameworks still treat emotion as a local target tied to external utterance or clip expressions. [Lee and Lee, 2022; Zhao *et al.*, 2022]. Mainstream ERC setups rarely make a distinction between a speaker's implicit emotion from explicit emotion [Stepputtis *et al.*, 2023; Brahman *et al.*, 2021; Anwar *et al.*, 2025]. Current research focuses on implicit and explicit emotions in short contexts, whereas emotional authenticity and long-horizon inconsistency reasoning remain largely underexplored in long-horizon dialogues [Li *et al.*, 2025b; Koga *et al.*, 2024; Xinyi *et al.*, 2024; Su *et al.*, 2025; Liu *et al.*, 2022].

2.2 Implicit Emotion Recognition

Related tasks also address the mismatch between feigned expressions and genuine intent or affect, including sarcasm detection, deception detection, metaphor understanding, and implicit emotion recognition. Sarcasm detection addresses pragmatic incongruity between literal content and intended meaning, often exhibiting sentiment polarity reversal. However, it is typically formulated at the sentence or short-context level. It also focuses on a narrow form of incongruity rather than explicit and implicit emotions [Farabi *et al.*, 2024; Yao *et al.*, 2025]. Deception detection aims to infer whether a speaker's statements or behaviors are deceptive. Its objective is usually veracity or credibility prediction, not the inference of paired implicit versus explicit emotions in long-horizon dialogues [Gupta *et al.*, 2019; Soldner *et al.*, 2019]. Implicit emotion recognition requires inferring emotion from context without explicit emotion words. Most frameworks, however, operate on short texts and a limited emotion taxonomy. They also fail to distinguish genuine from explicit emotion [Deng and Ren, 2021]. In contrast, EIA explicitly targets implicit and explicit emotion inconsistency analysis in long-horizon dialogues. It requires evidence-grounded reasoning over accumulated events and evolving character states.

3 Task Definition and Dataset Construction

This section formalizes the **EIA** task and metrics, then describes the **MaskDialog** construction process, which com-

bines LLM pre-annotation with human verification.

3.1 Task Definition: EIA

EIA moves beyond surface recognition by using an individual's long-horizon history to analyze their current emotional state. The core challenge is reasoning about the incongruence between explicit and implicit emotions, which involves three key tasks: detection, recognition, and causal analysis.

Problem Formulation. A long-horizon dialogue is segmented into an ordered sequence of scenes $\mathcal{S} = \{S_0, S_1, \ldots, S_t\}$, we denote the preceding scenes as $\mathcal{H}_t = \{S_0, \ldots, S_{t-1}\}$, and c denotes the target character. Each scene S_i contains a multi-speaker dialogue with utterances $S_i = \{u_{i,1}, \ldots, u_{i,n}\}$, where each utterance u is associated with a speaker identity $\mathrm{spk}(u)$.

At time step t, the input is:

$$x_t = \langle c,\ S_t,\ \mathcal{H}_t \rangle, \tag{1}$$

where c is the target character, S_t is the current scene, and $\mathcal{H}_t = \{S_0, \ldots, S_{t-1}\}$ is the long-horizon history.

Given x_t, we formulate EIA as the following mapping:

$$f_\theta(c, S_t, \mathcal{H}_t) \rightarrow \left(y_t^{\mathrm{inc}},\ E_t^{\mathrm{im}},\ E_t^{\mathrm{ex}},\ u_t^\star,\ r_t\right), \tag{2}$$

where $y_t^{\mathrm{inc}} \in \{0, 1\}$ indicates whether the character is inconsistent (1) or consistent (0). When $y_t^{\mathrm{inc}} = 0$, E_t^{im} (implicit emotional state) and E_t^{ex} (explicit emotional state) are treated as aligned, and $u_t^\star$ (the most indicative utterance) and r_t (evidence-grounded explanation) are omitted.

Subtask-I: Inconsistency Detection. The first subtask predicts whether the target character c exhibits implicit and explicit inconsistency in the current scene S_t:

$$y_t^{\mathrm{inc}} \in \{0, 1\}. \tag{3}$$

Subtask-II: Dual Emotion Recognition. If $y_t^{\mathrm{inc}} = 1$, two emotion-state descriptions for the target character are generated:

E_t^{ex} and E_t^{im}, both represented by open-vocabulary emotion words that are not restricted to a fixed label set.

Subtask-III: Trigger-Rationale Attribution. The third subtask localizes the most indicative utterance by c in the current scene, and generates an evidence-grounded explanation r_t that analyzes why the inconsistency occurs by integrating the current scene S_t and the long-horizon history $\mathcal{H}_t$:

$$r_t = \mathrm{Explain}(c, S_t, \mathcal{H}_t). \tag{4}$$

3.2 Evaluation Metrics

Subtask-I is a binary classification task where $y^{\mathrm{inc}} = 1$ denotes *inconsistent* and $y^{\mathrm{inc}} = 0$ denotes *consistent*. We adopt standard metrics (Accuracy, Precision, Recall, F_1), taking the inconsistent class F_1 as the primary metric and reporting Accuracy due to the balanced dataset.

Subtask-II evaluates predicted implicit ($\hat{E}_i^{\mathrm{im}}$) and explicit ($\hat{E}_i^{\mathrm{ex}}$) emotion sets against ground truths ($E_i^{\mathrm{im}}, E_i^{\mathrm{ex}}$). Exact matching is impractical for open-vocabulary terms; thus, we follow AffectGPT [Lian *et al.*, 2025a] and apply a canonicalization function $\phi(\cdot)$ with normalization, synonym merging, and taxonomic mapping before computing set-level metrics.

Specifically, suppose the dataset contains N samples. For each sample i and state $j \in \{implicit, explicit\}$, we compute precision, recall, and F_1 as:

$$P_s^j = \frac{1}{|N|} \sum_{i \in N} \frac{\left|\phi(E_i^j) \cap \phi(\hat{E}_i^j)\right|}{\left|\phi(\hat{E}_i^j)\right|}, \tag{5}$$

$$R_s^j = \frac{1}{|N|} \sum_{i \in N} \frac{\left|\phi(E_i^j) \cap \phi(\hat{E}_i^j)\right|}{\left|\phi(E_i^j)\right|}, \tag{6}$$

$$F1_s^j = \frac{2 P_s^j R_s^j}{P_s^j + R_s^j}. \tag{7}$$

We report the average state score as:

$$F1_s = \frac{1}{2}\left(F1_s^{im} + F1_s^{ex}\right). \tag{8}$$

Subtask-III is restricted to the inconsistent instances ($y_i^{\mathrm{inc}} = 1$), for which annotations are available. We denote this subset as $\mathcal{D}^+ = \{i \mid y_i^{\mathrm{inc}} = 1\}$.

Trigger Accuracy. We compute the exact match rate between the predicted trigger $\hat{u}_i^\star \in S_t$ and the ground truth $u_i^\star$ for all instances $i \in \mathcal{D}^+$:

$$\mathrm{EM}_i = \begin{cases} 1, & \hat{u}_i^\star = u_i^\star, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

$$\mathrm{Acc}_{\mathrm{Trig}} = \frac{1}{|\mathcal{D}^+|} \sum_{i \in \mathcal{D}^+} \mathrm{EM}_i. \tag{10}$$

Rationale score. We employ **GPT-5.2** as a judge to evaluate the predicted rationale $\hat{r}_i$ ($i \in \mathcal{D}^+$) against r_i and x_i based on *Relevance* (reasoning), *Accuracy* (faithfulness), and *Completeness* (coverage), the final score is averaged over $\mathcal{D}^+$:

$$\mathrm{Score}_{\mathrm{Rat}} = \frac{1}{|\mathcal{D}^+|} \sum i \in \mathcal{D}^+ \mathrm{LLMScore}(\hat{r}_i, r_i, x_i). \tag{11}$$

3.3 Dataset Construction

Figure 2 illustrates the construction of MaskDialog via two LLM-assisted pipelines: **Script-to-Instance** (mining existing TV scripts) and **LLM-to-Script** (generating synthetic scripts). Both pipelines undergo unified human verification.

Script-to-Instance Pipeline

MaskDialog is built from TV scripts, where each scene (narration and dialogue) forms an indexed instance. To manage script length, we use a sliding window of m scenes, maintaining history summaries ($S_0, \ldots, S_{t-1}$) and character cards, all updates restricted to preceding scenes to prevent data leakage. For each window, Gemini-2.5-pro processes the scenes, history, and character cards to generate evidence-based incongruence candidates, history summary for event tracking, and updated character cards. This process iterates through the entire script. To ensure dataset balance, we sample an equal number of **consistent** instances from scenes not flagged as inconsistent by the LLM.

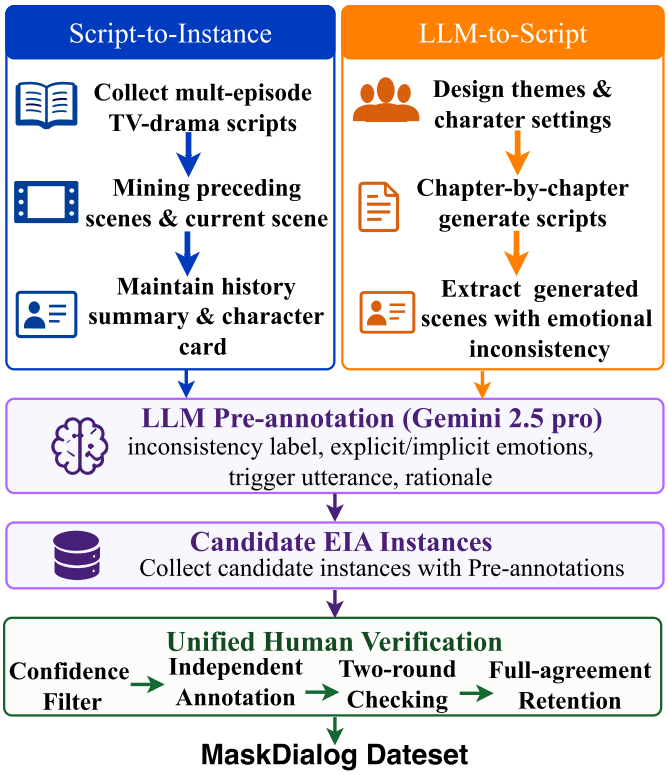

Figure 2: Overview of the MaskDialog construction pipeline. We build candidate EIA instances via two complementary pipelines: Script-to-Instance mines instances from TV drama scripts, while LLM-to-Script generates scripts iteratively. Both are processed by a shared LLM pre-annotation module, then refined through unified human verification to yield high-quality MaskDialog instances.

LLM-to-Script Augmentation

To enhance diversity with controllable patterns, we augment the dataset by prompting an LLM to generate thematic scripts. We first specify themes covering diverse social settings and instruct the LLM to create a high-level outline comprising a synopsis, chapters, and character profiles. This scaffold ensures narrative coherence for long-horizon reasoning.

The script is generated iteratively, chapter by chapter. For each chapter k, the LLM receives as input the outline, character profiles, and a running history summary m_{k-1} of all preceding chapters. It is then prompted to generate the dialogue scenes for chapter k. Subsequently, the summary is updated to m_k to incorporate the new content, and the process advances to chapter $k+1$. This continues until the entire script is complete.

During generation, we instruct the LLM to produce scenes with explicit–implicit emotional inconsistency for selected characters while preserving natural dialogue. For each scene, the LLM outputs preliminary EIA annotations as candidates for human verification. To maintain dataset balance, we also sample consistent scenes from generated scripts with aligned explicit and implicit emotional states. We adopt Gemini-2.5-pro to create these synthetic scripts and annotations.

Human annotation and label refinement

All LLM-generated candidates are subjected to the same rigorous human verification and quality-control pipeline as the script-based data. Specifically, we first filter candidates by the model-reported confidence. Subsequently, two annotators independently verify and refine the labels according to the same annotation guidelines. We conduct two rounds of checking and retain only instances with full annotator agreement, discarding disputed cases to ensure annotation quality. Finally, we adopted Cohen's Kappa for inter-annotator agreement, obtaining a high consistency score of 0.82.

Data Insights

We build MaskDialog, a curated dataset constructed via two complementary pipelines. Specifically designed to study explicit–implicit emotion divergence in long-horizon dialogues, MaskDialog provides fine-grained, open-vocabulary descriptions for both affective states. For inconsistent cases, it further supplies evidence-grounded rationales and trigger utterances, supporting deeper analysis beyond surface-level recognition. To enrich scenario diversity and approximate complex social interactions, the dataset covers a broad range of themes.

4 Methodology

As illustrated in Figure 3, we propose two LLM-based inference methods for EIA under the unified input setting. Both methods build a long-horizon reasoning context and maintain a dynamic profile memory for the target character.

4.1 Long-horizon Context and Dynamic Character Profile

Inference context. For a target character c at time step t, we construct the inference context from three components: (i) the dialogue of the current scene S_t, (ii) a set of relevant preceding scenes $\tilde{\mathcal{H}}_t$ drawn from $\{S_0, \ldots, S_{t-1}\}$, and (iii) a dynamic character profile $\mathcal{M}_t(c)$, which captures the character's evolving states.

Dynamic character profile. The dynamic character profile $\mathcal{M}_t(c)$ is structured along four dimensions: *Identity & Role*, *Personality & Traits*, *Motives & Goals*, and *Relationships*. To ensure it reflects recent developments, we update the profile using a local window of the most recent w scenes (we set w=5). At each update step, an LLM determines whether any dimension requires revision based on evidence from the new scenes; otherwise, the profile is retained. This evidence-driven update strategy enhances memory stability and mitigates temporal drift. Detailed prompts and formatting constraints are provided in the supplementary material.

4.2 One-shot Self-consistent Inference (OSI)

OSI performs joint inference for all three subtasks in a single prompt conditioned on $(S_t, \tilde{\mathcal{H}}_t, \mathcal{M}_t(c))$. To improve robustness, we employ a self-consistency mechanism with sampling and aggregation. Specifically, we generate N independent responses with a higher temperature. For Subtask-I, the inconsistency decision is aggregated via majority voting. We then filter the responses, retaining only those consistent with

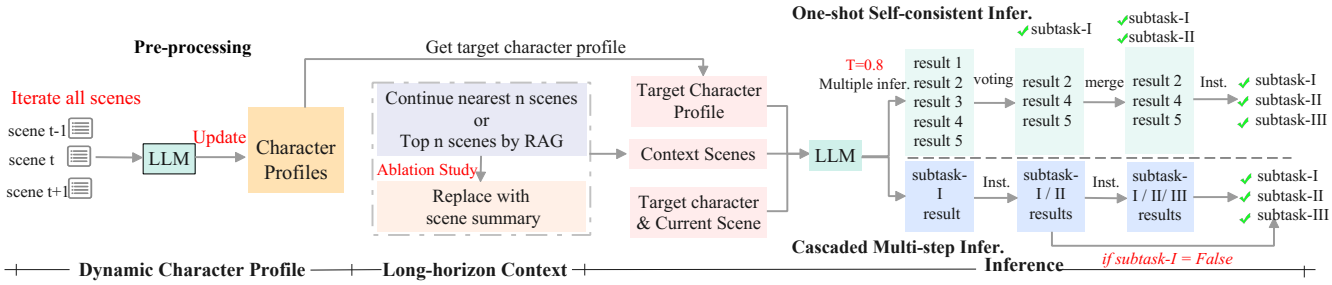

Figure 3: **Overview of Inconsistency Analysis (EIA) inference methods: One-shot Self-consistent Inference and Cascaded Multi-step Inference.** A shared context builder constructs the long-horizon input by combining the current scene, preceding scenes, and a dynamic character profile updated along the timeline.

the voted label. For Subtask-II, we merge and deduplicate the emotion words from the filtered responses to form more comprehensive sets for explicit and implicit emotions. For Subtask-III, we first determine the trigger utterance by majority vote. Subsequently, we prompt an LLM to select the most coherent rationale from the candidate set, taking into account the historical context and the current scene.

4.3 Cascaded Multi-step Inference (CMI)

CMI employs a three-step reasoning cascade under the same input setting, decomposing EIA into sequential specialized prompts:

Step 1. Determines whether the target character exhibits emotional inconsistency in the current scene.

Step 2. If inconsistency is detected, the model infers both explicit and implicit emotional states; otherwise, it infers a single shared state, setting $\hat{E}^{ex} = \hat{E}^{im}$.

Step 3. Triggered only when inconsistency is detected, this step localizes the trigger utterance and generates an evidence-grounded rationale.

This cascaded design isolates subtasks into dedicated prompts, thereby preventing the generation of unnecessary triggers and rationales for consistent cases. It is worth noting that CMI can also be combined with self-consistency by sampling multiple outputs at each step and aggregating the intermediate decisions. However, doing so would substantially increase the inference budget, since CMI already decomposes EIA into multiple sequential LLM calls. To ensure a roughly comparable computational cost between OSI and CMI, we do not apply self-consistency to CMI in our main experiments. This design allows us to compare the effect of joint versus cascaded task decomposition without confounding it with different inference budgets.

5 Experiments

5.1 Experimental Setup

Dataset. We conduct a benchmark on the full MaskDialog dataset using fixed prompts and hyperparameters, following the evaluation protocol and metrics outlined in Section 3.2.

Models and long-horizon inputs. We benchmark six mainstream LLMs for on EIA: GPT-5-mini, GPT-5.1-chat-latest, Gemini-3-Pro-Preview, Deepseek-v3.2 [Liu *et al.*, 2025], Qwen3-30B-A3B [Yang *et al.*, 2025], and Qwen-Plus. Dynamic character profiles are updated by Gemini-2.5-Pro ($w = 5$). For history, we compare two strategies: (i) **consecutive context**, which uses the most recent K preceding scenes, and (ii) **RAG-based context**, where we retrieve the top-20 relevant scenes using Text-embedding-v4 [Zhang *et al.*, 2025a] and re-rank them with Qwen3-Rerank. In an additional ablation study, we replace raw preceding scenes with compact summaries generated by Gemini-2.5-Pro to reduce context length.

Prompts and parameters. We employ a unified, model-agnostic prompting strategy across all LLMs. For OSI, we generate N=5 responses at temperature 0.8 and aggregate them via voting. For CMI, we use subtask-specific prompts with conservative decoding for stability.

5.2 Main Results

As shown in Figure 4, the two inference methods exhibit distinct performance profiles. OSI tends to obtain higher Accuracy on Subtask-I, which may be related to its self-consistency voting that favors more conservative decisions under uncertainty. In contrast, CMI shows a slightly higher F1. One possible explanation is that CMI appears more likely to predict inconsistency on ambiguous cases, which can increase recall and thereby improve F1.

Subtask-II remains challenging for both methods. When relying solely on unimodal (textual) clues, the dialogue and history often provide insufficient affective evidence for predicting nuanced emotional states. OSI exhibits a higher performance ceiling, as its multiple sampled generations yield a more diverse set of emotion words; merging these word lists can improve the recall of emotional states to some extent.

On Subtask-III, the flexibility of joint generation allows OSI to achieve higher peak performance on capable models. CMI, by contrast, delivers more consistent results across different LLMs by progressively constraining the reasoning flow and reducing the search space for trigger and rationale attribution. In summary, OSI favors peak performance potential, while CMI offers greater robustness and stability across model variants.

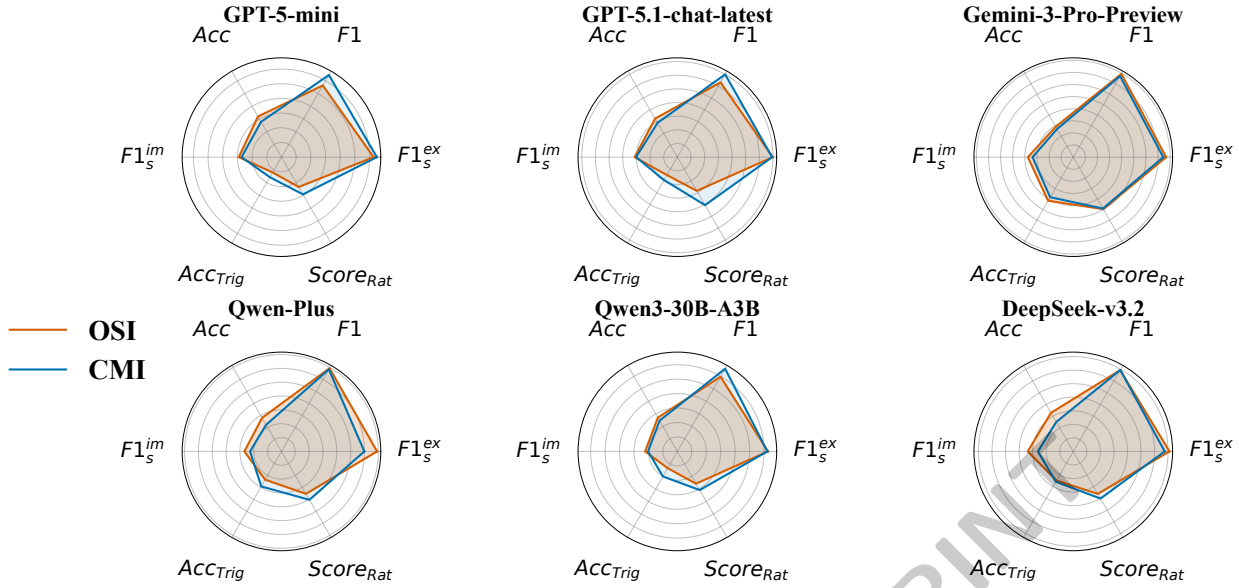

Figure 4: **Radar plots comparing One-shot Self-consistent Inference (OSI) and Cascaded Multi-step Inference (CMI) on MaskDialog.** Each subplot corresponds to one LLM and reports six evaluation metrics (Subtask-I: Acc and F1, Subtask-II: $F1_s^{ex}$ and $F1_s^{im}$, Subtask-III: Acc_{Trig} and $Score_{Rat}$).

Methods	Raw/Summary	Acc	F1	$F1_s^{ex}$	$F1_s^{im}$	Acc_{Trig}	$Score_{Rat}$
OSI + Consec	**Raw**	**.687**	**.662**	.303	.301	**.218**	**.332**
OSI + RAG	**Raw**	.675	.630	.300	.296	.187	.302
OSI + RAG + Re-ranking	**Raw**	.668	.624	.300	.294	.197	.304
OSI + Consec	**Summary**	.679	.640	**.307**	.301	.208	.307
OSI + RAG	**Summary**	.686	.651	.303	**.304**	.198	.315
OSI + RAG + Re-ranking	**Summary**	.679	.643	.305	.301	.206	.314
CMI + Consec	**Raw**	.664	.693	.259	.267	.247	.384
CMI + RAG	**Raw**	.667	.694	.259	.273	.236	.370
CMI + RAG + Re-ranking	**Raw**	.651	.681	.256	.274	.238	.371
CMI + Consec	**Summary**	**.681**	**.715**	**.265**	.278	**.284**	**.402**
CMI + RAG	**Summary**	.672	.706	.263	**.279**	.275	.385
CMI + RAG + Re-ranking	**Summary**	.670	.703	**.265**	.276	.271	.388

Table 1: Ablation study on long-horizon context construction for MaskDialog. We vary (i) history selection: using the consecutive scenes (Consec) versus RAG-based context (RAG), with an optional Re-ranking step (RAG+Re-ranking); and (ii) history summary: using raw preceding scenes (Raw) versus scene summaries (Summary). Reported scores are averaged across six evaluated LLMs (GPT-5-mini, GPT-5.1, Gemini-3-Pro-Preview, Qwen-Plus, Qwen3-30B-A3B, DeepSeek-v3.2).

5.3 Ablation and Analysis

To investigate the impact of long-horizon context construction on EIA performance, we perform controlled ablation studies. Unless stated otherwise, our default configuration employs consecutive history in its raw form. In each ablation run, we systematically vary a single factor while holding all other components constant.

Consecutive context vs. Retrieval context. We compare two strategies for long-horizon context construction. The consecutive strategy adopts the most recent 20 preceding scenes, while the retrieval strategy selects the top-20 relevant scenes from the full dialogue history. Retrieval is implemented via a hybrid retriever combining dense embeddings and a name-aware BM25 component, with character names in the current scene as the query. For each candidate scene, we compute a hybrid score as a weighted sum of normalized dense and BM25 scores, and keep the top-20 scenes. We further evaluate two RAG variants: (i) Embedding + BM25 and (ii) Embedding + BM25 + re-ranking. This design allows us to examine whether retrieval-based context selection benefits EIA more than simple consecutive context.

As shown in Table 1, **consecutive context shows a modest advantage over retrieval-based context in most settings**. The retrieval outputs reveal that approximately 65% of the top-20 retrieved candidates overlap with the most recent 20 consecutive scenes, suggesting that the hybrid retriever can recover semantically relevant content. However, semantic relevance alone is insufficient for EIA, which depends on temporal event coherence and character state evolution over long horizons. Consecutive context preserves chronological

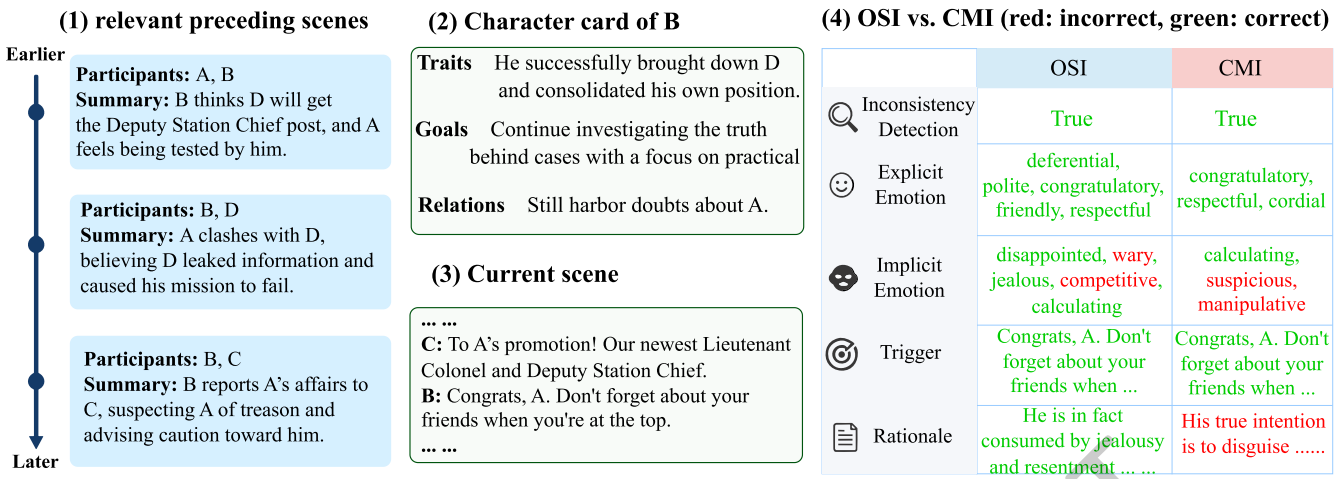

Figure 5: A representative example illustrating the difference between OSI and CMI on EIA. Red text indicates incorrect predictions, while green highlights correct results.

order and local discourse continuity, which may provide a more reliable basis for tracking emotional states and attributing their triggers.

By contrast, retrieval ranks scenes mainly by semantic similarity, which may disrupt narrative flow and make the original storyline harder to reconstruct. Although retrieved scenes are individually relevant, they may come from disparate time points, making coherent event-chain reconstruction more difficult. Consequently, the model must infer missing temporal and causal links from fragmented evidence, which may weaken inconsistency detection and attribution analysis. Furthermore, re-ranking yields only marginal gains for EIA. While it refines semantic relevance, re-ranking cannot restore chronological order; selected scenes remain temporally disordered and cannot form a coherent event chain.

Raw history vs. Summarized history. We evaluate a method that replaces raw historical dialogues (either consecutive or retrieved) with their summaries, aiming to reduce context length while preserving key information.

As shown in Table 1, summarized history appears to help OSI mainly in the RAG setting. Because retrieved history is fragmented and may contain irrelevant details, scene-level summaries can reduce noise and evidence conflicts across snippets. For consecutive history, however, raw dialogues already preserve temporal coherence and useful textual cues, so summarization brings limited gains and may discard informative details. CMI shows more consistent gains from summarization, possibly because its step-wise reasoning favors compact memory and relies less on raw historical text. Shorter inputs may also reduce redundancy and alleviate attention dilution at each stage.

Case Study. Figure 5 presents a representative example from MaskDialog. While both OSI and CMI correctly identify the instance as incongruent, their detailed outputs differ. OSI generates a more comprehensive set of implicit and explicit emotions by aggregating words across multiple samples. Furthermore, although both localize the trigger, OSI yields a superior rationale; its selection step produces an explanation that aligns more closely with the ground truth.

5.4 Limitations and Future Work

Our work has several limitations that highlight valuable future directions. First, MaskDialog is constructed from TV-drama and LLM-generated scripts, which may differ from naturally occurring interactions in language style, narrative structure, and emotional expression. Future work should extend EIA to more naturalistic dialogues. Second, current LLMs struggle to faithfully utilize long-horizon context, especially when critical evidence is sparse. Developing fine-grained memory mechanisms and robust evidence-tracking systems thus remains a priority [Zha *et al.*, 2025; Zhang *et al.*, 2025b]. Third, inferring genuine vs. feigned emotions from text-only dialogues is challenging due to the ambiguity of textual signals (e. g., missing prosodic and facial cues). Extending EIA to multimodal inputs is therefore a promising avenue [Li *et al.*, 2025a; Ai *et al.*, 2025; Xu *et al.*, 2025]. Finally, prompt-based inference remains brittle for complex, long-horizon reasoning. Future work could explore task decomposition, improved context construction, and learning from human feedback to enhance robustness and faithfulness [Lian *et al.*, 2025b].

6 Conclusion

This work takes a step toward modeling emotions as latent, long-evolving mind states rather than isolated surface signals. By framing Emotional Inconsistency Analysis (EIA) and introducing the MaskDialog-based benchmark, we highlight the limitations of LLMs in tracking implicit emotions over long-term interactions. Our findings suggest that robust emotion understanding requires integrating longitudinal reasoning, internal–external emotion disentanglement, and evidence-based inference. We hope this work encourages future research on emotionally aware dialogue systems that can faithfully reflect human internal states, advancing affective computing toward deeper cognitive and social intelligence.

Acknowledgements

The work leading to this research was supported by the National Natural Science Foundation of China under Grant No. U25A20447 (key program) and No. 62571184, the National Science and Technology Major Project of China under Grant 2026ZD1500300, the Science and Technology Innovation Program of Hunan Province under Grant No. 2025RC6003, the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515010112, and the Shenzhen Natural Science Foundation under Grant No. JCYJ20250604190534043.

References

[Ai *et al.*, 2025] Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11418–11426, Philadelphia, PA, USA, 2025.

[Anwar *et al.*, 2025] Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2838–2845, Atlanta, GA, USA, 2025.

[Brahman *et al.*, 2021] Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. Let your characters tell their story: A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic, 2021.

[Deng and Ren, 2021] Jiawen Deng and Fuji Ren. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67, 2021.

[Farabi *et al.*, 2024] Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 8020–8028, Jeju, Korea, 2024.

[Gao *et al.*, 2025] Zhiqiang Gao, Shihao Gao, Zixing Zhang, Yihao Guo, Hongyu Chen, and Jing Han. Structured prompting and llm ensembling for multimodal conversational aspect-based sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, page 14107–14113, Dublin, Ireland, 2025.

[Goffman, 2023] Erving Goffman. The presentation of self in everyday life. In *Social theory re-wired*, pages 450–459. Routledge, 2023.

[Gupta *et al.*, 2019] Viresh Gupta, Mohit Agarwal, Manik Arora, Tanmoy Chakraborty, Richa Singh, and Mayank Vatsa. Bag-of-lies: A multimodal dataset for deception detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (CVPR)*, pages 83–90, Long Beach, CA, USA, 2019.

[Han *et al.*, 2025] Jing Han, Zhiqiang Gao, Shihao Gao, Jialing Liu, Hongyu Chen, Zixing Zhang, and Björn W Schuller. Pioneering multimodal emotion recognition in the era of large models: From closed sets to open vocabularies. *arXiv preprint arXiv:2512.20938*, 2025.

[Koga *et al.*, 2024] Yurie Koga, Shunsuke Kando, and Yusuke Miyao. Forecasting implicit emotions elicited in conversations. In *Proceedings of the 17th International Natural Language Generation Conference (INLG)*, pages 145–152, Tokyo, Japan, 2024.

[Lee and Lee, 2022] Joosung Lee and Wooin Lee. 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 (NAACL)*, pages 5669–5679, Seattle, Washington, USA, 2022.

[Li *et al.*, 2017] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 986–995, Taipei, Taiwan, 2017.

[Li *et al.*, 2024] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, et al. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12468, Seattle, WA, USA, 2024.

[Li *et al.*, 2025a] Jiang Li, Xiaoping Wang, and Zhigang Zeng. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8786–8803, 2025.

[Li *et al.*, 2025b] Xinran Li, Yu Liu, Jiaqi Qiao, and Xiujuan Xu. Do llms feel? Teaching emotion recognition with prompts, retrieval, and curriculum learning. *arXiv preprint arXiv:2511.07061*, 2025.

[Lian *et al.*, 2025a] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vancouver, Canada, 2025.

[Lian *et al.*, 2025b] Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, and Xiaobai Li. AffectGPT-R1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2508.01318*, 2025.

[Liu *et al.*, 2022] Yuchen Liu, Jinming Zhao, Jingwen Hu, Ruichen Li, and Qin Jin. DialogueEIN: Emotion interaction network for dialogue affective analysis. In *Proceedings of the 29th International Conference on Computational Linguistics (ACL)*, pages 684–693, Gyeongju, Republic of Korea, 2022.

[Liu *et al.*, 2025] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[Luo *et al.*, 2024] Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. PanoSent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 7667–7676, Melbourne, Australia, 2024.

[Pereira *et al.*, 2024] Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. Deep emotion recognition in textual conversations: A survey. *Artificial Intelligence Review*, 58(1):10, 2024.

[Plaza-del Arco *et al.*, 2024] Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 5696–5710, Torino, Italia, 2024.

[Poria *et al.*, 2019] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–536, Florence, Italy, 2019.

[Schneeberger *et al.*, 2024] Tanja Schneeberger, Mirella Hladký, Ann-Kristin Thurner, Jana Volkert, Alexander Heimerl, Tobias Baur, Elisabeth André, and Patrick Gebhard. The deep method: Towards computational modeling of the social emotion shame driven by theory, introspection, and social signals. *IEEE Transactions on Affective Computing*, 15(2):417–432, 2024.

[Soldner *et al.*, 2019] Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1768–1777, Minneapolis, MN, USA, 2019.

[Stepputtis *et al.*, 2023] Simon Stepputtis, Joseph P Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Lewis, and Katia Sycara. Long-horizon dialogue understanding for role identification in the game of avalon with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11193–11208, Toronto, Canada, 2023.

[Su *et al.*, 2025] Yuting Su, Yichen Wei, Weizhi Nie, Sicheng Zhao, and Anan Liu. Dynamic causal disentanglement model for dialogue emotion detection. *IEEE Transactions on Affective Computing*, 16(1):1–14, 2025.

[Xinyi *et al.*, 2024] Wu Xinyi, Xu Changqing, Li Nan, Su Rongfeng, Wang Lan, and Yan Nan. Depression enhances internal inconsistency between spoken and semantic emotion: Evidence from the analysis of emotion expression in conversation. In *Proceedings of the 25th Annual Conference of the International Speech Communication Association (Interspeech)*, 2024.

[Xu *et al.*, 2025] Weixiang Xu, Zhongren Dong, Runming Wang, Xinzhou Xu, and Zixing Zhang. Gatem 2 former: Gated feature selection and expert modeling in multimodal emotion recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Hyderabad, India, 2025.

[Yang *et al.*, 2025] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[Yao *et al.*, 2025] Ben Yao, Yazhou Zhang, Qiuchi Li, and Jing Qin. Is sarcasm detection a step-by-step reasoning process in large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 25651–25659, Philadelphia, PA, USA, 2025.

[Zha *et al.*, 2025] Xupeng Zha, Huan Zhao, Guanghui Ye, and Zixing Zhang. Dual-view learning for conversational emotion recognition through context and emotion-shift modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 25823–25831, Philadelphia, PA, USA, 2025.

[Zhang *et al.*, 2024] Zixing Zhang, Zhongren Dong, Zhiqiang Gao, Shihao Gao, Donghao Wang, Ciqiang Chen, Yuhan Nie, and Huan Zhao. Open vocabulary emotion prediction based on large multimodal models. In *Proceedings of the International Workshop on Multimodal and Responsible Affective Computing (MRAC)*, pages 99–103, New York, NY, USA, 2024.

[Zhang *et al.*, 2025a] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[Zhang *et al.*, 2025b] Zeyu Zhang, Quanyu Dai, Xu Chen, Rui Li, Zhongyang Li, and Zhenhua Dong. MemEngine: A unified and modular library for developing advanced memory of LLM-based agents. In *Proceedings of the ACM on Web Conference (WWW)*, pages 821–824, Sydney, Australia, 2025.

[Zhao *et al.*, 2022] Weixiang Zhao, Yanyan Zhao, and Xin Lu. CauAIN: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4524–4530, Vienna, Austria, 2022.