Selective Reflection for Context-Sensitive Emotion Recognition in Conversation

Xinyu Wang, Weijie Feng†, Hongchuang Wang, Zhiyong Cheng
Hefei University of Technology
xinyuwang@mail.hfut.edu.cn, wjfeng@hfut.edu.cn
hongchuangwang@mail.hfut.edu.cn, jason.zy.cheng@gmail.com

Abstract

Recent advances in Large Language Models have transformed Emotion Recognition in Conversation, yet existing instruction-tuned approaches still produce high-confidence misjudgments on utterances whose surface text is weakly informative but whose emotion depends on the surrounding situation. We propose SRERC, a selective reflective correction framework for LLM-based ERC. SRERC uses a verdict teacher to assess contextual support for the first-pass prediction and triggers reflective decoding only when reflection has an actionable role. For selected cases, an SRC generator constructs a Scene-aware Reflection Context, which provides target-specific situational grounding and verification probes for an SRC-conditioned second pass. This gated design enables targeted correction while preserving non-gated predictions and limiting unnecessary over-correction. Experiments on IEMOCAP, MELD, and EmoryNLP show that SRERC achieves the highest average weighted-F1 across the three benchmarks, with diagnostic gains on context-sensitive cases where surface cues are insufficient or misleading.

1 Introduction

Large language models (LLMs) have become increasingly central to dialogue understanding, owing to their strong abilities in instruction following, contextual reasoning, and open-ended language generation. In emotion recognition in conversation (ERC), this trend has led recent studies to reformulate utterance-level emotion prediction as an instruction-following or generative task (Lei et al., 2024; Xue et al., 2024; Fu et al., 2025; Zhang et al., 2025). This formulation is attractive because conversational emotions are rarely determined by isolated lexical cues. Instead, they often depend on speaker interactions, topic progression, implicit intentions, and local affective state shifts (Poria et al., 2019b; Majumder et al., 2019). ERC therefore requires LLMs to ground their predictions in evolving conversational context.

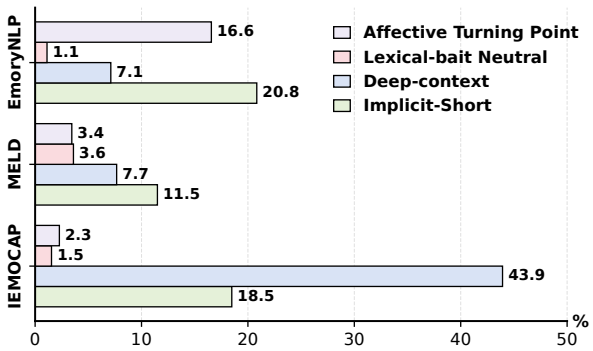


Figure 1: Prevalence of four context-sensitive diagnostic case types in the test sets.

Despite this progress, most LLM-based ERC systems still follow a one-pass prediction paradigm (Lei et al., 2024; Xue et al., 2024; Fu et al., 2025; Zhang et al., 2025; Jing et al., 2026). They provide the model with a dialogue history, sometimes augmented with speaker profiles or character descriptions (Xue et al., 2024; Fu et al., 2025), and ask it to generate the final emotion label in a single decoding step. The limitation of this paradigm is not that context is absent, but that the generated label is not explicitly verified against the target-specific dialogue situation. Once the model forms a plausible label hypothesis, there is no mechanism for checking whether this label is sufficiently supported by the surrounding context. As a result, predictions can become unreliable when the target utterance is lexically weak, pragmatically implicit, or marks a local shift in the affective state of the dialogue.

To examine whether such context-sensitive cases are merely anecdotal, we identify four recurring diagnostic patterns in the benchmark test sets: Implicit-Short, Deep-context, Lexical-bait Neutral,

†Corresponding Author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2466–2481
November 6-10, 2026 ©2026 Association for Computational Linguistics

and Affective Turning Point. Their formal definitions and construction rules are provided in Appendix A. As shown in Figure 1, these cases occur with non-trivial frequency across IEMOCAP, MELD, and EmoryNLP. This suggests that the reliability issue of one-pass ERC is not limited to rare corner cases, but repeatedly appears when target emotions are weakly expressed, require earlier-turn evidence, locally depart from nearby emotional cues, or are easily misread from surface words.

To address this issue, we propose **SRERC**, a **S**elective **R**eflective correction framework for LLM-based ERC. Rather than applying reflection to every utterance, SRERC uses an independent verdict teacher to identify predictions for which reflective decoding has an actionable role, thereby reducing unnecessary re-decoding and over-correction. Given a first-pass prediction and its confidence score from an instruction-tuned ERC backbone, the verdict teacher assesses whether the prediction is supported by the local dialogue context. For selected cases, a frozen SRC generator constructs a Scene-aware Reflection Context (SRC), which provides target-specific situational grounding and verification probes without revealing the emotion label. The same ERC backbone is then queried once more with the SRC-augmented prompt, while non-gated predictions are preserved.

In summary, our contributions are threefold:

- **(Framework)** We introduce **SRERC**, a selective reflective correction framework for LLM-based ERC that preserves standard first-pass prediction while adding a gated reflective decoding path for context-sensitive cases.
- **(Mechanism)** We design a verdict-guided reflection gate with a Scene-aware Reflection Context (SRC), enabling selective consolidation or correction based on contextual-support signals.
- **(Evaluation)** Experiments on three benchmark datasets show that SRERC achieves the best average weighted-F1 among compared LLM-based ERC methods, with further gains on diagnostic context-sensitive cases and positive sample-level correction effects.

2 Related Work

2.1 Emotion Recognition in Conversation

Early ERC research treats the task as a discriminative sequence-labelling problem and relies on recurrent (Majumder et al., 2019; Jiao et al., 2020; Hu et al., 2023; Li et al., 2024b; Gou et al., 2025), graph-based (Ghosal et al., 2019; Hu et al., 2021; Shen et al., 2021b; Ai et al., 2025; Li et al., 2025), or attention-based (Bao et al., 2022; Li et al., 2024b,a; Nguyen et al., 2024) models to capture intra- and inter-speaker context (Fu et al., 2023; Wu et al., 2025; Shou et al., 2025). The emergence of instruction-tuned large language models (Touvron et al., 2023; Wei et al., 2022a) has more recently shifted the field towards a generative paradigm, in which ERC is reformulated as label generation with a fine-tuned LLM (Lei et al., 2024; Xue et al., 2024; Fu et al., 2025; Zhang et al., 2025; Jing et al., 2026). Despite their strong performance, most existing LLM-based ERC systems still produce the final label through a single-pass decoding process. This makes them vulnerable to utterances whose surface form is weakly informative or misleading, but whose emotion depends on transient dialogue context.

2.2 Contextual and Speaker Modeling in ERC

A complementary line of work studies what auxiliary context should be supplied to ERC models. Existing approaches inject external commonsense or structured affective knowledge (Ghosal et al., 2020; Zhong et al., 2019; Zhu et al., 2021; Li et al., 2023), or rely on a frozen LLM to produce static speaker biographies or character descriptions that are concatenated to ERC prompts (Xue et al., 2024; Fu et al., 2025). While effective on average, these auxiliary signals are typically introduced before label generation and are not used to verify whether a particular generated label is supported by the target-specific situation. Our work instead focuses on prediction re-examination: it asks whether additional situational evidence should be introduced for selected first-pass predictions.

2.3 Self-Correction for LLMs

A growing body of work investigates how LLMs can refine or correct their own outputs through chain-of-thought prompting (Wei et al., 2022b; Yao et al., 2023), iterative self-critique (Madaan et al., 2023), or verbal reinforcement loops (Shinn et al., 2023). However, directly applying self-correction to every ERC prediction is both computationally expensive and vulnerable to over-correction, where additional generation rounds may overturn predictions that were already correct (Huang et al., 2024). This motivates a selective correction perspective for LLM-based ERC, where reflection is treated

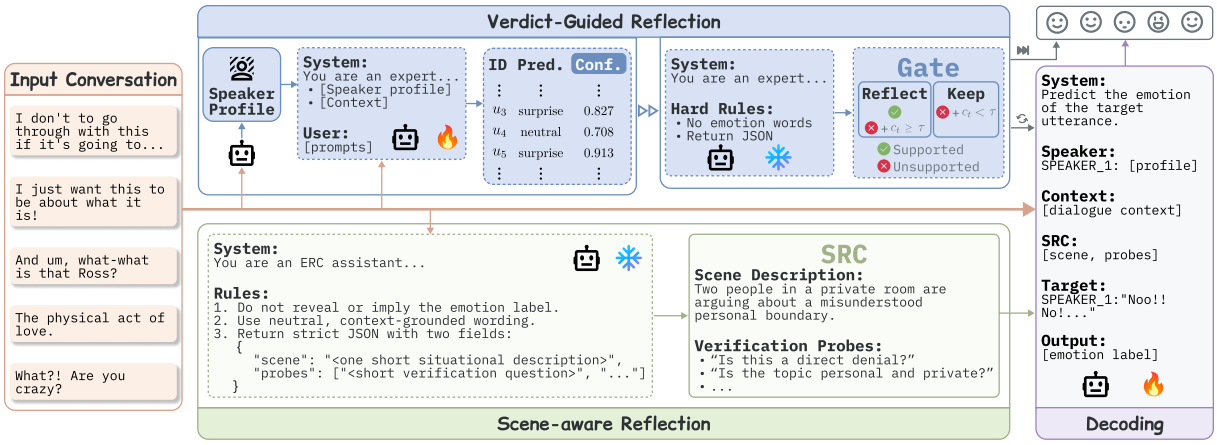


Figure 2: Overview of SRERC.

as a conditional operation rather than a universal post-processing step.

3 Methodology

Given a dialogue $C = \{u_1, \cdots, u_N\}$ uttered by a speaker set $\mathcal{P} = \{p_1, \cdots, p_M\}$, emotion recognition in conversation (ERC) aims to predict an emotion label $y_t \in \mathcal{E}$ for each target utterance u_t, where $\mathcal{E}$ is a predefined emotion-label set. Let $p_t \in \mathcal{P}$ denote the speaker of u_t. We propose SRERC, a selective reflective correction framework that improves the contextual reliability of LLM-based ERC. As illustrated in Figure 2, SRERC first obtains a first-pass prediction and a confidence score from an instruction-tuned ERC backbone. An independent verdict teacher then assesses whether the prediction is supported by the local dialogue context, and a reflection gate selects actionable cases for either consolidation or corrective re-examination (§3.1). For gated cases, SRERC constructs a Scene-aware Reflection Context (SRC) with target-specific situational grounding and verification probes (§3.2), and queries the same ERC backbone once more to obtain the final prediction (§3.3).

3.1 Verdict-Guided Reflection Gate

The reflection gate determines whether a first-pass prediction should enter reflective decoding with additional reflection context. Rather than applying reflection to every utterance, SRERC selects cases where reflection has an actionable role: either consolidating contextually supported predictions or re-examining high-confidence predictions that lack sufficient contextual support.

First-pass prediction. We use an instruction-tuned LLM M_{erc} as the ERC backbone. Following previous work (Xue et al., 2024), a frozen profile extractor M_{prof} derives a one-time textual speaker profile ρ_{p_t} for each speaker $p_t \in \mathcal{P}$ from dialogue C before first-pass prediction:

$$\rho_{p_t} = M_{\text{prof}}(C, p_t). \tag{1}$$

The profile extractor is queried once per speaker and remains frozen throughout training and inference. For each target utterance u_t produced by speaker p_t, we construct the first-pass prompt x_t as:

$$x_t = \text{prompt}\big(C_{<t}, u_t, \rho_{p_t}\big), \tag{2}$$

where $C_{<t}$ denotes the preceding dialogue context, and $\text{prompt}(\cdot)$ denotes the deterministic template used to assemble the ERC input. The ERC backbone M_{erc} is fine-tuned to generate the verbalized gold emotion label with the standard next-token cross-entropy loss. At inference time, M_{erc} produces a first-pass prediction:

$$\hat{y}_t = \arg\max_{e \in \mathcal{E}} P_{M_{\text{erc}}}(e \mid x_t), \tag{3}$$

where $\mathcal{E}$ is the emotion-label set. Since labels are generated as text, we estimate the confidence of $\hat{y}_t$ from the generation probability of its label-distinguishing token. Let $\lambda(e)$ denote the first token that uniquely identifies label e among the candidate labels. We define the first-pass confidence $c_t \in [0, 1]$ as

$$c_t = P_{M_{\text{erc}}}\big(\lambda(\hat{y}_t) \mid x_t\big). \tag{4}$$

Verdict teacher. To assess whether the first-pass prediction $\hat{y}_t$ is contextually supported, SRERC uses an independent frozen verdict teacher M_{ver}. We implement $M_{\text{ver}}^{\text{rat}}$ and $M_{\text{ver}}^{\text{judge}}$ as two structured

reasoning stages within one call to the same verdict teacher: evidence grounding and the binary judgment are elicited in a fixed JSON schema. Given the target utterance u_t, its preceding context $C_{<t}$, the target-speaker profile ρ_{p_t}, the first-pass prediction $\hat{y}_t$, and the confidence score c_t, the verdict teacher first derives an evidence-grounding rationale:

$$(h_t, \mathcal{D}_t) = M_{\text{ver}}^{\text{rat}}\big(C_{<t}, u_t, \rho_{p_t}, \hat{y}_t, c_t\big), \quad (5)$$

where $\mathcal{D}_t = \{d_{t,1}, \ldots, d_{t,K}\}$ is a small set of short textual evidence cues, with each $d_{t,k}$ grounded in the target utterance or its surrounding turns, and h_t is an optional non-label textual description of a potential behavioral mismatch between the first-pass prediction and the contextual evidence. Conditioned on the evidence-grounding rationale, the verdict teacher then verifies whether the current hypothesis $\hat{y}_t$ is contextually supported:

$$v_t = M_{\text{ver}}^{\text{judge}}\big(\hat{y}_t, c_t, h_t, \mathcal{D}_t\big), \quad (6)$$

where $v_t \in \{\texttt{Supported}, \texttt{Unsupported}\}$. Here, $v_t = \texttt{Supported}$ means that the evidence cues in $\mathcal{D}_t$ are compatible with the first-pass prediction $\hat{y}_t$. Conversely, $v_t = \texttt{Unsupported}$ means that the cues provide insufficient support for, or contradict, the current prediction, optionally accompanied by the behavioral mismatch signal h_t.

Reflection gate. Given the verdict v_t and confidence c_t, the reflection gate determines whether the first-pass prompt should be kept unchanged or sent to reflective decoding with SRC:

$$\begin{aligned} r_t = \mathbb{1}\big[\, & v_t = \texttt{Supported} \vee \\ & (\, v_t = \texttt{Unsupported} \wedge\, c_t \geq \tau\,)\,\big], \end{aligned} \quad (7)$$

where $r_t \in \{0, 1\}$ indicates whether reflective decoding is triggered, $\tau \in [0, 1]$ is a confidence threshold, and $\mathbb{1}[\cdot]$ denotes the indicator function. This rule separates first-pass predictions into three cases. If $v_t = \texttt{Supported}$, the prediction is judged compatible with the dialogue evidence, and reflective decoding functions as a controlled consolidation step. Although the final label is still produced by the ERC backbone, the added SRC is intended to make the decision depend more explicitly on scene-grounded evidence. If $v_t = \texttt{Unsupported}$ and $c_t \geq \tau$, the backbone is confident but the verdict teacher finds insufficient contextual support, so reflection is used to challenge a potentially misoriented high-confidence prediction. If $v_t = \texttt{Unsupported}$ and $c_t < \tau$, the prediction is both unsupported and low-confidence; such cases provide neither a supported hypothesis to consolidate nor a confident hypothesis worth explicitly challenging, so SRERC keeps the first-pass prompt unchanged.

3.2 Scene-aware Reflection Context

Reflective decoding is useful only when the second pass receives evidence that is not already encoded in the original prompt. Re-decoding with the same input would reduce reflection to stochastic re-sampling, while generating reflection context from the first-pass prediction could rationalize the initial hypothesis. Therefore, SRERC constructs a prediction-agnostic Scene-aware Reflection Context (SRC), which supplies target-specific situational grounding and verification probes without exposing an emotion label.

SRC generation. SRC is generated by a frozen LLM M_{src} and is a compact two-field structure consisting of a situational scene description and a small set of verification probes. For a gated utterance u_t with $r_t = 1$, we first obtain a raw SRC:

$$S_t^0 = \big(a_t,\ \mathcal{Q}_t\big) = M_{\text{src}}\big(C_{<t}, u_t\big), \quad (8)$$

where a_t is a short situational scene description and $\mathcal{Q}_t = \{q_{t,1}, \ldots, q_{t,K_t}\}$ is a set of verification probes. The scene field a_t is grounded in $(C_{<t}, u_t)$ and covers only situational factors such as the ongoing event, speaker relation, topic, or local interactional setting. The probe field $\mathcal{Q}_t$ is a small set of rule-like verification questions, e.g., whether the target response is literal, whether it follows the previous turn, or whether the topic has shifted. These probes specify what the ERC backbone should verify in the second pass, but they do not provide answers, external facts, or emotion-label suggestions. Different from the verdict teacher, M_{src} does not take $\hat{y}_t$ or c_t as input. This design prevents SRC from becoming a post-hoc rationalization of the first-pass prediction and keeps the second pass grounded in the dialogue situation itself.

Label-word firewall. To prevent SRC from leaking the target label, we apply a label-word firewall before reflective decoding. Let $\mathcal{L}_{\text{emo}}$ denote the set of dataset-specific emotion-label words and their close lexical variants. We define a sanitization operator $\Pi_{\mathcal{L}_{\text{emo}}}(\cdot)$ that masks spans matching $\mathcal{L}_{\text{emo}}$. The final SRC used by SRERC is

$$S_t = \big(\Pi_{\mathcal{L}_{\text{emo}}}(a_t),\ \{\Pi_{\mathcal{L}_{\text{emo}}}(q_{t,k})\}_{k=1}^{K_t}\big). \quad (9)$$

Dataset	# Conversations			# Utterances			# Cls.
	train	val	test	train	val	test	
IEMOCAP	108	12	31	5163	647	1623	6
MELD	1038	114	280	9989	1109	2610	7
EmoryNLP	659	89	79	7551	954	984	7

Table 1: Dataset statistics.

The firewall ensures that SRC contributes sanitized situational and reflective evidence while leaving emotion-label generation to M_{erc}.

3.3 Conditional Reflective Decoding

With the reflection gate and SRC construction, SRERC determines the final prediction through conditional reflective decoding. The ERC backbone M_{erc} is shared by both the first-pass prediction and reflective decoding. The only difference lies in the input: the first pass uses the original ERC input x_t, whereas reflective decoding uses an SRC-augmented input. For a gated utterance with $r_t = 1$, we construct the reflective input as

$$x_t^+ = x_t \oplus S_t, \tag{10}$$

where $\oplus$ denotes prompt concatenation. The same ERC backbone then produces a reflective prediction:

$$\bar{y}_t = \arg\max_{e \in \mathcal{E}} P_{M_{\text{erc}}}\left(e \mid x_t^+\right). \tag{11}$$

The final prediction is selected according to the reflection gate:

$$\tilde{y}_t = \begin{cases} \hat{y}_t, & r_t = 0, \\ \bar{y}_t, & r_t = 1. \end{cases} \tag{12}$$

Thus, SRERC preserves non-gated predictions and invokes reflective decoding only when the verdict signal supports either contextual consolidation or high-confidence re-examination.

4 Experimental Setup

4.1 Datasets

All experiments are conducted on three widely used English ERC benchmarks: **IEMOCAP** (Busso et al., 2008), **MELD** (Poria et al., 2019a), and **EmoryNLP** (Zahiri and Choi, 2018). We use only the textual modality for all datasets and follow the official data splits. Dataset statistics are reported in Table 1. Additional dataset details are provided in Appendix B.

Dataset	Trigger	BB passes	Verdict/utt.	SRC/utt.
IEMOCAP	22.1%	1.22	1.00	0.22
MELD	27.5%	1.28	1.00	0.28
EmoryNLP	78.3%	1.78	1.00	0.78

Table 2: Inference profile per utterance. Backbone passes and teacher calls are averages; speaker profiles are extracted once per speaker offline.

4.2 Baselines

We compare SRERC with three groups of baselines. **(1) Discriminative ERC methods** include DialogueRNN (Majumder et al., 2019), DialogueGCN (Ghosal et al., 2019), COSMIC (Ghosal et al., 2020), DAG-ERC (Shen et al., 2021b), SKAIG (Li et al., 2021), CauAIN (Zhao et al., 2022), DialogXL (Shen et al., 2021a), and SPCL (Song et al., 2022). **(2) General-purpose LLMs** include GPT-5.4-mini (OpenAI, 2026), MiMo-V2.5-Pro (Xiaomi MiMo Team, 2026), DeepSeek-V4-Pro (DeepSeek-AI, 2026), and Gemini-3-Flash (Google DeepMind, 2025). **(3) ERC-specific LLM methods** include InstructERC (Lei et al., 2024), BiosERC (Xue et al., 2024), LaERC-S (Fu et al., 2025), and Causal-ERC (Jing et al., 2026). Further details are provided in Appendix C.

4.3 Implementation Details

We instantiate the first-pass ERC backbone M_{erc} with `LLaMA-2-13B` (Touvron et al., 2023) and fine-tune it with QLoRA (Dettmers et al., 2023). Following BiosERC (Xue et al., 2024), we use a frozen MiMo model as the profile extractor M_{prof} to generate one-time speaker profiles, which are then fixed in all ERC prompts. We train for 3 epochs with a learning rate of 3×10^{-4} and an effective batch size of 16, and truncate $C_{<t}$ to a context window of $W = 5$ preceding utterances. We instantiate both the SRC generator M_{src} and the verdict teacher M_{ver} with `gpt-5.4-mini`, and generate teacher outputs offline without access to gold labels. The confidence threshold τ is tuned on the development set. Additional details are provided in Appendix D.

5 Results and Analysis

5.1 Overall Results

Table 3 reports the main results on the three ERC benchmarks. SRERC achieves the best average weighted-F1 among methods evaluated on all datasets, reaching 61.74 and outperforming the

Method	IEMOCAP	MELD	EmoryNLP	*Avg.*
Discriminative ERC methods				
DialogueRNN	64.65	65.30	37.54	55.83
DialogueGCN	62.11	62.68	36.43	53.74
COSMIC	65.03	63.43	38.49	55.65
DAG-ERC	66.54	63.36	38.29	56.06
SKAIG	66.96	65.18	38.88	57.01
CauAIN	65.01	64.89	37.87	55.92
DialogXL	65.94	62.41	34.73	54.36
SPCL	68.42	66.13	40.25	58.27
General-purpose LLMs				
GPT-5.4-mini	57.93	62.54	37.61	52.69
MiMo-V2.5-Pro	57.62	60.71	41.44	53.26
DeepSeek-V4-Pro	59.61	59.50	39.66	52.92
Gemini-3-Flash	60.22	63.94	40.87	55.01
ERC-specific LLM methods				
InstructERC$^\sharp$	71.39	69.15	41.37	60.64
BiosERC$^\sharp$	71.19	69.83	41.68	60.90
LaERC-S$^\sharp$	**72.40**	69.27	42.08	61.25
Causal-ERC$^\sharp$	71.89	70.25	–	–
SRERC (ours)	71.60	**70.37**	**43.25**	**61.74**

Table 3: Main results in weighted-F1 (%). Results marked with $^\sharp$ are taken from the original paper.

Method	IEMOCAP	MELD	EmoryNLP	*Avg.* Δ
InstructERC	68.76	64.48	39.01	–
+ SRERC	**69.48**	**65.05**	**39.81**	**+0.70**

Table 4: Transfer of the complete SRERC mechanism to an independently trained InstructERC backbone (LLaMA-2-7B + LoRA), without backbone-specific tuning (weighted-F1, %).

strongest ERC-specific LLM baseline LaERC-S by $+0.49$. The improvement is mainly reflected on MELD and EmoryNLP, where SRERC obtains the best scores of 70.37 and 43.25, respectively. On IEMOCAP, SRERC remains competitive but does not surpass LaERC-S. Compared with general-purpose zero-shot LLMs, SRERC shows a clear advantage, with a $+6.73$ average gain over the strongest model in this group, Gemini-3-Flash. These results indicate that selective reflective correction can improve strong LLM-based ERC systems at the benchmark level. Since the overall gains are moderate, we further examine whether the improvements are concentrated in the context-sensitive cases targeted by our framework, and whether they arise from beneficial reflective revisions rather than indiscriminate output changes.

Transfer to another backbone. Table 4 separates the reflection effect from our LLaMA-2-13B backbone. Applying the same verdict gate, prediction-agnostic SRC, and conditional re-decoding to an independently trained InstructERC backbone improves weighted-F1 on all three datasets, with an average gain of $+0.70$.

Method	IEMOCAP	MELD	EmoryNLP	*Combined*
Diagnostic 1: Implicit-Short				
# Samples	300	300	205	805
InstructERC	69.33	58.33	40.49	57.89
GPT-5.4-mini	52.67	**64.00**	37.07	52.92
SRERC	**73.67**	62.00	**40.98**	**60.99**
Diagnostic 2: Deep-context				
# Samples	713	200	70	983
InstructERC	68.02	50.50	38.57	62.36
GPT-5.4-mini	51.75	**73.00**	64.29	56.97
SRERC	**71.67**	59.00	**65.71**	**68.67**
Diagnostic 3: Lexical-bait Neutral				
# Samples	25	94	11	130
InstructERC	48.00	77.66	18.18	66.92
GPT-5.4-mini	**56.00**	53.19	45.45	53.08
SRERC	**56.00**	**80.85**	**54.55**	**73.85**
Diagnostic 4: Affective Turning Point				
# Samples	37	90	163	290
InstructERC	48.65	51.11	**38.65**	**43.79**
GPT-5.4-mini	45.95	**57.78**	27.61	39.31
SRERC	**62.16**	**57.78**	30.67	43.10

Table 5: Context-sensitive evaluation on diagnostic subsets. Model entries report accuracy (%). Combined denotes micro-averaged accuracy.

5.2 Context-Sensitive Evaluation

To evaluate whether SRERC improves the context-sensitive cases targeted by selective reflective correction, we construct four diagnostic subsets and compare SRERC with representative baselines. More details about context-sensitive cases can be found in Appendix A. As shown in Table 5, SRERC achieves the best combined accuracy on three of the four diagnostic settings, including Implicit-Short, Deep-context, and Lexical-bait Neutral. This pattern suggests that the main benefit of SRERC lies not in uniformly changing predictions, but in improving cases where the target emotion is underdetermined by surface text and requires target-specific contextual verification.

Diagnostic 1: Implicit-Short. Implicit-Short cases are short target utterances whose emotion is difficult to infer from lexical content alone and therefore requires contextual interpretation. On this subset, SRERC achieves the best combined accuracy of 60.99%, outperforming InstructERC by $+3.10$ points and GPT-5.4-mini by $+8.07$ points.

The gains on IEMOCAP and EmoryNLP suggest that reflective decoding is useful when the target utterance provides insufficient affective evidence.

Diagnostic 2: Deep-context. Deep-context cases are target utterances whose correct emotion cannot be resolved from the target sentence and adjacent turns alone, and therefore require relevant evidence from earlier turns. On this subset, SRERC obtains the strongest combined accuracy of 68.67%, exceeding InstructERC by $+6.31$ points and GPT-5.4-mini by $+11.70$ points. The gains on IEMOCAP and EmoryNLP indicate that verification probes can help the backbone recover relevant earlier-turn cues, although the MELD result shows that long-context reasoning remains dataset-dependent.

Diagnostic 3: Lexical-bait Neutral. Lexical-bait Neutral cases are gold-neutral utterances that contain polarity-bearing lexical cues identified by a fixed affective lexicon. These cases test whether a model can avoid overreacting to surface emotional words when the surrounding context supports a neutral interpretation. On this subset, SRERC achieves the best combined accuracy of 73.85%, improving over InstructERC by $+6.93$ points and GPT-5.4-mini by $+20.77$ points. This supports our motivation that target-specific contextual verification can reduce lexical shortcut errors in neutral utterances.

Diagnostic 4: Affective Turning Point. Affective Turning Point cases are target utterances whose coarse valence polarity differs from that of the immediately preceding utterance under a fixed label-to-valence mapping. These cases test whether a model can verify when the target utterance departs from the immediately preceding emotional cue. SRERC substantially improves over both baselines on IEMOCAP and ties with GPT-5.4-mini on MELD, but its combined accuracy is slightly lower than InstructERC due to the weaker result on EmoryNLP. This indicates that affective turning points remain difficult: the model must distinguish genuine emotional shifts from cases where local context should still constrain the prediction.

5.3 Sample-Level Correction Analysis

To examine whether the improvement of SRERC comes from meaningful error correction rather than indiscriminate output changes, we compare the first-pass prediction with the final prediction produced by SRERC at the sample level. We distinguish two types of prediction changes: *Wrong* $\rightarrow$ *Correct* cases, where an initially wrong prediction becomes correct, and *Correct* $\rightarrow$ *Wrong* cases, where an initially correct prediction becomes wrong. We further report the net correction, computed as the number of Wrong $\rightarrow$ Correct cases minus the number of Correct $\rightarrow$ Wrong cases, and the error recovery rate, computed as the percentage of first-pass errors corrected by SRERC.

	IEMOCAP	**MELD**	**EmoryNLP**
Total	1623	2610	984
First-pass Correct	1134	1838	462
Final Correct	1154	1868	477
Wrong $\rightarrow$ Correct	113	87	46
Correct $\rightarrow$ Wrong	93	57	31
Net Gain	$+20$	$+30$	$+15$
Error recovery rate	**23.11**	**11.27**	**8.81**
Exact McNemar p	0.185	**0.015**	0.110

Table 6: Sample-level correction statistics of SRERC over the first-pass prediction. McNemar tests are two-sided and exact.

Variant	**IEMOCAP**	**MELD**	**EmoryNLP**	*Avg.*
First-pass only	69.80	69.50	40.75	60.02
w/o Reflection Gate	71.32	70.10	41.00	60.81
w/o SRC	70.12	70.24	42.26	60.87
w/o Scene	71.32	69.97	43.07	61.45
w/o Probes	71.41	70.16	43.17	61.58
Full model	**71.60**	**70.37**	**43.25**	**61.74**

Table 7: Ablation study of SRERC on the three benchmarks in weighted-F1 (%).

Table 6 shows that SRERC yields positive net correction on all three datasets. Wrong $\rightarrow$ Correct cases consistently outnumber Correct $\rightarrow$ Wrong cases, leading to net corrections of $+20$, $+30$, and $+15$ instances on IEMOCAP, MELD, and EmoryNLP, respectively. These correspond to accuracy gains of $+1.23$, $+1.15$, and $+1.52$ points. In addition, SRERC recovers 23.11%, 11.27%, and 8.81% of first-pass errors on the three datasets. Exact paired McNemar tests give p=0.185, 0.015, and 0.110, respectively; the MELD improvement is significant at $p<0.05$, while the smaller IEMOCAP and EmoryNLP test sets do not reach this threshold. These results indicate that reflective decoding does not merely perturb first-pass outputs; its successful revisions consistently exceed the new errors it introduces.

5.4 Ablation Study

Table 7 reports the ablation results of SRERC.

Main Components. We first examine the two main components of SRERC. *w/o Reflection Gate* applies SRC-conditioned reflective decoding to every utterance without verdict-based selection, while *w/o SRC* keeps the reflection gate but removes SRC from the reflective input. Both variants improve over the first-pass-only baseline, increasing the average weighted-F1 from 60.02 to 60.81 and 60.87, respectively. However, both remain below the full SRERC, which reaches 61.74 on average. This indicates that neither verdict-based selection nor an additional reflective decoding pass alone fully explains the gain; the best performance is obtained when selective gating and SRC-conditioned evidence are combined. The lower score of *w/o Reflection Gate* further suggests that applying SRC-conditioned reflective decoding indiscriminately can introduce avoidable regressions, supporting the need for verdict-based selection.

Triggered-Subset and Gating Analysis. Because reflective decoding is applied only to gated utterances, full-set scores dilute the effect on the cases where the mechanism intervenes. On the triggered subset, adding the actual SRC improves weighted-F1 over the first-pass predictions by $+2.9$ and $+1.8$ points on IEMOCAP and MELD, respectively. In a separate empty-SRC re-decoding control, the second pass produces no prediction changes, showing that the gain comes from the additional scene-aware evidence rather than from merely decoding twice. The gate is also directionally useful: on IEMOCAP, reflection yields a net correction of $+4$ cases on gate-triggered utterances, whereas applying the same operation to gate-skipped utterances yields a net correction of -4 cases. Together, these results show that the gate preferentially selects cases that benefit from SRC-conditioned re-decoding and avoids unnecessary changes on the remaining utterances.

SRC Components. We further ablate the two fields of SRC. Removing the scene field reduces the average score from 61.74 to 61.45, while removing the probe field reduces it to 61.58. The drops are smaller than those caused by removing the main components, but both fields still contribute to the full model. This suggests that situational grounding and verification probes provide useful and non-redundant information, with the scene description contributing slightly more in this setting.

Verdict/SRC teacher	IEMOCAP	MELD	EmoryNLP
GPT-5.4-mini	**71.60**	70.37	43.25
GLM-5.1	71.48	**70.47**	**43.44**

Table 8: Teacher robustness after replacing both auxiliary teachers while keeping the ERC backbone and gate unchanged (weighted-F1, %).

IEMOCAP							
	Hap.	Sad	Neu.	Ang.	Exc.	Fru.	–
GPT-5.4-mini	49.84	69.73	62.49	44.31	49.77	61.29	–
InstructERC†	**56.39**	82.42	67.39	62.02	69.22	62.38	–
First-pass only	52.41	84.85	67.74	63.52	**73.07**	69.03	–
SRERC	55.75	**86.20**	**72.38**	**66.85**	71.67	**69.47**	–
MELD							
	Neu.	Joy	Sur.	Sad.	Ang.	Dis.	Fear
GPT-5.4-mini	71.64	57.23	59.16	44.87	56.69	**45.22**	33.06
InstructERC†	80.90	64.68	**61.56**	39.74	55.17	42.04	**35.64**
First-pass only	81.10	67.25	60.75	50.55	60.88	37.74	26.67
SRERC	**82.03**	**67.53**	60.74	**50.99**	**62.54**	39.66	30.99
EmoryNLP							
	Joy.	Mad	Neu.	Pea.	Pow.	Sad	Sca.
GPT-5.4-mini	56.00	43.24	52.86	1.69	7.21	21.36	30.49
InstructERC†	53.23	36.82	52.69	**16.37**	**17.49**	31.37	**39.71**
First-pass only	59.57	**44.79**	56.99	1.69	1.92	**35.19**	35.11
SRERC	**61.32**	43.82	**58.84**	8.20	10.53	34.29	36.36

Table 9: Per-class F1 (%) on the three benchmarks. †Per-class scores for InstructERC are reproduced by us because the original paper reports only dataset-level weighted-F1.

Teacher Robustness and Self-Teacher. As shown in Table 8, replacing GPT-5.4-mini with GLM-5.1 leaves performance essentially unchanged across all three datasets, indicating that the mechanism is not tied to one proprietary teacher. We also tested the fine-tuned ERC backbone itself in the verdict and SRC roles. Because it was trained to produce an emotion label rather than structured verdict and context outputs, it produced no usable auxiliary signal and the pipeline fell back to the first-pass prediction.

5.5 Fine-grained Emotion Resolution

Table 9 reports per-class F1 on the three benchmarks. We compare the zero-shot general-purpose LLM GPT-5.4-mini, the ERC-specific LLM baseline InstructERC, the first-pass-only backbone, and the full SRERC. Per-class scores of InstructERC are reproduced by us because the original paper reports only dataset-level weighted-F1. Overall, SRERC improves over the first-pass-only backbone on most emotion categories, suggesting that SRC-conditioned reflective decoding does not simply shift predictions toward a few dominant labels. The

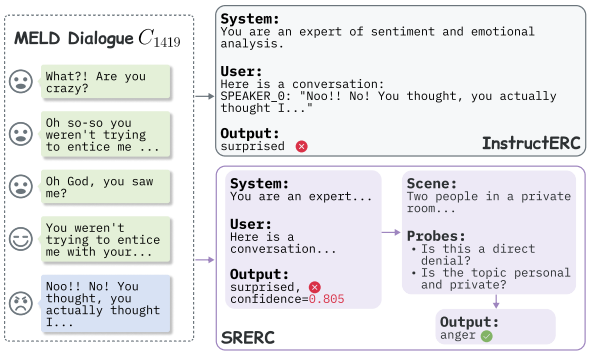


Figure 3: Case study on InstructERC and SRERC.

gains are especially visible on context-sensitive or low-resource categories. On EmoryNLP, the first-pass-only backbone obtains only 1.69 and 1.92 F1 on *peaceful* and *powerful*, respectively, while SRERC improves them to 8.20 and 10.53. On MELD, SRERC improves the rare *fear* class from 26.67 to 30.99. On IEMOCAP, the *neutral* class improves from 67.74 to 72.38, indicating better handling of utterances whose surface form alone is insufficient for emotion prediction. These results complement the hard-case analysis in Section 5.2 by showing that the benefit of SRERC also appears at the emotion-category level.

5.6 Case Study

Figure 3 presents a MELD example where the target utterance "`Noo!! No! You thought, you actually thought I...`" is labeled as *anger*. InstructERC predicts *excited*, likely distracted by the surrounding surprised reactions and the exclamatory surface form. By contrast, SRERC first produces the same erroneous first-pass prediction, but the reflection gate triggers an SRC-conditioned second pass. With SRC highlighting a misunderstood personal boundary and a direct denial, SRERC revises the prediction to the correct label. This example illustrates how selective reflective correction helps the model move beyond nearby-context interference and surface cues, and verify the target utterance against target-specific situational evidence.

6 Conclusion

In this work, we studied a context-sensitive reliability issue in LLM-based emotion recognition in conversation: first-pass predictions may be confidently misaligned when the target emotion depends on affective shifts or situational cues that are only weakly expressed in the utterance itself. We proposed SRERC, a selective reflective correction framework that uses a verdict-guided gate to decide whether a prediction should be revisited and a Scene-aware Reflection Context to provide target-specific situational evidence for gated cases. Experiments on IEMOCAP, MELD, and EmoryNLP show that SRERC achieves the best average weighted-F1. Further analyses show clear gains on Implicit-Short, Deep-context, and Lexical-bait Neutral cases, while sample-level correction and ablation results confirm that the gains come from beneficial reflective revisions and from combining selective revision with SRC-provided evidence.

Limitations

Despite its empirical gains, SRERC has three practical limitations. First, it introduces additional inference overhead. Although the verdict-guided gate avoids applying reflective decoding to every utterance, each gated case still requires an additional pass through the LLM backbone. This makes SRERC less suitable for strictly real-time scenarios, such as streaming dialogue systems with tight latency constraints. Reducing this cost through more efficient reflective decoding strategies or activation caching is an important direction for future work.

Second, SRERC is currently text-only. This is a natural starting point for LLM-based ERC, but some datasets contain acoustic and visual signals that can be essential for recognizing implicit emotions. Cases involving prosody, facial expressions, or other non-verbal cues may therefore remain unresolved when the textual context is insufficient. Extending SRC with multimodal evidence is a promising way to raise the perceptual ceiling of the framework.

Third, SRERC uses a lightweight binary reflection gate rather than a full reliability triage system. Unsupported low-confidence predictions are not sent to reflective decoding because they provide neither a supported hypothesis to consolidate nor a confident unsupported hypothesis to challenge. Future work may handle this regime with explicit escalation mechanisms, such as abstention, multi-teacher adjudication, or iterative verification.

References

Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. 2025. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. *Proceedings*

of the AAAI Conference on Artificial Intelligence, 39(11):11418–11426.

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

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

DeepSeek-AI. 2026. Deepseek v4 preview release. Released April 24, 2026. Accessed: 2026-05-19.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods. *Electronics*, 12(22).

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, Abu Dhabi, UAE. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 flash: Frontier intelligence built for speed. Released December 2025,Accessed: 2026-05-19.

Zhinan Gou, Yuchen Long, Jieli Sun, and Kai Gao. 2025. Tg-erc: Utilizing three generation models to handle emotion recognition in conversation tasks. *Expert Systems with Applications*, 268:126269.

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.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. 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 5666–5675, Online. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Exploiting Unsupervised Data for Emotion Recognition in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4839–4846, Online. Association for Computational Linguistics.

Ran Jing, Geng Tu, Yice Zhang, and Ruifeng Xu. 2026. Causal-erc: A multimodal framework with causal prompting for emotion recognition in conversations with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31383–31391.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2024. InstructERC: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *Preprint*, arXiv:2309.11911.

Jiang Li, Xiaoping Wang, Yingjian Liu, and Zhigang Zeng. 2024a. Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*, 15(4):1919–1933.

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

Jiang Li, Xiaoping Wang, and Zhigang Zeng. 2025. 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.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1204–1214, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. 2023. Skier: A symbolic knowledge integrated model for conversational emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13121–13129.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.

Cam-Van Thi Nguyen, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. 2024. Ada2i: Enhancing modality balance for multimodal conversational emotion recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pages 9330–9339, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2026. Introducing gpt-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/. Released March 2026,Accessed: 2026-05-19.

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

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13789–13797.

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

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

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. 2025. Multimodal large language models meet multimodal emotion recognition and reasoning: A survey. *Preprint*, arXiv:2509.24322.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

ChengYan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6257–6274, Suzhou, China. Association for Computational Linguistics.

Xiaomi MiMo Team. 2026. Xiaomi mimo-v2.5-pro. Released April 27, 2026. Accessed: 2026-05-19.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. 2024. BiosERC: Integrating biography speakers supported by LLMs for ERC tasks. In *Artificial Neural Networks and Machine*

Learning – ICANN 2024, pages 277–292, Cham. Springer Nature Switzerland.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Sayyed M. Zahiri and Jinho D. Choi. 2018. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, volume WS-18 of *AAAI Technical Report*, pages 44–52. AAAI Press.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2025. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks*, 192:107901.

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

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

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

A Context-Sensitive Diagnostic Cases

To probe ERC behavior beyond aggregate weighted-F1, we evaluate models on four context-sensitive diagnostic subsets that target distinct failure modes. All subsets are derived from the dialogue input and gold labels, and are therefore model-agnostic and computed once for each dataset. These subsets are used only for analysis and are not used for model training, model selection, or threshold tuning. For each benchmark, the emotion-label set $\mathcal{E}$ is partitioned into a positive pole $\mathcal{E}^+$ and a negative pole $\mathcal{E}^-$:

$$\begin{aligned}\mathcal{E}^+ = \{&\textit{happy}, \textit{joy}, \textit{joyful}, \textit{excited},\\ &\textit{peaceful}, \textit{powerful}\},\\ \mathcal{E}^- = \{&\textit{sad}, \textit{sadness}, \textit{angry}, \textit{anger}, \textit{mad},\\ &\textit{frustrated}, \textit{fear}, \textit{scared}, \textit{disgust}\}.\end{aligned}$$

The labels *neutral* and *surprise* are treated as non-polar and do not satisfy the polarity conditions below. We also use two small affective lexicons:

$$\begin{aligned}\mathcal{L}^+ = \{&\textit{happy}, \textit{great}, \textit{love}, \textit{lovely}, \textit{wonderful},\\ &\textit{fantastic}, \textit{amazing}, \textit{awesome}, \textit{joy},\\ &\textit{excited}, \textit{thanks}, \textit{thank you}, \textit{yay},\\ &\textit{haha}, \textit{perfect}, \textit{delighted}, \textit{glad},\\ &\textit{cheerful}\},\\ \mathcal{L}^- = \{&\textit{sad}, \textit{angry}, \textit{anger}, \textit{hate}, \textit{afraid},\\ &\textit{scared}, \textit{damn}, \textit{sorry}, \textit{stop}, \textit{shut up},\\ &\textit{fuck}, \textit{fight}, \textit{argue}, \textit{stupid}, \textit{kill},\\ &\textit{cry}, \textit{frustrated}, \textit{annoyed}, \textit{upset}\}.\end{aligned}$$

Given a target utterance u_t at position t in dialogue $C = \langle u_1, \ldots, u_N \rangle$, with gold label y_t and gold label history $\{y_1, \ldots, y_{t-1}\}$, u_t is selected into a diagnostic subset if it satisfies all conditions of that subset. The four subsets are not constrained to be disjoint, but in practice their selection rules yield largely non-overlapping utterances.

Diagnostic 1: Implicit-Short. A target utterance u_t is selected if (i) u_t contains at most four whitespace-tokenized words, and (ii) its gold label is polar, i.e., $y_t \in \mathcal{E}^+ \cup \mathcal{E}^-$. This subset probes whether a model can recover a polar emotion from an extremely short and lexically underspecified target utterance.

Diagnostic 2: Deep-context. A target utterance u_t is selected if (i) $t \geq 6$, meaning that at least five preceding utterances are available in the dialogue history, (ii) u_t contains at least five whitespace-tokenized words, and (iii) its gold label is polar, i.e., $y_t \in \mathcal{E}^+ \cup \mathcal{E}^-$. This subset targets cases where the model has access to a longer preceding context and must avoid relying only on the target utterance or immediately adjacent turns.

Diagnostic 3: Lexical-bait Neutral. A target utterance u_t is selected if (i) $y_t = \textit{neutral}$, and (ii) the lower-cased surface form of u_t contains at least one token or multi-token phrase from $\mathcal{L}^+ \cup \mathcal{L}^-$.

The surface text therefore contains affective lexical cues that may bait a polar prediction, while the gold label remains neutral. This subset tests whether a model can override surface lexical signals using the surrounding dialogue context.

Diagnostic 4: Affective Turning Point. A target utterance u_t is selected if (i) y_t is polar, and (ii) the immediately preceding gold label y_{t-1} is also polar but has the opposite valence to y_t. Formally, exactly one of $\{y_{t-1}, y_t\}$ belongs to $\mathcal{E}^+$ and the other belongs to $\mathcal{E}^-$. This subset captures local affective pivots inside a dialogue, where the immediately preceding emotional cue may mislead models toward the previous opposite-valence state.

B Detailed Dataset Descriptions

IEMOCAP The *Interactive Emotional Dyadic Motion Capture* (Busso et al., 2008) corpus is one of the most widely adopted resources for affective computing research. It consists of approximately twelve hours of audiovisual recordings produced by ten professional actors (five male and five female) performing both improvised and scripted dyadic scenarios across five sessions. Following the standard ERC convention adopted by recent LLM-based work (Lei et al., 2024; Xue et al., 2024), we use the six-category subset of emotion labels (*angry*, *happy*, *neutral*, *sad*, *excited*, and *frustrated*). The first four sessions are used for training, with a small held-out portion for validation, and the fifth session is reserved for testing, yielding $108/12/31$ conversations and $5163/647/1623$ utterances. Although the dataset is recorded multimodally, we use only the textual transcription in this work.

MELD The *Multimodal EmotionLines Dataset* (Poria et al., 2019a) is an extension of the textual EmotionLines corpus, augmented with aligned audio and visual signals, and built upon roughly $1{,}400$ multi-party conversations and $13{,}000$ utterances sampled from the sitcom *Friends*. Each utterance is annotated with one of Ekman's six basic emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*) plus a *neutral* class. The official split contains $1038/114/280$ conversations and $9989/1109/2610$ utterances for training, validation, and testing, respectively. Compared with IEMOCAP, MELD is markedly more challenging in two respects: (i) it is multi-party, with up to eight distinct speakers per conversation, and (ii) the class distribution is highly imbalanced, with the *neutral* class accounting for over 47% of all utterances and minority classes such as *fear* below 3%. We use only the textual modality.

EmoryNLP The EmoryNLP (Zahiri and Choi, 2018) corpus is also drawn from the *Friends* TV series but adopts an annotation scheme based on the Plutchik wheel and a primary–color decomposition of emotions. It contains seven emotion categories: *joyful*, *peaceful*, *powerful*, *neutral*, *sad*, *scared*, and *mad*. The corpus is partitioned into $659/89/79$ conversations and $7551/954/984$ utterances for training, validation, and testing, respectively. EmoryNLP is the most fine-grained of the three benchmarks: several of its label names—particularly *peaceful* and *powerful*—are not strictly correlated with explicit lexical sentiment markers, making the corpus a stringent test of a model's ability to recover context-dependent emotional states from situational evidence rather than surface text alone.

Licenses IEMOCAP is distributed by USC under a proprietary license that requires a formal data-access agreement with the SAIL laboratory and restricts usage to non-commercial academic research. MELD is released under the GNU General Public License v3.0 (GPL-3.0).[*] The EmoryNLP emotion-detection corpus is released under the Apache License 2.0 as part of the Character Mining project.[†] Our use of all three datasets is consistent with their respective license terms.

C Detailed Baseline Descriptions

C.1 Discriminative ERC models

DialogueRNN (Majumder et al., 2019) jointly tracks party state, global dialogue context, and emotional state with three coupled GRUs. **DialogueGCN** (Ghosal et al., 2019) models intra- and inter-speaker dependencies as edges in a relational graph convolutional network. **COSMIC** (Ghosal et al., 2020) enriches the conversational representation with commonsense relations extracted from a frozen COMET knowledge generator. **DAG-ERC** (Shen et al., 2021b) encodes the dialogue as a directed acyclic graph in which each utterance attends to a fixed-size predecessor window. **SKAIG** (Li et al., 2021) integrates psychological commonsense covering past, present, and future

[*] https://github.com/declare-lab/MELD

[†] https://github.com/emorynlp/emotion-detection

emotional states of the speaker. **CauAIN** (Zhao et al., 2022) introduces a causal-aware interaction module that explicitly traces the causal contributors of each emotional trigger. **DialogXL** (Shen et al., 2021a) extends XLNet with utterance-level recurrence and dialogue-aware self-attention designed for multi-party conversations. **SPCL** (Song et al., 2022) casts ERC as supervised prototypical contrastive learning under a difficulty-aware curriculum that mitigates class imbalance.

C.2 General-purpose LLMs

We additionally evaluate four flagship closed-source LLMs in a strict zero-shot setting, using the same prompt template as our backbone but without any fine-tuning: **GPT-5.4-mini** (OpenAI, 2026), **MiMo-V2.5-Pro** (Xiaomi MiMo Team, 2026), **DeepSeek-V4-Pro** (DeepSeek-AI, 2026), and **Gemini-3-Flash** (Google DeepMind, 2025). All four models are queried at temperature 0.2 with the same dialogue context window of $W{=}5$ utterances used by SRERC. Their outputs are parsed into each dataset's emotion vocabulary by exact string matching against the canonical label names.

C.3 ERC-specific LLMs

InstructERC (Lei et al., 2024) reformulates ERC as instruction-following generation on top of a fine-tuned LLM and additionally augments the prompt with retrieved similar examples and an auxiliary speaker-identification task. **BiosERC** (Xue et al., 2024) attaches an offline-produced speaker biography to every prompt, providing the backbone with a per-speaker persona. **LaERC-S** (Fu et al., 2025) pushes speaker characterization further by exposing the LLM to a structured speaker profile during fine-tuning. **Causal-ERC** (Jing et al., 2026) integrates causal reasoning into the LLM prompt so that the model can leverage cause–effect chains across utterances. These LLM-based baselines, like SRERC, build on fine-tuned LLM backbones, but they mainly inject static auxiliary signals before label generation. None of them performs verdict-gated reflective decoding or utterance-level dynamic context injection.

D Training Details

We use LLaMA-2-13B as the first-pass ERC backbone and employ QLoRA for parameter-efficient fine-tuning. One-time speaker profiles are generated using a frozen MiMo-V2.5-Pro model and remain fixed during both training and inference. The first-pass ERC backbone is optimized with AdamW under a linear learning-rate scheduler. We use an initial learning rate of 3×10^{-4} with a warm-up ratio of 0.03 and train the model for 3 epochs with an effective batch size of 16. Following standard QLoRA settings, the backbone adopts 4-bit NF4 quantization with double quantization enabled. For parameter-efficient adaptation, LoRA adapters are applied to all linear projection layers with rank $r = 32$, scaling factor $\alpha = 128$, and dropout probability 0.05. In addition, NEFTune embedding perturbation with $\alpha = 5$ is incorporated to improve robustness during low-rank fine-tuning.

For SRERC-specific settings, we use a local context window consisting of five preceding utterances. The Scene-aware Reflection Context (SRC) generator and verdict teacher are instantiated using GPT-5.4-mini, and all teacher-generated outputs are produced offline without access to gold labels. During inference, reflective decoding is triggered only for predictions selected by the verdict-guided gate, including supported predictions and high-confidence unsupported predictions. Dataset-specific settings differ only in maximum sequence length and the confidence threshold. The maximum sequence length is set to 1024 for IEMOCAP and EmoryNLP, and 2048 for MELD. The confidence threshold τ is set to 0.6 on IEMOCAP and 0.4 on both MELD and EmoryNLP. Best-model selection is based on weighted-F1 on the validation set.

All experiments are conducted on a single NVIDIA RTX 5090 GPU with QLoRA 4-bit quantization and bf16 mixed precision. A complete pipeline run takes approximately 2 hours for IEMOCAP, 12 hours for MELD, and 5 hours for EmoryNLP.

E Sensitivity to the Confidence Threshold

The confidence threshold τ in Eq. (7) is the only free hyperparameter introduced by SRERC. We sweep it across a meaningful operating range on each dataset and report the resulting weighted-F1 in Figure 4. The optimum is reached at $\tau{=}0.6$ on IEMOCAP and at $\tau{=}0.4$ on both MELD and EmoryNLP. This pattern is consistent with the corpus characteristics: the dyadic dialogues of IEMOCAP contain fewer genuinely ambiguous utterances and therefore favor a more conservative gate, whereas the multi-party, scene-rich exchanges of MELD and EmoryNLP yield a larger pool of high-confidence but contextually contested predictions,

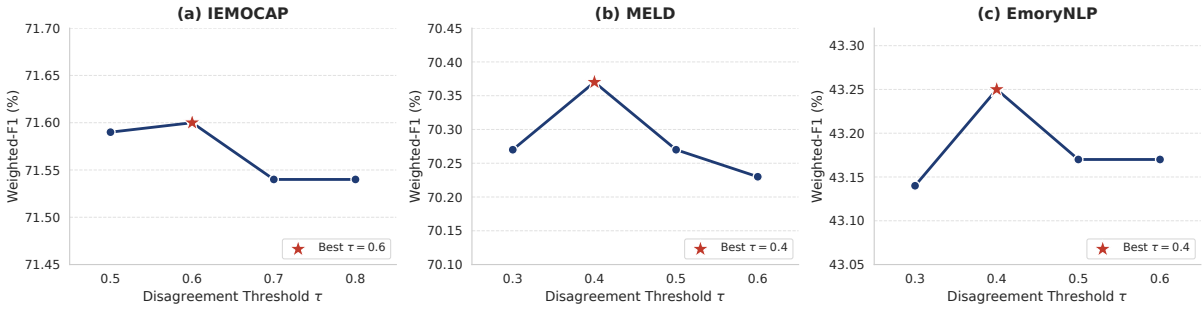


Figure 4: Sensitivity analysis of the confidence threshold τ on the three benchmarks (weighted-F1, %).

which a lower threshold captures more fully. Beyond these optima, however, SRERC remains relatively stable across nearby threshold values, indicating that the method is not overly sensitive to τ.

F Use of AI Assistants

AI assistants were used only for limited writing support, including language polishing, formatting refinement, and consistency checking. All research ideas, method designs, model implementations, experimental analyses, and conclusions were independently developed and verified by the authors.

G Discussion on Potential Risks

Although SRERC is developed for research on emotion recognition in conversation, emotion prediction may still be inaccurate, oversimplified, or misinterpreted in real-world use, especially for implicit, sarcastic, or culturally dependent expressions. The reflective correction process may also make predictions appear more reliable than they actually are, while speakers' emotional states can remain inherently ambiguous. Therefore, SRERC should not be directly used in high-stakes scenarios such as mental-health assessment, recruitment, education, or legal judgment without human oversight and further validation. Our experiments use only publicly available benchmark datasets and do not involve new personal data collection or real-world deployment.

H Prompt Templates

This section reports schematic templates used by the ERC backbone, verdict teacher, and SRC generator; dataset-specific formatting instructions are omitted for readability. Bracketed fields are replaced for each instance. The verdict rationale and judgment in Eqs. (5)–(6) are returned in one structured call.

First-pass ERC backbone prompt.

```
[System]
You are an expert at analyzing the emotion of
    utterances among
speakers in a conversation.
Given the characteristic of this speaker, [
    TARGET_SPEAKER]:
[SPEAKER_PROFILE]
Given the following conversation as a context:
[DIALOGUE_WINDOW]

[User]
Based on the above conversation and speaker
    characteristics, which
emotional label describes [TARGET_SPEAKER] in
    the utterance
"[TARGET_UTTERANCE]"?

[Assistant output]
[PREDICTED_EMOTION_LABEL]
```

Verdict teacher prompt. The following template receives the first-pass hypothesis but no gold label. Dataset-specific versions differ only in the dataset description and short confidence-calibration guidance.

```
[System]
You are a senior emotion annotator for [
    DATASET]. A smaller model has
already produced a FIRST-PASS prediction for
    the target utterance.
Supply richer context so the smaller model can
     re-decide, especially
when a literal reading has misled it.

Return strict JSON only:
{
  "scene": "one sentence (<=20 words) on
    setting/stakes/register",
  "speaker_vibe": "2--3 descriptive adjectives
    ",
  "surface_vs_deep": "<=25 words comparing
    literal and contextual readings",
  "correction_hint": {
    "agrees_with_baseline": true | false,
    "if_disagree_signal_toward": "<=15-word
    behavioral description"
```

```
  },
  "decisive_cues": ["grounded cue 1", "
     optional grounded cue 2"]
}

Base agrees_with_baseline on the target
utterance, preceding turns, and baseline
     confidence. Each decisive cue
must quote or paraphrase concrete dialogue
     evidence.

[User]
[Speaker description]
[SPEAKER_PROFILE]
[Recent conversation window]
[DIALOGUE_WINDOW]
[Target speaker] [TARGET_SPEAKER]
[Target utterance] "[TARGET_UTTERANCE]"
[First-pass prediction] [PREDICTED_LABEL]
[First-pass confidence] [CONFIDENCE]
```

Prediction-agnostic SRC generator prompt. This prompt omits both the first-pass label and its confidence.

```
[System]
You are an annotator providing very brief
     context hints for an emotion
recognition model on [DATASET]. Describe the
     social context, including
the setting, topic, speaker relationship, role
      in the turn, and
formality, to support emotion prediction.

Return strict JSON with exactly two keys:
{
  "atmosphere": "one concise scene sentence
     (<=20 words)",
  "self_check": ["evidence question 1",
                 "optional evidence question
     2"]
}
Keep each question within 12 words and the
     total below 50 words.

[User]
[Speaker description]
[SPEAKER_PROFILE]
[Conversation window]
[DIALOGUE_WINDOW]
[Target speaker] [TARGET_SPEAKER]
[Target utterance] "[TARGET_UTTERANCE]"
```

Reflective decoding prompt. For a gated item, the backbone receives the same first-pass prompt with the following context block appended to its system message and predicts the emotion label again.

```
### Scene atmosphere:
[SCENE_ATMOSPHERE]

### Self-check (do silently, do not output):
- [PROBE_1]
- [OPTIONAL_PROBE_2]
```