ERCThinker: Fast-Slow Thinking for Emotion Recognition in Conversation

Yumeng Fu[1], Weitao Huang[1], Junjie Wu[2],
Hao Teng[1], Shouduo Shang[1], Meishan Zhang[3], Bingquan Liu[1]*,
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China,
[2]School of Computer Science and Technology, Soochow University, Suzhou, China,
[3]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China,
24b303004@stu.hit.edu.cn, huangweitao@insun.hit.edu.cn, 20224027010@stu.suda.edu.cn
25s103237@stu.hit.edu.cn, 24s103317@stu.hit.edu.cn, mason.zms@gmail.com, liubq@hit.edu.cn

Abstract

Emotion Recognition in Conversation (ERC) aims to identify the emotional states of speakers in conversations. Existing ERC methods perform either fast thinking or slow thinking for emotion predictions. The former lacks interpretability of emotion predictions, and the latter focuses on emotion analysis at shallow semantics. Such insufficient reasoning chains fall short in capturing deep semantics within conversations. To address these limitations, we propose ERCThinker, a Fast-Slow thinking framework for the task of ERC. First, we design different thinking strategies with fine-grained emotion reasoning chains to capture deep semantics that contain topic, discourse structure, speaker characteristic, scene, and emotion shift. Second, we develop an adaptive thinking mechanism in both strategy-level and utterance-level, guiding the model to dynamically perform thinking switching across various scenarios. Furthermore, we utilize Agent-as-Judge to score reasoning chains as reward signals for more accurate emotion predictions. To support training, we construct EmotionCue-CoT, the emotion reasoning dataset with supervision in both explanation and judgment. Extensive experiments on various ERC benchmark datasets demonstrate that ERCThinker achieves state-of-the-art performance in both explanation and judgment, making progress in the realm of ERC.

1 Introduction

Emotion Recognition in Conversation (ERC) involves identifying the speaker's emotional state in a conversation. This task is crucial for machines to understand human-machine interactions in the real world (Cowie et al., 2001), which is applied in various potential applications, such as virtual assistants (Chatterjee et al., 2021), customer service (Shen et al., 2025a) and mental health monitoring (Zhang and Tan, 2025). Faced with complex conversation scenarios, accurately identifying and reasoning emotions is urgently needed.

Discriminative methods typically formulate the ERC task as a classification problem, which are roughly classified into four categories, including Transformer-based methods (Chudasama et al., 2022; Liu et al., 2023), Recurrent-based methods (Lei et al., 2023b; Hu et al., 2023), GNN-based methods (Shen et al., 2021; Li et al., 2024), and LLMs-assisted methods (Lei et al., 2023a; Xue et al., 2024; Li et al., 2025). Despite their impressive performance, these methods provide emotion labels without explicit rationales. This limits the interpretability of emotion predictions.

In contrast to discriminative ERC methods that follow fast thinking (Zheng et al., 2025) to provide intuitive responses, recent generative ERC methods adopt slow thinking (Deng et al., 2025) to generate emotion reasoning chains and then derive emotion predictions (Lian et al., 2025). Nevertheless, these reasoning chains focus on shallow semantics (e.g., facial actions or vocal prosody) and under-exploit deep semantics arising from conversational interactions, such as topic, discourse structure, speaker characteristic, scene, and emotion shift. Therefore, we focus on equipping ERC models with diverse thinking strategies that can produce fine-grained emotion reasoning chains for both explanation and judgment.

To address these limitations, we propose ERCThinker, a novel Fast-Slow thinking framework for emotion recognition in conversation. Specifically, inspired by human cognitive behavior in conversational emotion recognition, we design two thinking strategies consisting of fast thinking and slow thinking. The former refers to the direct answers related to emotion labels, and the latter performs fine-grained emotion reasoning chains for predicting emotions. After then, to make the model switch thinking strategies in coping with different conversation scenarios, we introduce an adaptive thinking

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39344–39359

mechanism to incorporate both strategy-level and utterance-level information into advantage estimation. Finally, we introduce Agent-as-Judge that scores reasoning chains as reward signals, effectively guiding the model to capture the trajectories of emotion triggering. Additionally, we construct high-quality chain-of-thought data via an automated pipeline that aggregates multi-dimensional emotional cues (e.g., topic, speaker characteristics, scene, discourse structure, and emotion shifts) from multiple open-source datasets, and retains LLM-generated rationales with high self-reported confidence. Overall, ERCThinker significantly improves reasoning efficiency and token utilization while preserving strong task performance.

The primary contributions of this work can be outlined as follows:

- We propose ERCThinker, a novel fast-slow thinking framework for ERC. ERCThinker uses an adaptive thinking mechanism to perform thinking strategy switching across various conversation scenarios, and introduces Agent-as-Judge to refine emotion reasoning chains for accurate emotion predictions.
- We construct EmotionCueCoT, an emotional chain-of-thought that integrates multi-aspect emotional clues, including topic, speaker characteristics, scene, discourse structure, and emotion shifts to enhance the interpretability and coherence of emotion reasoning.
- Extensive experiments on different ERC benchmark datasets demonstrate that the proposed ERCThinker achieves state-of-the-art results across evaluation metrics.

2 Related Works

2.1 Emotion Recognition in Conversation

In the era of large language models, emotion recognition in conversation have increasingly adopted instruction tuning paradigms (Lei et al., 2023a; Li et al., 2025; Xue et al., 2024). Under such settings, models are driven by instructions to capture emotional cues in dialogue, which are then explicitly injected into predefined templates to generate emotion predictions (Fu et al., 2025; Tu et al., 2024; Shen et al., 2025b).

However, these studies remain confined to a pipeline of "clue extraction–template concatenation–label output." Emotional cues are merely exploited in a simplistic manner, lacking a mechanism that couples these cues with an interpretable emotion reasoning process. As a result, it is difficult to effectively bridge the gap between "observed cues" and the "rationale for emotion judgments," which limits the model's ability to perceive deep semantic structures in conversational text, thereby affecting emotion prediction.

2.2 Reasoning Models

Building on the dual-process view, System 1 supports fast and intuitive judgments, whereas System 2 enables slow and deliberative reasoning (Evans, 1984; Kahneman, 2003). Recent reasoning large language models, such as DeepSeek-R1(Guo et al., 2025) and OpenAI o1 (Jaech et al., 2024), suggest that explicitly strengthening reasoning can improve performance across diverse tasks. These tasks include mathematical reasoning(Qiao et al., 2025; Yang et al., 2025a), code generation(Yue et al., 2025; Yang et al., 2025b), knowledge-intensive question answering(Zhang and Zhao, 2025).

Despite these advances, transferring such capabilities to conversational emotion understanding remains challenging, as this task critically depends on complex semantic structures in textual dialogues. In particular, effective emotion recognition requires the model to adaptively determine when fast intuitive inference suffices and when deeper deliberative reasoning is necessary, based on the underlying semantic content of the conversation. Motivated by this gap, we investigate conversational emotion recognition from a fast–slow thinking paradigm, enabling adaptive switching between System-1-style efficient inference and System-2-style deep reasoning. This design allows the model to capture subtle emotional cues more effectively while maintaining controlled inference costs.

3 Method

We propose the ERCThinker framework to achieve interpretable emotion recognition in conversations via adaptive reasoning. We first construct EmotionCueCoT, a reasoning dataset with semantic-level emotional cues, and build a Llama3-8B–based baseline by supervised fine-tuning with a fast–slow thinking cold start. We then introduce reinforcement learning with an adaptive strategy formulation (strategy-level and utterance-level) to optimize emotion-reasoning strategies and enhance reasoning capability. Finally, we adopt an dynamically agent-based score mechanism to direct the model

to the right reasoning process.

3.1 EmotionCueCoT Dataset

3.1.1 Dataset Collection

Existing emotion reasoning datasets are limited in their ability to capture semantic-level emotional cues. To enable model to better capture such cues and to enhance interpretable dialogue emotion recognition, we construct EmotionCueCoT, an emotion reasoning dataset with multi-dimensional semantic cues. It contains 25,086 reasoning instances. Each instance is annotated with an emotion label from seven common categories: happy, sad, neutral, angry, surprise, fear, and disgust. The dataset is derived from IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2018), and EmoryNLP (Zahiri and Choi, 2018). By unifying three datasets into a consistent dialogue schema, our composite dataset spans diverse speakers and conversational settings. It provides rich emotion cues and pairs utterance-level emotion labels with emotion-centric reasoning, enabling text-based ERC models to output both predictions and cue-aware explanations.

3.1.2 Emotional Cues Extraction

As illustrated in Figure 1, accurate emotion understanding requires the integration of multiple semantic emotional cues that are tightly grounded in the conversational context. In this work, we focus on several key cue types, including topic, speaker characteristics, scene, discourse structure, and emotion shifts across dialogue. These cues capture complementary aspects of conversational semantics and jointly support fine-grained emotion reasoning. We adopt established and widely used processing pipelines to extract each cue and incorporate them into the model in a unified manner. To maintain clarity in the main text, we present only a high-level description here, while detailed implementations of cue extraction and modeling are provided in the Appendix A.2.

3.1.3 The pipeline of dataset construction.

Compared with traditional manual annotation, automated annotation reduces costs and enables the construction of large-scale datasets. Moreover, it is not influenced by human bias. We designed a set of templates that provide five types of emotional cues to a large language model as prompt inputs to DeepSeek-R1 (Guo et al., 2025), thereby inducing a step-by-step reasoning process. During output generation, we perform a self-evaluation (Shinn et al., 2023) step to evaluate the model's initial output. The model assigns a confidence score ranging from 0 to 5 to its own reasoning results. If the confidence score falls below 4, the chain of thought will be regenerated. (Prompt is in Figure 5)

3.2 Cold-start SFT for Fast–Slow Thinking.

To enable stable learning of fast- versus slow-thinking output formats during the reinforcement learning stage, we first perform a supervised fast-slow thinking warm-up.

Each sample is scored along multiple dimensions, including slow-thinking criteria, fast-thinking criteria, model confidence, and model emotion prediction accuracy. The slow-thinking criteria capture higher expressive and reasoning complexity, covering expression length, causal/contrastive connectives (e.g., *because*, *but*, *so*), structural complexity (number of sentences/clauses), coreference (e.g., *he/she*, *this/that*), metacognitive markers (e.g., *I thought carefully*, *maybe*, *reflection*, *calm down*, *perspective taking*), and the length of utterance. In contrast, the fast-thinking criteria characterize more direct, affective, and terse expressions, including the density of strong emotion words, punctuation (e.g., *exclamation/question marks*), emojis, discourse particles/interjections, and extremely short text. The formula for determining whether each sample requires fast or slow thinking is as follows (Prompts are in Figs. 6– 8):

$$S_i = C_i^{\text{slow}} - C_i^{\text{fast}} - \text{Acc}_i, \qquad (1)$$

where C_i^{slow} and C_i^{fast} are the slow-/fast-thinking scores, and Acc_i is the prediction correctness.

We select the top 40% of samples ranked by the final score as slow-thinking data. We represent the slow-thinking format as "<think>...</think> <answer>emotion label</answer>", and the fast-thinking format as "<think> There is no thinking process"</think> <answer>emotion label</answer>". Mixing the two formats yields the final cold-start dataset, which is used to perform supervised fine-tuning. This stage teaches the model to (1) faithfully follow fast–slow output constraints and (2) acquire an initial policy prior for switching between fast and slow modes, thereby enabling more stable and controllable optimization in the reinforcement learning stage.

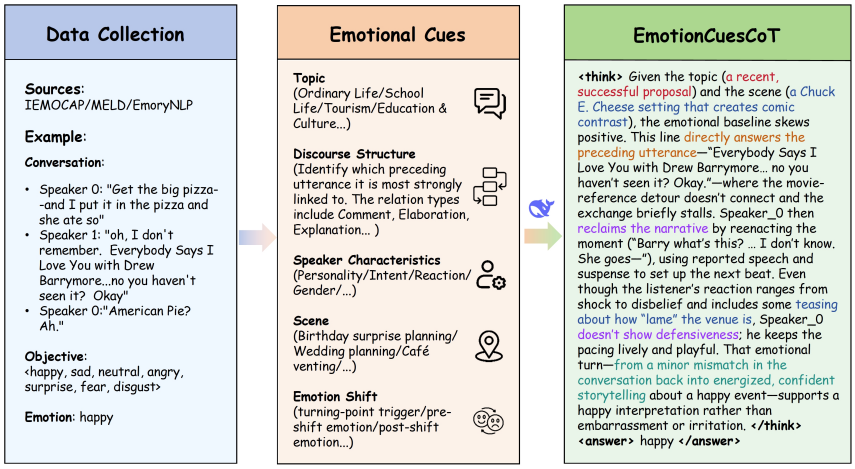

Figure 1: The pipeline of EmotionCueCoT dataset construction.

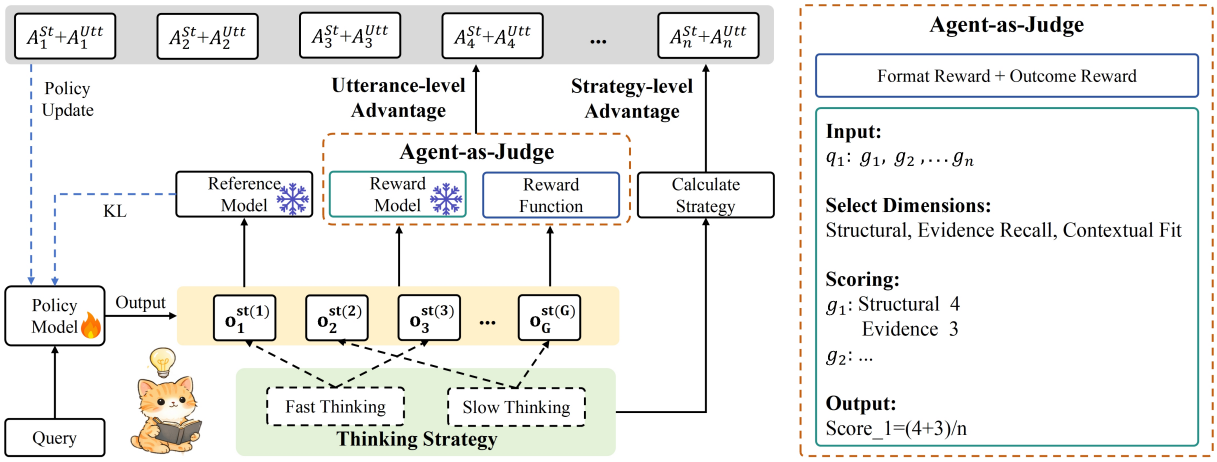

Figure 2: The overview framework of ERCThinker.

3.3 Fast-slow Thinking for Emotion Reasoning

In conversational emotion recognition, fast–slow thinking is necessary because some emotional cues are explicit and can be handled by immediate strategy-based responses. However, others are implicit, context-dependent, requiring deeper inference and self-verification. We combine fast and slow thinking to not only reduce unnecessary computational overhead but also enhance emotion understanding capability.

Strategy-level and Utterance-level Advantage To enable both the ability to select appropriate thinking strategies and the capacity to accurately analyze the emotion expressed in each utterance within those strategies, we propose an *Adaptive Advantage* mechanism that enables the model to adaptively choose between two thinking strategies (fast and slow). The proposed method decomposes the advantage signal at two levels, namely the *strategy-level* and the *utterance-level*, and injects them into a unified policy-optimization objective built upon GRPO. The objective is defined as follows:

$$\max_{\pi_\theta} \mathbb{E}_{o\sim\pi_{\text{old}}}\left[\sum_{j=1}^{G}\frac{\pi_\theta(o_j)}{\pi_{\text{old}}(o_j)}\left(A^{\text{St}}_{j,t}+A^{\text{Utt}}_{j,t}\right)\right] - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \tag{2}$$

At the strategy level, we introduce a *strategy advantage* to quantify the relative utility of different inference strategies. For each strategy $k \in \{\text{fast}, \text{slow}\}$, let St_k denote the set of rollout samples generated under strategy k within the current rollout group. We compute its average reward and average output length as:

$$\bar{r}_{St_k} = \frac{1}{|St_k|}\sum_{i\in St_k} r_i, \tag{3}$$

$$\bar{l}_{St_k} = \frac{1}{|St_k|}\sum_{i\in St_k} l_i. \tag{4}$$

Here, $|St_k|$ is the number of rollouts in St_k, r_i is

the reward assigned to the i-th rollout, and l_i is the generated token length of that rollout.

Based on these statistics, the strategy advantage is defined in a conditional form:

$$A_{i,t}^{\mathrm{St}} = \frac{\bar{r}_{\mathrm{St}(i)} - \mathrm{mean}(\{\bar{r}_{\mathrm{fast}}, \bar{r}_{\mathrm{slow}}\})}{\mathrm{std}(\{\bar{r}_{\mathrm{fast}}, \bar{r}_{\mathrm{slow}}\})} + \beta\left[-\tanh\left(\frac{\bar{l}_{\mathrm{St}(i)} - \mathrm{mean}(\{\bar{l}_{\mathrm{fast}}, \bar{l}_{\mathrm{slow}}\})}{\mathrm{std}(\{\bar{l}_{\mathrm{fast}}, \bar{l}_{\mathrm{slow}}\})}\right)\right], \tag{5}$$

$$\beta = \begin{cases} 1, & \mathrm{std}\left(\{r_1^{\mathrm{St}(1)}, \ldots, r_G^{\mathrm{St}(G)}\}\right) = 0, \\ 0, & \mathrm{std}\left(\{r_1^{\mathrm{St}(1)}, \ldots, r_G^{\mathrm{St}(G)}\}\right) > 0. \end{cases} \tag{6}$$

This definition reflects an adaptive trade-off between effectiveness and efficiency: when `fast` and `slow` exhibit a noticeable gap in sentiment prediction accuracy, the model is guided to select the strategy with higher average reward; when their prediction quality is similar, the length-based term encourages the more efficient `fast` strategy.

Specifically, the utterance advantage measures, within a rollout group, how an individual instance performs relative to the group average. It is defined as:

$$A_{i,t}^{\mathrm{Utt}} = \frac{r_i^{St(i)} - \mathrm{mean}\left(\left\{r_1^{St(1)}, \ldots, r_G^{St(G)}\right\}\right)}{\mathrm{std}\left(\left\{r_1^{St(1)}, \ldots, r_G^{St(G)}\right\}\right)} \tag{7}$$

Finally, the token-level update uses the combined advantage:

$$A_{i,t}^{\mathrm{total}} = A_{i,t}^{St} + A_{i,t}^{\mathrm{Utt}} \tag{8}$$

thereby enabling strategy selection and utterance emotion prediction optimization to proceed jointly under a unified learning framework. Strategy-level advantage allows the LLM to dynamically choose the most suitable thinking strategy by balancing efficiency (shorter reasoning when performance is similar) and effectiveness (higher-reward modes when performance differs), with a tanh-bounded advantage to stabilize training under large length variations.

Process Reward with Agent. Our reward function consists of three components: the answer reward r_i^a, the format reward r_i^f, and the process reward r_i^p. The overall reward r_i is as follows:

$$r_i = \begin{cases} \alpha_a r_i^a + \alpha_f r_i^f, & \text{if } \mathrm{St}(i) = \texttt{fast}, \\ \alpha_a r_i^a + \alpha_f r_i^f + \alpha_p r_i^p, & \text{if } \mathrm{St}(i) = \texttt{slow}. \end{cases} \tag{9}$$

Notably, conversational inputs vary substantially in semantic complexity and structural dependency.

Table 1: Evaluation Indicators and Dependency Hierarchy. The description is in Appendix.

ID	Parent	Metric Type	Abbreviation
1	—	Structural	ST.
2	1	Evidence Recall	ER.
3	2	Factual Alignment	FA.
4	2	Conflict Check	CC.
5	2	Logical Structure	LS.
6	3	Comprehensive Coverage	CoC.
7	3	Fine-Grained Coverage	FC.
8	5	Answer Consistency	AC.
9	5	Language Quality	LQ.
10	6	Conciseness	CO.
11	7	Logical Rigor	LR.
12	4	Intensity Match	IM.
13	4	Internal Consistency	IC.
14	4	Contextual Fit	CF.

As a result, a single, fixed reasoning trajectory is not universally suitable. Instead, the model should select an appropriate reasoning path conditioned on the current context. As illustrated in Figure 2, we adaptively activate different groups of evaluation dimensions to guide and assess the reasoning process.

Concretely, we adopt a two-stage, agent-based evaluation pipeline consisting of dimension selection and scoring. First, we synthesize prior criteria and, tailored to the characteristics of conversational emotion recognition, define a set of 14 interdependent evaluation dimensions organized into a dependency hierarchy (Table 1). The hierarchy enforces that selecting a child dimension requires prior selection of its parent, which supports a logically consistent and progressively refined assessment. Next, for each input instance, we prompt an evaluator to select a context-relevant subset of dimensions. The generated reasoning is then scored on each selected dimension using a 0–5 scale, and the overall quality score is computed as the average of the selected-dimension scores. This adaptive design alleviates the limitations of fixed-dimension evaluation and improves assessment effectiveness (Prompt is in Figure 4).

Specifically, when cues are explicit and ambiguity is low, we adopt a lightweight path, prioritizing Concisenness and Language Quality to avoid unnecessary verbosity and improve readability. When the dialogue exhibits emotional conflicts or pragmatic reversals, we switch to a verification path, placing greater emphasis on Conflict Check, Internal Consistency, and Contextual Fit to suppress misjudgments and ensure alignment with the conversational context. When evidence is dispersed across

Models	Reasoning	Emotion Recognition ↑				Emotion Reasoning ↑					
		IEMOCAP	MELD	EmoryNLP	Avg.	ST.	AC.	FA	CC.	CoC.	Avg.
General Large Language Models											
Llama3-8B (Grattafiori et al., 2024)	✓	55.94	53.69	42.19	50.61	4.02	2.35	2.79	1.68	2.95	2.76
Qwen2.5-7B (Yang et al., 2024)	✓	57.66	57.34	45.38	53.46	3.96	2.51	2.85	1.57	2.84	2.75
Ministral-3-8B-Reasoning (Rastogi et al., 2025)	✓	60.35	57.64	44.53	54.17	4.26	2.71	3.16	1.65	3.11	2.98
GLM-4.1V-9B-Thinking (Hong et al., 2025)	✓	55.30	62.17	43.14	53.54	3.72	2.80	3.22	1.60	2.89	2.85
GPT-4o (Hurst et al., 2024)	✓	59.97	61.36	46.08	55.80	4.15	2.73	3.19	1.69	3.35	3.02
DeepSeek-R1 (Guo et al., 2025)	✓	66.93	60.82	44.69	57.48	4.30	2.87	3.49	1.97	3.44	3.21
Emotion Large Language Models											
InstructERC (Lei et al., 2023a)	✗	63.06	66.35	44.35	57.92	-	-	-	-	-	-
LaERC-S (Fu et al., 2025)	✗	65.27	66.01	44.53	58.60	-	-	-	-	-	-
BiosERC (Xue et al., 2024)	✗	64.13	65.99	44.01	58.04	-	-	-	-	-	-
CoE (Shen et al., 2025b)	✓	65.01	64.03	44.29	57.78	3.12	2.76	2.56	1.16	2.01	2.32
Emotion-LLaMA (Cheng et al., 2024)	✓	57.55	49.23	39.21	48.66	2.18	1.22	1.27	1.17	1.71	1.51
R1-Omni (Zhao et al., 2025)	✓	55.76	50.13	41.02	48.97	2.20	1.21	1.25	1.18	1.69	1.51
Ours	✓	**74.17**	**66.48**	**49.86**	**63.50**	**4.39**	**3.20**	**3.64**	**2.10**	**3.49**	**3.36**

Table 2: Model performance is compared from two perspectives. Emotion recognition is evaluated by weighted-F1 (%), whereas reasoning quality is rated on the full test set using five 5-point criteria: Structural (ST.), Answer Consistency (AC.), Factual Alignment (FA.), Conflict Check (CC.) and Comprehensive Coverage(CoC.). The best scores are bolded, respectively. Details of the compared methods can be found in the Appendix A.1.

multiple clues, we employ an integration path, emphasizing Complete Coverage, Fine-grained Coverage, and Logical Rigor so the model can aggregate evidence and provide step-by-step, well-supported explanations.

4 Experiments

4.1 Experiments Set up

Datasets. We conduct experiments on the following three benchmarks: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2018), EmoryNLP (Zahiri and Choi, 2018).

Evaluation Metrics. Following previous works (Lei et al., 2023a), we choose the weighted average F1 score (w-F1) as the emotion recognition evaluation metric. Reasoning quality is scored by DeepSeek on a 5-point scale. We assess reasoning quality across all samples using *Structural* and *Answer Consistency* as core dimensions. For samples with explicit cues and low ambiguity, we additionally consider *Factual Alignment* to verify the correctness of factual grounding. For samples involving emotional reversals or internal contradictions, we introduce *Conflict Check* as an extra dimension to ensure consistency in the reasoning process. For samples that require support from multiple dispersed cues, we activate *Comprehensive Coverage* to encourage multi-evidence, chain-based argumentation.

Implementation Details. We use Llama3-8B as our backbone. During supervised fine-Tuning, we set the batchsize as 8, a learning rate of $2e{-}4$. During reinforcement learning, we set batchsize as 16, and generate $G = 8$ candidate responses per input. We implement experiments on 80G Nvidia A800 GPUs.

4.2 Main Results

We introduce the main experiments from three perspectives, including general reasoning large language models, emotion large language model with and without reasoning ability. We evaluate all models from two complementary angles: emotion recognition weighted-F1 and reasoning quality. For fairness, all comparisons are conducted under the same input and identical evaluation metrics.

Compared with general reasoning LLMs. The results show that although general LLMs exhibit strong generic capabilities, their performance on both metrics lags behind emotion-specialized models. This gap can be attributed to their difficulty in capturing task-relevant emotional cues.

Compared with emotion LLMs without reasoning capability. Conventional text-based emotion classifiers methods (InstructERC, LaERC-S, BiosERC) can achieve strong w-F1 in practice, but most of them lack reasoning capability. This limitation hinders their understanding of challenging samples, thereby constraining their overall emotion recognition performance.

Compared with emotion LLMs with reasoning capability. In the video-based and audio-based emotion domains, several methods (Emotion-LLaMA, R1-Omni) have successfully incorporated

Variations	*Experiment Setting*	Conversational Emotion Recognition (w-F1↑)				Emotion Reasoning	
		IEMOCAP	MELD	EmoryNLP	Avg.	Score↑	Token Len↓
Baseline 1	SFT-Fast	68.50	63.22	48.23	59.98	–	8.42
Baseline 2	SFT-Slow	72.51	65.13	49.07	62.24	3.01	238.61
Baseline 3	SFT-Fast-Slow	72.80	65.73	49.23	62.59	3.19	50.23
	(a) Reward Model						
V1	GRPO-Fast-Slow-w/o-outcome reward	72.59	65.11	47.71	61.80	3.12	41.05
V2	GRPO-Fast-Slow-w/o-format reward	72.20	65.32	47.65	61.72	3.09	40.16
V3	GRPO-Fast-Slow-w/o-process reward	72.89	66.16	48.23	62.43	3.15	42.13
V4	GRPO-Fast-Slow	**74.17**	**66.48**	**49.86**	**63.50**	3.36	38.94
	(b) Thinking Strategy						
V5	GRPO-Slow	73.56	66.35	49.50	63.14	3.07	235.63
V6	GRPO-Fast	69.95	64.83	48.90	61.23	–	8.53

Table 3: Ablation study on reward function and thinking strategy for conversational emotion recognition and emotion reasoning on the overall test set. V1–V4 are built upon Baseline 3. V5 is built upon Baseline 2. V6 is built upon Baseline 1.

emotion reasoning mechanisms. However, when evaluated using semantics-oriented criteria, these approaches exhibit limited capability and generally underperform text-based methods. This is because their emotion recognition improvements largely stem from shallow multimodal manifestations (e.g., facial expressions), which are inherently coarse-grained and fail to capture complex linguistic phenomena.

Additionally, compared with CoE, which mainly constructs supervision by post-hoc filtering generated rationales and thus cannot optimize reasoning. Our method uses GRPO with a adaptive fast-slow thinking schedule to directly control reasoning behavior and jointly improve prediction accuracy and explanation quality under unified reward learning.

4.3 Ablations

SFT vs. GRPO As shown in Table 3, SFT mainly relies on static supervision. SFT-Fast performs worse, while SFT-Slow improves recognition but generates overly long reasoning. The hybrid SFT-Fast-Slow offers a trade-off, yet it remains limited by the "supervision-as-target" paradigm and cannot jointly optimize accuracy and reasoning quality.

In contrast, GRPO explicitly shapes the model's reasoning behavior through reward-driven policy optimization, leading to more consistent improvements and better efficiency. For example, GRPO-Fast-Slow (V4) consistently outperforms the strongest SFT baseline (Baseline 3) by +1.37/+0.75/+0.63 on three datasets, respectively. While also yielding higher-quality reasoning (Score 3.35) with more concise outputs (Token Len 38.94). These results indicate that GRPO can simultaneously enhance emotion recognition performance and reasoning quality while effectively controlling reasoning length, thereby demonstrating a clear overall advantage over SFT.

Effect of Reward. We further analyze the contribution of different reward components by ablating outcome-, format-, and process-level rewards (V1–V3). Removing the outcome reward (V1) lowers CER performance, showing that supervising the final emotion prediction is important for stable optimization. Removing the format constraint (V2) also hurts performance, indicating that structured outputs help regularize training and reduce noisy exploration.

Notably, removing the process-level reward (V3) leads to a substantial decline in emotional reasoning quality scores and a decrease in CER performance compared to the full model. We hypothesize that, in the absence of adaptive process supervision, the model is more likely to converge to a single, rigid reasoning pattern, making it difficult to dynamically switch reasoning paths across dialogues with varying semantic characteristics. As a result, the selection and organization of critical cues become less targeted, ultimately impairing overall performance.

Effect of Thinking Strategy As shown in Table 3, integrating fast and slow thinking in a unified manner achieves a more favorable trade-off between performance and efficiency. Although employing slow thinking alone (V5) yields competitive emotion recognition performance, it substantially increases the length of reasoning traces,

Settings (α weights in Eq. (9))	IEMOCAP	MELD	EmoryNLP	Avg.
Slow: $(\alpha_a, \alpha_f, \alpha_p) = (0.4, 0.2, 0.4)$, Fast: $(\alpha_a, \alpha_f) = (0.8, 0.2)$	73.56	66.36	49.72	63.21
Slow: $(\alpha_a, \alpha_f, \alpha_p) = (0.25, 0.25, 0.5)$, Fast: $(\alpha_a, \alpha_f) = (0.75, 0.25)$	73.93	66.38	49.77	63.36
Slow: $(\alpha_a, \alpha_f, \alpha_p) = (0.45, 0.1, 0.45)$, Fast: $(\alpha_a, \alpha_f) = (0.9, 0.1)$	**74.17**	**66.48**	**49.86**	**63.50**

Table 4: Sensitivity analysis on fast–slow thinking parameters.

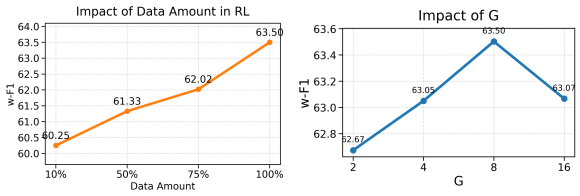

Figure 3: The impact of G and data amount in RL.

resulting in higher computational overhead. In contrast, relying solely on fast thinking (V6) incurs minimal reasoning cost but tends to overlook critical cues in complex emotional scenarios, leading to limited performance. By comparison, the combined fast–slow thinking strategy (V4) consistently improves performance on three datasets while maintaining relatively compact reasoning lengths. These results indicate that the integration of fast and slow thinking enables the model to better adapt to variations in conversational semantic complexity and, consequently, enhances both emotion recognition and reasoning quality.

Computational overhead and efficiency. The results reveal a clear efficiency–depth trade-off: slow thinking greatly increases token length and overhead for only moderate gains, whereas fast thinking is cheap but less effective. GRPO-Fast-Slow offers the best balance, improving performance with compact reasoning and far lower token cost than pure slow reasoning.

4.4 Parameter Sensitivity Analysis

Analysis of Reward. Table 4 indicates that model performance is somewhat sensitive to the allocation of the α weights in Eq. (9), and that relatively balanced weighting across objectives is typically more effective than either uniform averaging or overly conservative settings. We find that assigning a lower weight to the format reward leads to better performance, suggesting that the model is not short of signals for format compliance during training. Therefore, the reward design should prioritize process reward and outcome rewards.

Analysis of RL data amount. In Figure 3, by varying the amount of RL training data, the left plot shows a clear positive correlation between RL data size and weighted F1. Performance steadily improves and is nearly monotonic as more samples are used. Notably, even a very small RL set (e.g., 10% samples) brings noticeable gains, indicating that policy optimization can benefit from limited but informative feedback. As the data scale increases, improvements become more stable and consistent, suggesting that broader trajectory coverage enhances optimization stability and model generalization.

Analysis of G. In Figure 3, the right plot shows a non-monotonic dependence on the hyperparameter G: performance improves as G increases from small values, peaks at an intermediate range, and then degrades when G becomes larger. This behavior is consistent with a bias–variance trade-off in the RL procedure. When G is too small, the learning signal may be under-expressed, yielding insufficient optimization pressure, whereas an overly large G can amplify update noise and induce training instability. Therefore, G serves as a key control knob that must be tuned to balance learning strength and stability.

5 Conclusion

In conclusion, ERCThinker advances conversational emotion recognition by unifying interpretable fast–slow reasoning with accurate prediction. By introducing fine-grained reasoning chains that model topic, discourse structure, speaker traits, scene context, and emotion shifts, the framework captures deeper conversational semantics than prior fast-only or shallow slow approaches. Its adaptive switching at both strategy and utterance levels enables efficient allocation of reasoning depth to diverse scenarios. In addition, the Agent-as-Judge provides process-aware rewards that refine explanations and stabilize learning. Experiments across multiple benchmarks confirm consistent gains in accuracy and reasoning quality.

Limitations

This work has several limitations. First, to ensure a fair comparison between open-source and closed-source LLMs, we evaluated publicly available models only through direct inference rather than fine-tuning. As a result, we did not report fine-tuning results on mainstream open-source models, such as 7B-scale LLMs, which may limit the completeness of the assessment of open-source models' emotional reasoning capability. Second, our definition of fast thinking is relatively simplified and mainly based on surface-level cues such as punctuation, emojis, and sentence length. While this design improves feasibility, it may overlook cases where short sentences still contain complex emotions and require deeper analysis. Such simplification may introduce classification bias and should be further refined in future work.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Rajdeep Chatterjee, Saptarshi Mazumdar, R. Simon Sherratt, Rohit Halder, Tanmoy Maitra, and Debasis Giri. 2021. Real-time speech emotion analysis for smart home assistants. *IEEE Transactions on Consumer Electronics*, 67(1):68–76.

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.

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 (CVPR) Workshops*, pages 4652–4661.

Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *Preprint*, arXiv:2503.17352.

Jonathan St BT Evans. 1984. Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75(4):451–468.

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

Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 248:108861.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475.

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

Shanglin Lei, Xiaoping Wang, Guanting Dong, Jiang Li, and Yingjian Liu. 2023b. Watch the speakers: A hybrid continuous attribution network for emotion recognition in conversation with emotion disentanglement. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 881–888.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2024. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *Trans. Multi.*, 26:77–89.

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

Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, and Xiaobai Li. 2025. Affectgpt-r1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2508.01318*.

Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. 2023. Hierarchical dialogue understanding with special tokens and turn-level attention. *Preprint*, arXiv:2305.00262.

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

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, and 1 others. 2025. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, and 1 others. 2025. Magistral. *arXiv preprint arXiv:2506.10910*.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560, Online. Association for Computational Linguistics.

Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. 2025a. 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, Vienna, Austria. Association for Computational Linguistics.

Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. 2025b. 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.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Geng Tu, Jun Wang, Zhenyu Li, Shiwei Chen, Bin Liang, Xi Zeng, Min Yang, and Ruifeng Xu. 2024. Multiple knowledge-enhanced interactive graph network for multimodal conversational emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3861–3874.

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.

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, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025a. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Yongxin Wang, Dongchun Xie, Yiwei Wang, Xiaodan Liang, and Jing Tang. 2025b. Depth-breadth synergy in rlvr: Unlocking llm reasoning gains with adaptive exploration. *arXiv preprint arXiv:2508.13755*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

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

Tao Zhang and Zhenhua Tan. 2025. ECERC: Evidence-cause attention network for multi-modal emotion recognition in conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2064–2077, Vienna, Austria. Association for Computational Linguistics.

Zhiqiang Zhang and Wen Zhao. 2025. A collaborative reasoning framework powered by reinforcement learning and large language models for complex questions answering over knowledge graph. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10672–10684.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*.

Haoyu Zheng, Zhuonan Wang, Yuqian Yuan, Tianwei Lin, Wenqiao Zhang, Zheqi Lv, Juncheng Li, Siliang Tang, Yueting Zhuang, and Hongyang He. 2025. Fast thinking for large language models. *Preprint*, arXiv:2509.23633.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582.

A Appendix

A.1 Baseline Details

InstructERC (Lei et al., 2023a). This framework reformulates emotion recognition in conversation by transitioning from a discriminative to a generative paradigm, utilizing a retrieval template module and multi-task emotion alignment to integrate multi-granularity dialogue supervision into LLMs.
LaERC-S (Fu et al., 2025). This framework leverages the extensive world knowledge of LLMs through a two-stage learning paradigm to explicitly model interlocutor mental states and behavioral characteristics, facilitating more precise reasoning of emotional dynamics within complex conversational contexts.
BiosERC (Xue et al., 2024). This framework enhances emotion classification by utilizing LLMs to extract speaker "biographical information," injecting these high-level personality traits and characteristics as supplementary contextual knowledge to model complex emotional interactions.
CoE (Shen et al., 2025b). This framework introduces a multi-stage auxiliary learning strategy that progressively integrates role-playing, speaker identification, and emotion reasoning tasks to enhance a LLM's capacity for interpreting complex emotional clues and contextual dynamics within conversational streams.
Emotion-LLaMA (Cheng et al., 2024). This framework introduces a multimodal architecture that integrates audio, visual, and textual modalities via emotion-specific encoders and aligns these features into a shared embedding space, utilizing a modified LLaMA backbone through instruction tuning to enhance fine-grained emotional recognition and reasoning.
R1-Omni (Zhao et al., 2025). This approach introduces a Reinforcement Learning with Verifiable Reward (RLVR) paradigm to optimize Omni-multimodal large language models, enhancing the interpretability of cross-modal feature contributions while simultaneously boosting recognition accuracy and zero-shot generalization in complex emotional contexts.

A.2 Cues Extraction Pipeline

We prioritize contextual cues that are most relevant to conversational emotion understanding. Standard processing pipelines are employed to extract the following five types of cues:

- **Topic Information**: Each utterance is encoded using a pretrained language model, and a continuous latent topic representation is learned via a VAE-style objective. Attention mechanisms are used to model topic transitions across dialogue turns (Zhu et al., 2021).
- **Speaker Characteristics**: Following (Ghosal et al., 2020), we encode each utterance with a pretrained speaker characteristic inference model to derive latent representations capturing speaker intent, effects, and reactions. These are incorporated into downstream models as structured semantic features.
- **Discourse Structure**: We apply deep sequential discourse modeling (Shi and Huang, 2019) to induce explicit discourse dependency trees for each conversation, enabling identification of the specific preceding utterance to which each utterance responds.

- **Scene Information**: As in (Shen et al., 2025b), scene context, including location and environmental descriptions annotated at the beginning of each dialogue segment, is parsed and directly incorporated into the model input.
- **Emotion Shifts**: A binary label is assigned to each pair of adjacent utterances to indicate whether their emotion labels differ, following the emotion shift detection approach of (Gao et al., 2022).

ID	Parent	Dimension	Description
1	—	Structural	Follows a common reasoning structure (e.g., specific-to-general, sequential, or conclusion-supported-by-analysis).
2	1	Evidence Recall	Extracts at least one concrete cue from the input utterance.
3	2	Factual Alignment	The cited cue aligns with the true emotion and avoids speculation or fabrication.
4	2	Conflict Check	None of the cited cues contradict the true emotion.
5	2	Logical Structure	Clearly links evidence to the conclusion without structural gaps or jumps.
6	3	Comprehensive Coverage	Uses multiple distinct and meaningful emotional cues to support the final judgment.
7	3	Fine-Grained Coverage	Closely ties reasoning to verbatim phrases or precise details from the utterance.
8	5	Answer Consistency	Ends with a single, clear emotion word that matches the preceding reasoning.
9	5	Language Quality	Uses fluent, natural, and grammatically sound language.
10	6	Conciseness	Is brief and avoids redundancy or irrelevant content.
11	7	Logical Rigor	Every reasoning step is grounded in explicit evidence or cited cues.
12	4	Intensity Match	The expressed emotion intensity (e.g., “very angry” vs. “a little upset”) matches that of the utterance, helping distinguish similar emotions.
13	4	Internal Consistency	Contains no self-contradictory statements within the reasoning.
14	4	Contextual Fit	Takes into account speaker identity, setting, and other contextual factors.

Table 5: Evaluation Dimensions and Dependency Hierarchy.

Prompt for evaluating reasoning process.

You are an expert evaluator of emotion recognition reasoning. Assess how well the model-generated explanation (in `<think>...</think>`) justifies its predicted emotion label (in `<answer>...</answer>`), using the ground-truth answer, reference explanation, and any provided metadata.

Rate each selected dimension on a **1–5 scale**:
- **5**: Better than reference
- **4**: On par with reference (minor factual issues allowed)
- **3**: Weak reasoning or structure; multiple factual errors affect *only* Factual Alignment
- **2**: Irrelevant, hallucinated, or no real reasoning
- **1**: Severely flawed reasoning

Context
Reference Explanation: {gt_reasoning}

Evaluation
Conversation
{conversation}
Speaker and Sentence
{speaker} : {sentence}
Ground-Truth Answer:{gt_answer}

Model Explanation:
####Thought
<think>{thoughts}</think>
<answer>{predict_answer}</answer>

Evaluation Indicators (with dependency hierarchy)
{evaluation_indicators}

Output Instructions
Select **exactly five metrics** from the **Type** column in the Evaluation Indicators table that best reflect the strengths and weaknesses of **this specific model explanation**.
- Only include a metric if it is meaningfully applicable (e.g., skip *Fine-Grained Coverage* if no exact quotes appear).
- **Respect dependency rules**: selecting a child node requires including all its ancestors (e.g., choosing *Logical Rigor* (ID 11) implies including *Fine-Grained Coverage* (7), *Factual Alignment* (3), and *Evidence Recall* (2)).
- Use the **exact Type names** (e.g., `"Factual Alignment"`) as keys—**not placeholders or IDs**.
- If the groundtruth is conflicting with the predict answer, the thinking process you need to evaluate is not good.

```json
{{
    "comment": "2–3 sentences explaining why these specific Type-named metrics were chosen based on the actual content, structure, and alignment of the model's reasoning.",
    "selected_metrics": {{
        "<Type_Name>": int,
        "<Type_Name>": int,
        "<Type_Name>": int,
        "<Type_Name>": int,
        "<Type_Name>": int
    }}
}} ```

Figure 4: The prompt for evaluating reasoning process.
```

Prompt for generating emotion reasoning process.

Now you are an expert in emotional discourse analysis.You are analyzing a conversation characterized by the following dimensions:
- topic: {topic}
- discourse structure: {discourse_structure}
- speaker characteristics: {speaker_characteristics}
- scene: {scene}
- emotion shift: {emotion_shift}

Conversation:
conversation

Task
Your task is to provide a detailed reasoning explanation of how the speaker's emotions evolve, leading to the specific shift labeled as emotion.

Analysis guidelines
- Simulate a human reader's critical thinking. Let the emotional interpretation emerge gradually through observation and analysis.
- Weave the speaker's characteristics and the scene's atmosphere into your narrative to justify the emotional intensity.
- Explain how the specific discourse structure and the progression of the topic trigger the emotional change.
- Explicitly reference or quote key phrases from the conversation to ground your reasoning in the actual text.
- Ensure the explanation is a cohesive, well-structured piece of writing between 80 and 250 words.

Figure 5: The prompt for generating emotion reasoning process.

Prompt for scoring fast-thinking characteristics of a sample.

You are an expert evaluator assessing whether a given utterance exhibits characteristics of **fast thinking**—immediate, affective, and intuitive expression. Score the sample on a **1–5 scale** for each criterion below, based solely on the utterance and its conversational context.

Evaluation
Conversation
{conversation}
Speaker and Sentence
{speaker} : {sentence}
Ground-Truth Answer: {gt_answer}

Fast-Thinking Indicators
Rate each dimension independently (1–5):
- **Emotion Word Density**: Frequency of strong emotion words (e.g., "angry", "thrilled", "devastated")
- **Punctuation Intensity**: Use of exclamation marks (!), question marks (?), or multiple punctuation (!!)
- **Emoji Usage**: Presence and relevance of emojis (e.g., [laugh], [cry])
- **Discourse Particles/Interjections**: Words like "wow", "ugh", "hey", "oh no"
- **Brevity**: Extremely short utterance (e.g., < 10 words or < 30 characters)

Output Instructions
Provide a JSON object with exact keys as listed above. Use integers 1–5 only.

```json
{
    "Emotion Word Density": int,
    "Punctuation Intensity": int,
    "Emoji Usage": int,
    "Discourse Particles/Interjections": int,
    "Brevity": int
}
```

Figure 6: Scoring prompt for fast-thinking characteristics.

Prompt for scoring slow-thinking characteristics of a sample.

You are an expert evaluator assessing whether a given utterance exhibits characteristics of **slow thinking**—deliberate, reflective, and cognitively complex reasoning. Score the sample on a **1–5 scale** for each criterion below, based solely on the utterance and its conversational context.

Evaluation
Conversation
{conversation}
Speaker and Sentence
{speaker} : {sentence}
Ground-Truth Answer: {gt_answer}

Slow-Thinking Indicators
Rate each dimension independently (1–5):
- **Expression Length**: Number of characters / tokens (longer → higher score)
- **Structural Complexity**: Use of multiple clauses, compound/complex sentences
- **Causal/Contrastive Connectives**: Presence of words like "because", "but", "although", "so"
- **Coreference**: Use of pronouns or demonstratives requiring inference (e.g., "he", "this")
- **Metacognitive Markers**: Phrases indicating reflection, uncertainty, or perspective-taking (e.g., "I think", "maybe", "from their view", "calm down")

Output Instructions
Provide a JSON object with exact keys as listed above. Use integers 1–5 only.

```json
{
    "Expression Length": int,
    "Structural Complexity": int,
    "Causal/Contrastive Connectives": int,
    "Coreference": int,
    "Metacognitive Markers": int
}
```

Figure 7: Scoring prompt for slow-thinking characteristics.

Prompt for emotion analysis with reasoning.

You are an expert in emotion recognition. Your task is to analyze the emotional state expressed by a specific speaker in a given utterance, based on the full conversational context. Carefully consider the speaker's words, implied tone, interpersonal dynamics, and situational cues.
Evaluation
Conversation
{conversation}
Speaker and Sentence
{speaker} : {sentence}

Task
Based on the conversation above, analyze the emotion expressed in the sentence spoken by {speaker}. Choose **one** emotion label from the following set: <{labels}>. Then, provide a brief reasoning for your choice, followed by your confidence level in this prediction on a scale from 0 to 5, where: - 0 = completely uncertain or no basis for judgment, - 5 = extremely confident, with clear and unambiguous evidence.
Output Format
You **must** output exactly in the following format—no additional text:
<think>{your concise reasoning here}</think> <answer>{selected emotion label}</answer> <confidence>{integer from 0 to 5}</confidence>

Figure 8: Emotion analysis prompt with reasoning.