Why Do Emotions Change? Appraisal-Guided Reasoning for Emotion–Cause Triplet Extraction in Conversations

Qiao Liang[2] **Ying Shen**[2*] **Yao Liu**[2] **Tiantian Chen**[2] **Lin Zhang**[1,2]
[1]IEIT SYSTEMS (Beijing) Co., Ltd.
[2]School of Computer Science and Technology, Tongji University
{2333091, yingshen, 2253215, 2111287, cslinzhang}@tongji.edu.cn [†]

Abstract

Multimodal Emotion–Cause Triplet Extraction in Conversations (MECTEC) is fundamental for fine-grained affect understanding, yet it remains challenging in multi-turn, multi-speaker settings. Existing methods often make locally plausible predictions but struggle to maintain conversation-level consistency under within-speaker emotion shifts and core events. To address this, we propose **ECFlow**, a unified framework that combines appraisal-guided structured generation with graph-structured reinforcement learning. ECFlow operationalizes cognitive appraisal theory into a controllable intermediate reasoning trace and constructs **UMECS**, a unified supervision dataset with cognitively grounded traces. It then lifts predicted and gold triplets into an Emotion–Cause Flow Graph and optimizes verifiable, structure-aware rewards for emotion-shift coherence and core-event consistency, together with task-oriented triplet rewards. Experiments on public MECTEC benchmarks show that ECFlow consistently outperforms strong baselines, achieving state-of-the-art triplet extraction and improved structure-aware metrics on emotion shifts and core events. Our code and dataset are available at `https://github.com/redifinition/ECFlow`.

1 Introduction

Multimodal Emotion Cause Triplet Extraction in Conversations (MECTEC) aims to extract (*emotion utterance, cause utterance(s), emotion label*) triplets from multi-turn multimodal conversations (Wang et al., 2022). Unlike emotion recognition tasks, it requires modeling long-range context and evolving speaker states, which is crucial for affective applications such as empathetic dialogue systems and mental-health support (Tao and Tan, 2005; Soleymani et al., 2017; Peng et al., 2025).

Despite its importance, MECTEC faces two fundamental challenges, one of which is speakers' *emotion shifts* across conversation turns. Frequent emotion shifts break within-speaker consistency and induce incorrect emotion–cause links. Yet existing approaches (Wang et al., 2022; Ding et al., 2020a,b; Liang et al., 2025) still predict triplet elements independently without explicitly modeling within-speaker emotional trajectories.

Second, emotional dynamics in conversation often revolve around **core events**. Core events are pivotal utterances that introduce key state changes (e.g., decisions, conflicts, revelations, or responsibility attribution) and serve as shared causal sources for multiple speakers' emotional responses across subsequent turns. In MECTEC, core events act as conversation-level causal anchors, enabling models to align emotions across turns and speakers to the same underlying cause. Without this perspective, models are easily misled by locally salient but non-causal cues, fragmenting emotions triggered by the same event across different utterances. Recent studies use LLMs via prompting or supervised fine-tuning (SFT) (Luo et al., 2024; Zhang et al., 2024), but most do not model core events or enforce conversation-level causal structure, leading LLMs to default to locally salient cues (Yang et al., 2024). As a result, emotions triggered by the same event may be attributed to different utterances, yielding fragmented and distorted triplet assignments.

Cognitive appraisal theory (Lazarus, 1982) explains emotions as outcomes of subjective appraisals with respect to personal goals, beliefs, and contextual factors. Under this view, emotion shifts correspond to dynamic reappraisals of unfolding events, often signaled by subtle non-verbal cues, which is largely overlooked by existing MECTEC

*Corresponding authors.

†This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant 2025ZD0123701, in part by the National Natural Science Foundation of China under Grant 62476202 and 62272343, and in part by the Fundamental Research Funds for the Central Universities.

models. This theoretical insight is particularly relevant to the second challenge of MECTEC, namely the lack of explicit reasoning mechanisms for linking emotions to their causes. Without a principled account of how emotions arise and change, models are prone to spurious correlations especially in conversations involving multiple events or speakers.

We therefore propose the **Appraisal-Guided Reasoning Paradigm**. Rather than treating emotions as independent labels, our approach guides the model to reason about how speakers appraise and reappraise events across conversation turns, enabling coherent tracking of speaker-level emotional trajectories. This theory-guided reasoning paradigm allows the model to better disambiguate emotion–cause relations under frequent emotion shifts and complex conversational dynamics.

While appraisal-guided reasoning provides a structured inductive bias, SFT alone is insufficient: it favors locally plausible traces without verifying global consistency, often violating within-speaker emotion shift coherence and core event consistency. To bridge this gap, we propose our LLM-based MECTEC framework **Emotion-Cause Flow** (ECFlow) by employing Reinforcement Learning with Verifiable Reward (RLVR) (Shao et al., 2024). We transform abstract constraints including emotion shift coherence and core event consistency into rigorous, structure-aware rewards, which enables ECFlow to optimize global coherence without incentivizing degenerate behaviors.

The contributions of this work are as follows:

- We propose an Appraisal-Guided Reasoning Paradigm that operationalizes cognitive appraisal theory into a structured and controllable intermediate reasoning trace, making within-speaker emotion shifts and core-event–based causal attribution explicit and thus amenable to verifiable, structure-aware optimization for MECTEC. We further build **UMECS**, which augments existing benchmarks with cognitively grounded traces to support training under this paradigm.
- We propose **ECFlow**, an RLVR framework designed to overcome the myopia of standard SFT. By abstracting conversation dynamics into an ECFlow Graph, we verify discrete structural constraints, specifically emotion shift coherence and core-event causality, converting them into rigorous rewards to directly optimize long-horizon logical consistency.
- Extensive experiments on ECF and MECAD demonstrate that ECFlow achieves new state-of-the-art performance. Notably, even in the *Zero-RL* setting (without SFT training), ECFlow still substantially outperforms the strongest prior baseline, highlighting the effectiveness of our method.

2 Related Work

2.1 Multimodal Emotion Cause Triplet Extraction

Prior MECTEC work falls into two lines: discriminative link prediction and LLM-based generation. Early works predominantly cast MECTEC as discriminative link prediction over candidate emotion–cause pairs, often supported by sequence or graph encoders. For instance, Li et al. (2024) introduces transition-style constraints to regularize predictions across turns, while Liang et al. (2025) and Hu et al. (2024) use multimodal graph representations to aggregate cross-utterance evidence. However, they treat links largely independently, without enforcing a global emotion–cause flow.

The SemEval-2024 Task 3 (Wang et al., 2024a) further promoted LLM-based solutions for MECTEC-style extraction. Leading systems, such as Samsung (Zhang et al., 2024) and NUS-Emo (Luo et al., 2024), apply instruction tuning and modular pipelines to incorporate multimodal signals (e.g., vision encoders or multimodal embeddings) for triplet generation. These systems are mostly SFT-based and can produce well-formed outputs while still violating long-horizon consistency, leading to locally plausible but globally inconsistent cause assignments (Yang et al., 2024). Emotion dynamics have also been studied in Emotion Recognition in Conversations (ERC), which tracks speakers' emotion evolution across turns (Poria et al., 2019; Wang et al., 2024b; Bansal et al., 2022). Unlike ERC, MECTEC requires causal attribution, which motivates us to explicitly enforce emotion shift coherence and core event consistency in triplet extraction.

2.2 Reasoning in Affective Computing

Traditional affective computing methods often cast the problem as static classification, overlooking the underlying cognitive processes. Cognitive Appraisal Theory (Lazarus, 1982) suggests that emotions arise from continuous evaluations of unfolding events, motivating explicit appraisal-based reasoning for interpretability. Recent LLM-based approaches typically rely on prompting or supervised fine-tuning (SFT) to generate rationales or structured outputs, but such rationales can be unconstrained and inconsistent with task constraints (Yang et al., 2024). More fundamentally, SFT optimizes token-level likelihood and may not adequately enforce conversation-level structural objectives (e.g., cross-turn coherence and shared causal anchors). To address this mismatch, recent advances in LLM alignment adopt reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024) to optimize solution validity beyond token-level supervision. However, adapting RLVR to open-ended conversational extraction with verifiable structural constraints remains underexplored. Our work bridges this gap by introducing **ECFlow**, which lifts appraisal-driven emotion dynamics into an *Emotion–Cause Flow Graph* and optimizes verifiable, graph-structured rewards for emotion-shift coherence and core-event consistency.

3 Methodology

3.1 Task Definition

Given a multi-speaker conversation $C = \{(S_i, U_i)\}_{i=1}^n$, where S_i denotes the speaker of the i-th utterance U_i. Each utterance is represented as $U_i = (U_i^t, U_i^a, U_i^v)$, corresponding to the text, audio, and visual modalities, respectively. MECTEC aims to extract all emotion–cause triplets in C:

$$\mathcal{P} = \{(U_j^e, U_j^c, y_j^e)\}_{j=1}^m, \quad (1)$$

where U_j^e is an emotion utterance with non-neutral emotion label y_j^e, U_j^c is the corresponding cause utterance, which typically precedes U_j^e but may also appear after it (post-cause). $y_j^e \in \{\text{Anger}, \text{Disgust}, \text{Fear}, \text{Joy}, \text{Sadness}, \text{Surprise}\}$ follows Ekman's basic emotions (Ekman, 1992). An utterance may participate in multiple triplets (e.g., multiple causes for one emotion, or one cause triggering multiple emotions).

Optionally, for LLM-based models, they may output a reasoning trace $\mathcal{T}$. In our framework we generate structured traces while evaluation is conducted on the extracted triplets.

3.2 Appraisal-Guided Reasoning Paradigm

> *"Cognitive appraisal means that the way one interprets an event at any given moment is crucial to the emotional response." — Lazarus, 1982*

Cognitive appraisal theory posits that emotions arise from a speaker's subjective evaluation of perceived cues in a social context. Standard LLM-based approaches often map the conversation context directly to triplets ($C \to \mathcal{P}$), bypassing this critical cognitive process. This direct mapping is hard to control under strict constraints and often yields cross-turn inconsistencies, such as missing subtle emotion shifts or misidentifying the core event of multiple reactions.

Therefore, we formalize MECTEC as a structured intermediate reasoning process. The goal is not to elicit unconstrained explanations, but to expose intermediate structural decisions that can later be supervised and indirectly verified through conversation-level rewards. We introduce an explicit appraisal trace $\mathcal{T}$ as a controllable bridge between the multimodal context C and the final triplet set $\mathcal{P}$. Concretely, our model generates $\mathcal{T}$ and $\mathcal{P}$ in a single structured output, which can be viewed as a factorization:

$$P(\mathcal{T}, \mathcal{P} \mid C) = P(\mathcal{T} \mid C)\, P(\mathcal{P} \mid \mathcal{T}, C). \quad (2)$$

Here, $P(\mathcal{T}|C)$ models the generation of the appraisal path, and $P(\mathcal{P}|\mathcal{T}, C)$ represents constrained triplet generation conditioned on the trace.

We design the structure of $\mathcal{T}$ to mirror the human cognitive process, focusing explicitly on capturing emotion shifts and core events:

(1) Macro-Contextual Appraisal ($\mathcal{A}_{\text{macro}}$). Before analyzing specific utterances, the model evaluates the conversational environment based on C, identifying cultural norms and social dynamics. This step provides a shared interpretive frame for resolving ambiguous multimodal cues in later turns.

(2) Sequential Reappraisal ($\mathcal{A}_{\text{reup}}$). Crucially, this step targets the accurate modeling of **emotion shifts**. We treat emotion as a dynamic flow where a speaker's state at utterance U_i depends on their state at the previous utterance by the same speaker. For each utterance U_i, the model explicitly asks: *"Given the speaker's previous state, does the new stimulus trigger a reappraisal?"* By forcing the model to articulate the transition (e.g., *Stability* vs. *Shift*), we ensure the extracted trajectory is coherent rather than fragmented.

(3) Causal Attribution ($\mathcal{A}_{\text{cause}}$). Based on the shifts identified in $\mathcal{A}_{\text{reup}}$, this step links emotions to antecedent triggers, facilitating the identification of **core events**. Instead of predicting isolated triplets, the model attributes reactions to specific

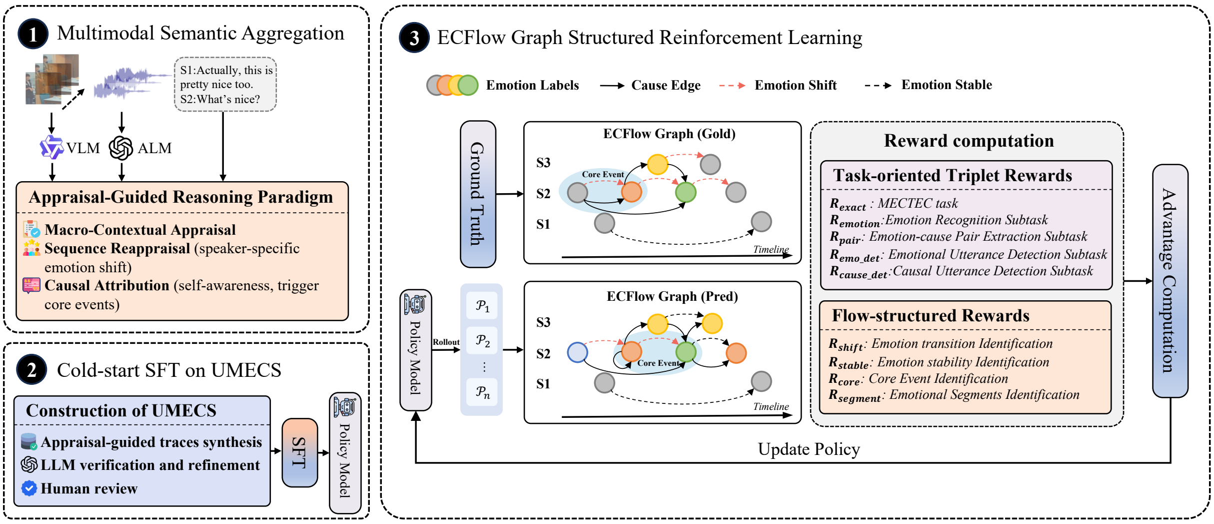

Figure 1: The framework of proposed ECFlow. (1) Stage1: Multimodal Semantic Aggregation. (2) Stage2: Cold-Start SFT on UMECS. (3) Stage3: Emotion–Cause Flow Graph Structured Reinforcement Learning.

stimuli. When multiple speakers' emotion shifts are attributed to the same utterance through this cognitive pathway, that utterance naturally emerges as a conversation-level core event (i.e., a shared cause linked to multiple emotion utterances).

Formally, the reasoning trace is structured as $\mathcal{T} = \langle \mathcal{A}_{\text{macro}}, \mathcal{A}_{\text{reup}}, \mathcal{A}_{\text{cause}} \rangle$. This paradigm defines a reasoning template that serves as the backbone of ECFlow. The full prompt template is provided in Appendix A. By externalizing the logic of emotion shifts and core events into the trace $\mathcal{T}$, the reasoning process becomes explicit and amenable to direct verification and optimization.

3.3 Overview of ECFlow

To extract emotion–cause triplets that are globally coherent across turns, ECFlow enforces emotion-shift coherence and core-event consistency by combining appraisal-structured generation with flow-graph verifiable rewards. As shown in Figure 1, ECFlow operates in three stages:

Multimodal Semantic Aggregation verbalizes audio and visual signals into textual cue descriptions and fuses them with transcripts, so that non-verbal evidence is explicitly available.

In *Cold-start SFT*, we initialize reasoning capabilities by fine-tuning on **UMECS**, a unified dataset constructed with explicit appraisal traces to transform the model into a structured cognitive reasoner.

In *Emotion–Cause Flow Graph Structured RL*, we utilize a carefully designed graph structure to capture the emotional shifts and core events within the conversation, guiding the model optimization via fine-grained structured rewards, thereby achieving globally consistent emotion–cause extraction.

3.4 Multimodal Semantic Aggregation

This stage aims to bridge the modality gap by enriching the textual conversation with explicitly verbalized audio and visual cues. While Multimodal LLMs (MLLMs) can directly ingest audio/visual inputs, we empirically find that structured, long-horizon textual reasoning is more reliable when performed by strong text-only LLMs, which is crucial for the constraint-heavy logic in MECTEC (Wang et al., 2024c; Cui et al., 2025). Moreover, prior work indicate that text often provides the strongest signal for MECTEC, while audio/visual cues typically play a complementary role with inconsistent marginal gains (Wang et al., 2024a; Cheng et al., 2024; Liang et al., 2025). Therefore, we adopt a verbalization strategy, utilizing LLMs as the core reasoner while augmenting the input with semantic descriptions of other modalities.

For each utterance U_i, an audio–language model (ALM) and a video–language model (VLM) are applied to obtain modality-grounded descriptions:

$$\hat{D}_i^a = f_a(U_i^a), \qquad \hat{D}_i^v = f_v(U_i^v), \tag{3}$$

where $\hat{D}_i^a$ describes utterance-level paralinguistic cues (e.g., intonation, loudness, speech rate) and $\hat{D}_i^v$ describes visible behaviors (e.g., facial actions, gaze, gestures) observed in the aligned clip. The prompting details are provided in Appendix B. We structure the enriched utterance $\tilde{U}_i$ to embed the paralinguistic and behavioral cues directly alongside the transcript:

$$\tilde{U}_i = [U_i^t;\ \hat{D}_i^a;\ \hat{D}_i^v]. \tag{4}$$

Then, we format the fused conversation into the appraisal-guided template (Section 3.2) to obtain the input prompt, denoted as X. This yields a unified text input with a consistent reasoning format across datasets and modalities.

3.5 Cold-start SFT on UMECS

This stage initializes the model with the ability to generate structured appraisal traces $\mathcal{T}$, providing a necessary reasoning foundation for subsequent RL optimization. Since original MECTEC datasets only provide isolated triplets, we employ a teacher LLM to synthesize $\mathcal{T}$s that operationalize cognitive appraisal and connect the conversation to the gold triplets. Concretely, UMECS is constructed in three stages: appraisal-guided trace synthesis from multimodal-enriched dialogues and gold triplets, automatic and LLM-based verification/refinement to remove unsupported or structurally invalid content, and expert human review to correct psychologically implausible or cross-turn inconsistent appraisals. We further verify and refine the traces with automatic validators, LLM-based checks, and expert human review to improve trace quality. Importantly, this human review stage is used for expert revision rather than sample filtering: reviewers edit problematic reasoning spans while keeping the gold triplets fixed and without introducing new causal relations. Detailed data verification and refinement procedures are illustrated in Appendix C.

We denote the resulting dataset as **UMECS** (**U**nified **MEC**TEC **S**FT dataset), which unifies existing benchmarks with a shared trace format and ground truth-consistent triplets. This supervision encourages controllable intermediate reasoning about shifts and causes. Given UMECS, we optimize the policy using the standard negative log-likelihood objective:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{UMECS}}} \sum_{i=1}^{|y|} \log \pi_\theta(y_i \mid x, y_{<i}) \,, \tag{5}$$

where π_θ is the optimized policy, x is the structured prompt *without* ground truth, and y contains the appraisal trace and the final boxed triplets. Policy π_θ can reliably produce appraisal-guided outputs.

3.6 ECFlow Graph Structured RL

With π_θ, we further refine the model with RLVR to optimize global conversational consistency, specifically *emotion-shift coherence* and *core-event consistency*. A reward defined only on final triplet matching is largely structure-agnostic: it ignores within-speaker temporal continuity and shared causes, and can be sparse when partial mistakes occur. Therefore, we lift both predictions and ground truth into an ECFlow Graph that jointly encodes emotion–cause links and within-speaker shift/stability transitions across turns. We then compute rewards by comparing graph structures rather than treating each predicted pair in isolation, which encourages globally coherent emotion trajectories and consistent identification of shared causal events.

3.6.1 ECFlow Graph Construction

Triplet-level matching does not explicitly model within-speaker shift/stability or shared causes. We therefore construct an ECFlow Graph that encodes emotion–cause links and within-speaker transitions, enabling structure-aware rewards beyond exact-triplet matching.

As shown in Figure 1, we place each utterance node by (speaker, time) and assign it an utterance-level emotion state: if the utterance appears as an emotion utterance in $\mathcal{P}$ we use its label, otherwise we set it to `neutral`. Given C and $\mathcal{P} \in \hat{\mathcal{P}}, \mathcal{P}^*$, we deterministically construct

$$G = (V, E_{\text{cause}}, E_{\text{shift}}, E_{\text{stable}}), \tag{6}$$

where $V = v_i{}_{i=1}^{n}$ contains one node per utterance. We define three edge types:

- **Cause edges** encode emotion–cause relations: for each $(U_j^e, U_j^c, y_j^e) \in \mathcal{P}$, we add a directed edge from the node of U_j^c to the node of U_j^e:

$$\begin{aligned} E_{\text{cause}} = \{\, & v(U_j^c) \to v(U_j^e) \mid \\ & (U_j^e, U_j^c, y_j^e) \in \mathcal{P} \,\}. \end{aligned} \tag{7}$$

- **Emotion-shift edges** capture within-speaker emotion transitions over time. For each speaker s, let $\mathcal{I}_s = (i_1, \ldots, i_{|\mathcal{I}_s|})$ be the indices of utterances spoken by s in chronological order, and let e_i denote the emotion state assigned to node v_i (i.e., the Ekman label if U_i is an emotion utterance in $\mathcal{P}$, otherwise `neutral`). For each consecutive pair (i_k, i_{k+1}) in $\mathcal{I}_s$, we add a directed edge $v_{i_k} \to v_{i_{k+1}}$ to E_{shift} if $e_{i_k} \neq e_{i_{k+1}}$, where e_i denotes the emotion category.

- **Emotional stable edges** E_{stable} connect consecutive utterances of the same speaker when the emotion state remains unchanged.

This construction yields a single graph representation that supports graph-level comparison and reward computation.

3.6.2 Reward Design

We define a composite, rule-based reward that combines flow-level structural agreement on graphs and task-level agreement on triplets. Let $\hat{G}$ and G^* denote the predicted and gold Emotion–Cause Flow Graphs, and $\hat{\mathcal{P}}$ and $\mathcal{P}^*$ the corresponding triplet sets. Flow-structured rewards R_{ECFlow} and task-oriented triplet rewards R_{Triplet} are computed.

(1) R_{ECFlow} is a weighted sum of four $F1$ scores computed on graph-derived sets which include:

i) **Emotion-Shift reward** (R_{shift}).

$$R_{\text{shift}} = F1\left(E_{\text{shift}}(\hat{G}), E_{\text{shift}}(G^*)\right). \tag{8}$$

ii) **Emotion-Stable reward** (R_{stable}).

$$R_{\text{stable}} = F1\left(E_{\text{stable}}(\hat{G}), E_{\text{stable}}(G^*)\right). \tag{9}$$

iii) **Core-event reward** (R_{core}). Some cause utterances act as *core events* that explain a large fraction of emotion–cause links in a conversation (Sun et al., 2024). Rather than thresholding local out-degree, we define core events by their *global explanatory coverage* over the cause-edge set.

Given an emotion–cause flow graph G, let $E_{\text{cause}}(G)$ denote its directed cause edges. For a candidate cause node $v_i \in V$, we define the set of cause edges it covers as

$$\mathcal{C}(v_i; G) = \{(v_i \to v_e) \in E_{\text{cause}}(G)\}. \tag{10}$$

For a set of cause nodes $H \subseteq V$, its edge coverage is

$$\mathrm{Cov}(H; G) = \frac{\left|\bigcup_{v \in H} \mathcal{C}(v; G)\right|}{|E_{\text{cause}}(G)|}. \tag{11}$$

We then define the *x-coverage core-event set* as the minimum-size set of cause nodes that covers at least an x fraction of cause edges:

$$\mathcal{H}_x(G) = \arg\min_{H \subseteq V} |H| \quad \text{s.t.} \quad \mathrm{Cov}(H; G) \geq x, \tag{12}$$

where $x \in (0, 1]$ is a fixed coverage threshold.[1]

Finally, the core-event reward compares the predicted and gold core-event sets:

$$R_{\text{core}} = F1\big(\mathcal{H}_x(\hat{G}), \mathcal{H}_x(G^*)\big). \tag{13}$$

iv) **Segment reward** (R_{segment}). We additionally compare speaker-specific non-neutral emotion segments (contiguous spans), which are robust to isolated edge errors:

$$R_{\text{segment}} = F1\left(\mathcal{S}(\hat{G}), \mathcal{S}(G^*)\right). \tag{14}$$

Overall, R_{ECFlow} is computed as a weighted combination of the above four rewards, providing dense and interpretable feedback that links local predictions to global conversational emotional dynamics, and complementing the triplet-level supervision.

(2) R_{Triplet} aggregates five $F1$ terms computed from $\hat{\mathcal{P}}$ and $\mathcal{P}^*$:

i) **Exact reward** (R_{exact}). $F1$ on exact triplet matching (U_j^e, U_j^c, y_j^e), measuring final task correctness.

ii) **Pair reward** (R_{pair}). $F1$ on utterance pairs (U_j^e, U_j^c), ignoring the emotion label y_j^e.

iii) **Emotion reward** (R_{emotion}). $F1$ on (U_j^e, y_j^e), ignoring the cause utterance U_j^c, isolating utterance-level emotion classification quality.

iv) **Emotion-detection reward** ($R_{\text{emo\_det}}$). $F1$ between predicted and gold sets of emotion utterance indices, encouraging correct detection of which turns carry emotions.

v) **Cause-detection reward** ($R_{\text{cause\_det}}$). $F1$ between predicted and gold sets of cause utterance indices, encouraging correct identification of causal turns even when exact pairing is imperfect.

R_{Triplet} is computed as a weighted sum of the five signals, and is combined with R_{ECFlow} to form the final RL reward for ECFlow; we further analyze the sensitivity to reward weights in Appendix G.3.

3.6.3 RL Optimization via GRPO

We optimize the policy with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Following the cold-start stage, we reuse the structured prompts from UMECS as RL inputs, i.e., $x \in \mathcal{D}_{\text{UMECS}}$. For each prompt x, we sample a group of G trajectories $\{y^{(i)}\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}(\cdot \mid x)$. Each trajectory $y^{(i)}$ contains an appraisal-guided reasoning trace $\mathcal{T}^{(i)}$ and the final predicted triplet set $\hat{\mathcal{P}}^{(i)}$. We score each trajectory with reward $R = w_1 R_{\text{ECFlow}} + w_2 R_{\text{Triplet}}$ with fixed weights, and obtain the corresponding group-normalized token-level advantages $\hat{A}_{i,t}$.

We then update θ by minimizing the GRPO sur-

[1] We obtain $\mathcal{H}_x(G)$ using a standard greedy set-cover approximation.

rogate objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{\substack{x\sim\mathcal{D}_{\text{UMECS}},\\ \{y^{(i)}\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\Big(\min\Big(r_{i,t}\hat{A}_{i,t}, \mathrm{clip}(r_{i,t},1-\epsilon,1+\epsilon)\hat{A}_{i,t}\Big) - \beta\, D_{\text{KL}}\Big(\pi_\theta \,\|\, \pi_{\text{ref}}\Big)\Big)\Bigg], \tag{15}$$

where T_i is the length of $y^{(i)}$ and $r_{i,t} = \frac{\pi_\theta(y_t^{(i)}|x,y_{<t}^{(i)})}{\pi_{\theta_{\text{old}}}(y_t^{(i)}|x,y_{<t}^{(i)})}$. Here π_{ref} denotes the frozen reference policy initialized by cold-start SFT on UMECS.

4 Experimental Setup

4.1 Implementation Details

We conduct experiments on two public MECTEC benchmark datasets, i.e., ECF (Wang et al., 2022) and MECAD (Liang et al., 2025), both of which provide aligned text, audio, and video modalities for multi-speaker conversations. Unless otherwise specified, we use Qwen2.5-7B-Instruct[2] as the backbone LLM for our SFT and RL training. For evaluation, following prior MECTEC work (Wang et al., 2022; Liang et al., 2025), we report the standard F1 score on utterance-level emotion–cause triplets. Specifically, we compute F1 for triplets within each of the six Ekman emotion categories separately (Anger, Disgust, Fear, Joy, Sadness, Surprise), and then report the weighted average F1 across all six categories, denoted as 6 Avg, where weights are proportional to the number of ground truth in each category. In addition, considering the strong class imbalance in MECTEC datasets (notably for Disgust and Fear), we also report the weighted average F1 over the four major categories excluding Disgust and Fear, denoted as 4 Avg, as commonly adopted in previous studies (Wang et al., 2023). Detailed parameter settings regarding cold-start SFT and RL training are provided in Appendix D.

4.2 Baselines

Due to the limited research on MECTEC, we also include representative approaches from related tasks such as Emotion Cause Pair Extraction (ECPE) and Emotion Cause Pair Extraction in Conversations (ECPEC). We group baselines into four categories: **(1) Pipeline:** MC-ECPE-2steps (Wang et al., 2022). **(2) End-to-End (E2E):** RankCP (Wei et al., 2020), SHARK (Wang et al., 2023), M^{3}HG (Liang et al., 2025). **(3) LLM-based Methods (non-RL):** GPT-5, NUS-Emo (Luo et al., 2024), UMECS-SFT. **(4) RL Methods:** GRPO w/ R_{Triplet} (Shao et al., 2024), DAPO w/ R_{Triplet} (Yu et al., 2025). Detailed descriptions of baseline models are provided in Appendix E.

5 Experimental Results

5.1 Main Results

Table 1 compares ECFlow with a range of baselines on MECTEC. ECFlow achieves the best results on both benchmarks. On ECF, ECFlow improves 6 Avg from 40.25 (NUS-Emo) to 46.95, yielding a **17%** relative gain. On MECAD, ECFlow increases 6 Avg from 38.57 to 42.45, achieving a **10%** relative gain. The gains are stable across multiple emotion categories, indicating that ECFlow improves the reliability of joint emotion and cause extraction at the conversation level.

In addition, although NUS-Emo relies on a multi-stage decomposition based on ERC and ECPE, UMECS-SFT as an end-to-end generative model achieves stronger performance on both datasets. This suggests that structuring generation with appraisal-guided reasoning traces provides effective intermediate constraints, reducing the dependence on task decomposition and other modules.

We then evaluate the Zero-RL setting to isolate the contribution of RLVR and structured optimization objectives. Without UMECS-based initialization, ECFlow still outperforms NUS-Emo on ECF, improving 6 Avg by 9%. This indicates that triplet-level matching alone is insufficient to enforce conversation-level constraints, while lifting predictions to a structured graph representation and optimizing verifiable rewards over global properties such as emotion shifts and shared causes provides a more effective learning signal.

Finally, adding audio and visual cues typically yields additional gains, but the improvements are moderate, suggesting that text remains the primary information source for MECTEC and non-textual cues mainly play a complementary role. This aligns with our design of centering reasoning on text while explicitly verbalizing audio and visual signals as readable semantic descriptions.

[2] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

Dataset		Method	Modality	Anger	Disgust	Fear	Joy	Sadness	Surprise	6 Avg.	4 Avg.
ECF	Pipeline	MC-ECPE-2steps (TAC'22)$^{\triangle}$	T, A, V	24.39	0.00	0.71	38.84	21.60	40.24	29.32	31.92
	E2E	RankCP (ACL'20)	T	28.29	12.03	3.52	38.69	22.17	37.67	30.58	32.48
		SHARK (EMNLP'23)*	T	28.65	10.42	5.33	40.41	25.35	40.45	32.24	34.33
		M^{3}HG (ACL'25)*	T, A, V	36.08	23.33	9.88	49.03	32.41	47.46	40.07	41.96
	LLM	GPT-5	T	27.09	16.82	8.81	24.55	22.60	35.39	26.32	26.32
		NUS-Emo (SemEval'24)	T, V	34.02	20.81	15.19	47.48	34.78	50.94	40.25	42.15
		UMECS-SFT	T, A, V	39.41	**35.90**	25.42	45.79	37.06	43.70	41.14	41.90
	RL	GRPO w/ R_{Triplet}	T	28.76	11.02	18.46	49.04	37.31	47.61	38.60	40.86
		DAPO w/ R_{Triplet}	T	33.67	20.55	11.43	48.34	37.57	46.92	39.91	41.91
		ECFlow (Zero-RL)	T	38.61	16.11	20.93	**52.67**	38.62	48.83	42.98	45.25
			T, A	38.46	19.05	22.45	51.82	38.87	49.80	43.14	45.20
			T, V	39.50	20.95	24.49	50.86	39.18	50.51	43.51	45.43
			T, A, V	39.71	23.81	24.49	50.86	39.81	51.26	43.99	45.78
		ECFlow (SFT + RL)	T, A, V	**44.40**	30.14	**37.04**	52.65	**41.75**	**51.99**	**46.95**	**48.20**
MECAD	Pipeline	MC-ECPE-2steps (TAC'22)	T, A, V	28.43	0.00	0.23	22.45	27.67	45.14	22.01	24.83
	E2E	RankCP (ACL'20)	T	29.79	12.50	3.06	21.79	29.31	32.36	26.29	28.32
		SHARK (EMNLP'23)	T	30.22	10.16	4.10	25.84	30.21	34.59	27.58	29.99
		M^{3}HG (ACL'25)*	T, A, V	38.34	21.89	8.79	28.10	31.17	43.29	32.82	34.59
	LLM	GPT-5	T	34.97	16.82	11.07	27.29	15.32	45.87	27.20	28.72
		NUS-Emo (SemEval'24)	T, V	44.98	32.16	5.26	37.28	34.48	46.81	38.57	40.19
		UMECS-SFT	T, A, V	44.05	**35.90**	**20.59**	34.78	35.19	48.45	39.02	39.88
	RL	GRPO w/ R_{Triplet}	T	37.80	18.80	11.76	38.74	36.24	45.36	36.28	38.76
		DAPO w/ R_{Triplet}	T	40.18	21.37	11.76	37.15	35.19	50.00	37.20	39.53
		ECFlow (ours)	T	44.28	21.70	13.95	38.79	35.11	55.59	39.45	41.99
			T, A	45.24	23.93	14.71	37.15	36.24	56.19	40.02	42.38
			T, V	44.94	27.35	11.76	39.13	35.19	57.73	40.47	42.64
			T, A, V	44.94	28.21	11.76	39.13	36.24	58.25	40.93	43.07
		ECFlow (SFT + RL)	T, A, V	**45.54**	34.19	17.65	**39.92**	**37.04**	**60.31**	**42.45**	**44.01**

Table 1: Performance comparison of different methods on the MECTEC task. $\triangle$ denotes the results are from (Wang et al., 2023). $*$ denotes the results are from the original paper (Wang et al., 2023; Liang et al., 2025). The best results and the second best results are in bold and underlined, respectively. Rows corresponding to our proposed methods are highlighted in light blue.

5.2 Structure-Aware Evaluation on Emotion Shifts and Core Events

This experiment tests whether lifting predicted and gold triplets into an ECFlow Graph and optimizing verifiable graph-structured rewards improves conversation-level consistency beyond local triplet matching. We report three structure-aware metrics computed on ECFlow Graphs induced from predicted and ground-truth triplets: Shift-F1, Core-F1, and Seg-F1, aligned with R_{shift}, R_{core}, and R_{segment}. For Core-F1, we use $x = 0.7$ and report additional x in Appendix F.

As shown in Table 2, ECFlow (SFT+RL) performs best on all three structure-aware metrics. ECFlow improves Shift-F1 by more accurately localizing within-speaker emotion change points. On Core-F1 ($x = 0.7$), ECFlow scores 66.36, outperforming NUS-Emo (63.03), indicating stronger core-event identification at the conversation level. On Seg-F1, ECFlow reaches 44.84 versus DAPO (40.53), suggesting more stable recovery of contiguous emotion segments and speaker trajectories.

Additional analyses are provided in the appendix, including Core-F1 sensitivity to the coverage threshold x in Appendix F, ablations on ECFlow reward components and the appraisal-guided reasoning format in Appendix G, and reward-weight sensitivity analysis in Appendix 2. We also report training dynamics for both Zero-RL and SFT+RL in Appendix H and qualitative case studies illustrating reasoning trace behaviors in Appendix I.

Table 2: Structure-aware evaluation on ECFlow-graph metrics. Shift-F1, Core-F1 (x=0.7), and Seg-F1 are computed on ECFlow Graphs induced from predicted and gold triplets; best results are in bold and second-best results are underlined.

Method	Shift-F1	Core-F1	Seg-F1
NUS-Emo	68.05	63.03	40.60
M^{3}HG	64.21	57.12	36.10
UMECS-SFT	68.26	59.00	39.67
GRPO w/ R_{Triplet}	69.24	61.89	40.27
DAPO w/ R_{Triplet}	70.85	62.37	40.53
ECFlow (SFT+RL)	**73.39**	**66.36**	**44.84**

6 Conclusion

In this paper, we propose ECFlow, an appraisal-guided and flow-structured reinforcement learning framework for MECTEC task. By operationalizing cognitive appraisal theory into a controllable intermediate reasoning paradigm, we enable models to explicitly track speaker-level emotion shifts and identify conversation-level core events through structured appraisal traces, supported by our constructed UMECS dataset. Furthermore, we lift both predictions and ground truth into an Emotion–Cause Flow Graph, transforming global conversational constraints, including emotion-shift coherence and core-event consistency into verifiable, structure-aware rewards optimized via GRPO. Extensive experiments on the ECF and MECAD benchmarks demonstrate that ECFlow consistently achieves state-of-the-art performance in both task-level accuracy and structure-aware evaluation, validating the effectiveness of combining appraisal-guided reasoning with graph-structured reinforcement learning for globally coherent emotion–cause extraction.

Limitations

Despite the superiority of the proposed ECFlow framework, its effectiveness is, to some extent, influenced by the long-context reasoning and structured generation capabilities of the underlying backbone model. A model that can reliably follow the appraisal template and maintain cross-turn consistency will naturally amplify the benefits of ECFlow. Conversely, models that drift in long-horizon generation may not fully exploit the potential of our framework. In addition, ECFlow relies on a multi-stage training and alignment pipeline (multimodal cue aggregation, UMECS cold-start SFT, and structured-reward RLVR), which increases computational cost and engineering complexity compared to pure SFT or prompting. Finally, while cue verbalization helps stabilize constraint-heavy, conversation-level reasoning, it introduces an intermediate representation bottleneck and may limit end-to-end modeling of fine-grained non-verbal signals; future work may explore integrating our graph-structured consistency objectives with stronger MLLMs as their long-context and structured reasoning capabilities mature.

Ethical Considerations

This research aims to improve the performance and consistency of MECTEC task, supporting more coherent and interpretable affective understanding. We adhered to standard research ethics throughout the study. All datasets, models, and tools used in our experiments are publicly available, ensuring transparency and reproducibility. No private or personally identifiable data are involved, and the proposed approach is not intended for harmful, deceptive, or privacy-invasive applications.

References

Keshav Bansal, Harsh Agarwal, Abhinav Joshi, and Ashutosh Modi. 2022. Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 44–56.

Zebang Cheng, Fuqiang Niu, Yuxiang Lin, Zhi-Qi Cheng, Xiaojiang Peng, and Bowen Zhang. 2024. Mips at semeval-2024 task 3: Multimodal emotion-cause pair extraction in conversations with multimodal language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 667–674.

Yiming Cui, Xin Yao, Yuxuan Qin, Xin Li, Shijin Wang, and Guoping Hu. 2025. Evaluating large language models on multimodal chemistry olympiad exams. *Communications Chemistry*.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020a. Ecpe-2d: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3574–3583.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Guimin Hu, Zhihong Zhu, Daniel Hershcovich, Hasti Seifi, and Jiayuan Xie. 2024. Unimeec: Towards unified multimodal emotion recognition and emotion cause. *arXiv preprint arXiv:2404.00403*.

Richard S Lazarus. 1982. Thoughts on the relations between emotion and cognition. *American psychologist*, 37(9):1019.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Qiao Liang, Ying Shen, Tiantian Chen, and Lin Zhang. 2025. M3hg: Multimodal, multi-scale, and multi-type node heterogeneous graph for emotion cause triplet extraction in conversations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11416–11431.

Meng Luo, Han Zhang, Shengqiong Wu, Bobo Li, Hong Han, and Hao Fei. 2024. Nus-emo at semeval-2024 task 3: Instruction-tuning llm for multimodal emotion-cause analysis in conversations. *arXiv preprint arXiv:2501.17261*.

Sancheng Peng, Lihong Cao, Guojun Wang, Zhouhao Ouyang, Yongmei Zhou, and Shui Yu. 2025. A survey on textual emotion cause extraction in social networks. *Digital Communications and Networks*, 11(2):524–536.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Yidan Sun, Qin Chao, and Boyang Li. 2024. Event causality is key to computational story understanding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3493–3511.

Jianhua Tao and Tieniu Tan. 2005. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2022. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024a. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2039–2050.

Fanfan Wang, Jianfei Yu, and Rui Xia. 2023. Generative emotion cause triplet extraction in conversations with commonsense knowledge. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 3952–3963.

Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024b. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024c. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3171–3181.

Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Shen Zhang, Haojie Zhang, Jing Zhang, Xudong Zhang, Yimeng Zhuang, and Jinting Wu. 2024. Samsung research china-beijing at semeval-2024 task 3: A multi-stage framework for emotion-cause pair extraction in conversations. *arXiv preprint arXiv:2404.16905*.

A A Reference Prompt for the Appraisal-Guided Reasoning Paradigm

We use a fixed checklist-style prompt to instantiate the appraisal-guided trace, the reference template is provided in Table 8.

B Implementation Details for Multimodal Semantic Aggregation

We instantiate the audio large model $f_a(\cdot)$ and the video large model $f_v(\cdot)$ with off-the-shelf foundation models to extract $\hat{D}_i^a$ and $\hat{D}_i^v$ for each utterance. Specifically, we use GPT-audio for $f_a(\cdot)$ to summarize paralinguistic cues from the aligned audio segment, and Gemini-2.5-Pro for $f_v(\cdot)$ to summarize observable behaviors from the aligned video clip. Table 10 and Table 9 provide simplified reference templates.

C Verification and Refinement for Synthesized Traces

To construct UMECS, we synthesize appraisal-guided reasoning traces using a teacher LLM conditioned on the fused multimodal conversation and the ground-truth triplets. Because synthesized rationales may contain formatting errors, invalid indices, or unsupported psychological explanations, we apply a multi-step quality control pipeline consisting of (i) rule-based validation, (ii) LLM-based verification and refinement, and (iii) expert human review.

Rule-based validation. We first run deterministic validators to ensure that each synthesized sample is strictly parsable and satisfies task constraints, including valid formatting of the `ANSWER` block, legal utterance indices and emotion labels, and exact agreement between the parsed boxed triplets and the provided ground-truth set. Samples that fail validation are discarded and re-generated with the teacher model until they pass the checks (up to a fixed retry budget).

LLM verification and refinement. For samples that pass rule-based validation, we further audit trace quality with an LLM-as-a-judge and refine problematic spans via a separate GPT-5 call.

Verification. Given the conversation (with multimodal cue descriptions), the ground truth, and the generated trace, the verifier checks whether (i) utterance-level appraisals are supported by explicit evidence in the dialogue/cues (no invented events, intentions, or triggers), (ii) speaker states and trigger explanations are internally consistent across turns, and (iii) the trace does not introduce invalid utterance indices or unsupported causal claims beyond the provided triplets. The verifier returns a structured diagnostic report with a pass/fail decision, error categories, and the utterance-level spans that should be revised. The verification prompt is shown in Table 11.

Refinement. When the verifier flags a trace as invalid, we invoke a separate refinement call to correct the problematic spans. The refiner takes the conversation (with multimodal cues), the locked ground truth, the original trace, and the verifier's structured diagnostics, and is instructed to revise only the specified spans while preserving the checklist organization. To prevent drift, the refiner is forbidden to introduce new utterance indices or unsupported events, and must copy the gold triplet set verbatim into the final boxed answer. We then re-run the rule-based validators to ensure strict parsability and constraint satisfaction after refinement. The refinement prompt is provided in Table 12.

Expert human review and manual refinement. Finally, we perform targeted expert review to improve psychological plausibility and grounding. We recruit 10 reviewers with psychology-related training and provide guidelines aligned with our appraisal checklist. They inspect the conversation, multimodal cue descriptions, and the synthesized trace, correcting ungrounded or inconsistent appraisals while keeping the final triplets unchanged. We prioritize high-risk samples (e.g., long contexts, rare emotions, dense many-to-many links, or traces that required LLM refinement). Experts either accept the trace, apply light edits, or request re-generation when major unsupported content is found.

Overall, this pipeline yields strictly valid and parsable samples with gold-consistent triplets and higher-quality appraisal traces that are better grounded in multimodal conversational evidence, providing reliable structured supervision for subsequent training phrases.

D Hyperparameter settings of the cold-start SFT and RL training of ECFlow

We summarize the hyperparameter settings used in the cold-start SFT and RL stages of ECFlow in Table 4 and Table 3, respectively.

E Detailed Description of Baseline Models in MECTEC task

We group baselines into four categories: (1) Pipeline, (2) End-to-End (E2E), (3) LLM-based Methods (non-RL), and (4) RL Methods.

- **MC-ECPE-2steps** (Wang et al., 2022) is a two-step MECTEC architecture, which first extracts emotion utterances and cause utterances separately, and then performs pairing and filtering to identify emotion-cause pairs.
- **RankCP** (Wei et al., 2020) is a graph attention network (GAT)-based approach for ECPE to extract emotion-cause pairs by ranking.

Category	Value
RL Optimization (GRPO)	
Optimizer	AdamW
Learning rate	1×10^{-6}
KL coefficient β	0
Clip range ϵ_{low}	0.2
Clip range ϵ_{high}	0.28
GRPO group size G	16
Training epochs	5
Batch and Sequence Settings	
Global batch size	64
Mini-batch size	16
Maximum prompt length	6144
Maximum response length	6144
Sampling temperature	0.7
Reward Weights: $R_{Triplet}$	
R_{exact}	1.00
R_{pair}	0.35
R_{emotion}	0.25
$R_{\text{emo\_det}}$	0.10
$R_{\text{cause\_det}}$	0.10
Reward Weights: R_{ECFlow}	
R_{shift}	0.20
R_{stable}	0.10
R_{core}	0.15
R_{segment}	0.15

Table 3: Hyperparameters for the RL stage of ECFlow.

- **SHARK** (Wang et al., 2023) is one of the state-of-the-art methods for ECPEC that incorporates commonsense into GATs to improve the model's semantic understanding of emotions and causes.
- **M^{3}HG** (Liang et al., 2025) is one of the state-of-the-art methods for MECTEC. It models the conversation structure as a multimodal heterogeneous graph, and fuses semantic information from different modalities at inter-utterance and intra-utterance granularities.
- **GPT-5** serves as a strong closed-source LLM baseline to assess the zero-shot capability of proprietary foundation models on MECTEC. We use the GPT-5.1 API and prompt it with the fused multimodal conversation together with our appraisal-guided reasoning paradigm, requiring it to perform zero-shot reasoning and output the final triplet set.
- **NUS-Emo** (Luo et al., 2024) is the strongest open-source baseline in SemEval-2024 Task 3. It decomposes the task into emotion recognition in conversation (ERC) and emotion–cause pair extraction (ECPE), selects ChatGLM (GLM et al., 2024) via pilot screening, and applies emotion–cause-aware instruction tuning with ImageBind-based (Girdhar et al., 2023) multimodal perception.

Category	Value
Data & Sequence	
Max sequence length	32768
Batching	
Micro-batch size per GPU	4
Global train batch size	32
Gradient accumulation steps	1
Optimization & Schedule	
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
LR scheduler	cosine (with linear warmup)
Warmup ratio	0.03
Training epochs	5

Table 4: Hyperparameters for the cold-start SFT stage of ECFlow.

Table 5: Core-F1 of different models under varying coverage thresholds x. We report $x \in \{0.4, 0.9\}$ here. Best results are in bold and second-best results are underlined.

Method	Core-F1 ($x = 0.4$)	Core-F1 ($x = 0.9$)
NUS-Emo	<u>45.81</u>	<u>68.15</u>
M^{3}HG	40.41	63.10
UMECS-SFT	43.18	67.56
GRPO (R_{Triplet})	45.54	67.79
DAPO (R_{Triplet})	44.20	66.13
ECFlow (SFT+RL)	**52.10**	**70.13**

- **UMECS-SFT** is our supervised baseline trained on the same UMECS dataset using standard SFT training.
- **GRPO** w/ R_{Triplet} (Shao et al., 2024): In our experiments, we apply GRPO to the same backbone and the same appraisal-guided reasoning format as ECFlow, and optimize the policy using only the task-oriented triplet reward R_{Triplet}.
- **DAPO** w/ R_{Triplet} (Yu et al., 2025) (Decoupled Clip and Dynamic Sampling Policy Optimization) is an advanced RL algorithm. In our experiments, we apply DAPO to the same backbone and the same appraisal-guided reasoning format as ECFlow, and optimize the policy using only the task-oriented triplet reward R_{Triplet}.

F Core-F1 Results with Different Coverage Thresholds x

To further examine the sensitivity of Core-F1 to the coverage threshold x and the robustness of model comparisons, we additionally report Core-F1 results under $x \in \{0.4, 0.9\}$. This experiment assesses the robustness of Core-F1 with respect to the coverage threshold x and examines whether ECFlow's advantage is consistent under varying global coverage requirements. As shown in Table 5, under the lower threshold $x = 0.4$, ECFlow achieves a Core-F1 of 52.10, outperforming all baselines, indicating its ability to identify core events even when only partial coverage of emotion–cause links is required. Under the setting $x = 0.9$, all methods obtain higher Core-F1 scores, yet ECFlow remains the best-performing model with a score of 70.13, maintaining a consistent margin over competing approaches.

Overall, the relative ranking of models remains stable across different values of x, and ECFlow consistently shows superior performance, suggesting that the gain in core-event identification is not tied to a particular threshold, but reflects a more robust ability to capture conversation-level causal structure.

G Ablation Study

G.1 Ablation Study on ECFlow Reward Components

To analyze the reward design of ECFlow and quantify the contribution of each verifiable graph-structured reward term, we conduct an ablation study on ECF and MECAD by selectively removing individual components from the ECFlow reward, including R_{shift}, R_{stable}, R_{core}, and R_{segment}, while keeping the backbone model and training setup unchanged.

As shown in Table 6, the full ECFlow (SFT+RL) achieves the best task performance on both datasets and also attains the highest Shift-F1 scores (73.39 on ECF and 69.17 on MECAD), indicating that lifting predictions to graphs and optimizing with structural rewards enables more accurate localization of emotion change points within speakers, thereby reducing cross-turn inconsistencies. Removing any reward component leads to performance degradation to varying degrees. For example, on ECF, removing R_{shift} reduces Shift-F1 from 73.39 to 71.45 and lowers 6 Avg. from 46.95 to 45.10. Removing R_{core} decreases Core-F1 from 66.36 to 63.79, accompanied by a drop in task performance, suggesting that global constraints on core triggering events support more consistent causal assignments. Removing R_{segment} also leads to declines in both task and structure metrics, highlighting the role of modeling contiguous emotion segments in mitigating fragmented predictions.

Overall, the results demonstrate that the different structural rewards are complementary, and their joint optimization yields more stable conversation-level consistency and more reliable triplet predictions than any single reward component alone.

G.2 Ablation Study on Appraisal-Guided Reasoning Paradigm

To evaluate the effect of the appraisal-guided reasoning paradigm on triplet extraction and conversation-level structural consistency, we compare UMECS-SFT with two alternative output formats under the same SFT setting: Direct Answer, which outputs only final triplets, and Free-form CoT, which allows unconstrained reasoning. We report both task metrics and structure-aware metrics, as shown in Table 7.

Overall, UMECS-SFT yields more stable and consistent gains than Direct Answer and Free-form CoT across both datasets. In particular, it achieves higher task performance while also improving Shift-F1, indicating a better recovery of emotion shift within speakers and reduced cross-turn inconsistencies. In contrast, Direct Answer attains the highest Core-F1 (e.g., 61.43 on ECF) but suffers clear drops in task metrics and Shift-F1, suggesting that outputting answers alone tends to focus on a small set of core triggers at the expense of overall triplet coverage and conversation-level consistency. Free-form CoT generally performs between the two, with improvements over Direct Answer on several metrics but still falling short of UMECS-SFT, indicating that unconstrained reasoning is less effective than AGRP in aligning predictions with the underlying emotion–cause organization of the conversation.

G.3 Reward Weight Sensitivity Analysis

Figure 2 analyzes the sensitivity of ECFlow to reward weight scaling by independently rescaling each reward term to $0.5\times$ and $2.0\times$ while keeping all others fixed. Overall, the results show that ECFlow is not sensitive to the precise choice of reward weights: performance variations remain limited, and the model consistently stays at a strong performance level across all configurations.

On the MECTEC task, most reward components introduce only minor fluctuations in 6-Avg F1 when rescaled. The most noticeable degradation occurs when reducing the weight of R_{exact}, which is directly aligned with the final triplet-matching

Table 6: Ablation on reward design of ECFlow. We report both task metrics (6 Avg., 4 Avg.) and flow-based test metrics (Shift-F1, Core-F1). Best results are in bold and second-best results are underlined.

Setting	Reward terms				ECF				MECAD			
					Task		Flow metrics		Task		Flow metrics	
	$\mathbf{R}_{\text{shift}}$	$\mathbf{R}_{\text{stable}}$	$\mathbf{R}_{\text{core}}$	$\mathbf{R}_{\text{segment}}$	6 Avg.	4 Avg.	Shift-F1	Core-F1	6 Avg.	4 Avg.	Shift-F1	Core-F1
ECFlow (SFT+RL)	✓	✓	✓	✓	**46.95**	**48.20**	**73.39**	**66.36**	**42.45**	44.01	**69.17**	62.32
GRPO w/ R_{Triplet}	×	×	×	×	38.60	40.86	69.24	61.89	36.28	38.76	66.19	58.30
w/o R_{shift}	×	✓	✓	✓	45.10	46.32	71.45	64.32	41.31	42.89	67.97	**62.78**
w/o R_{stable}	✓	×	✓	✓	45.90	46.96	72.59	65.24	42.15	**44.10**	68.74	62.54
w/o R_{core}	✓	✓	×	✓	45.53	46.72	72.98	63.79	42.69	44.24	68.08	61.10
w/o R_{segment}	✓	✓	✓	×	45.95	47.07	73.08	64.10	41.92	43.32	68.25	61.67

Table 7: Ablation on the appraisal-guided reasoning paradigm. "Direct Answer" outputs only boxed triplets; "Free CoT" uses unconstrained chain-of-thought.

Setting	ECF				MECAD			
	Task		Flow metrics		Task		Flow metrics	
	6 Avg.	4 Avg.	Shift-F1	Core-F1	6 Avg.	4 Avg.	Shift-F1	Core-F1
UMECS-SFT	**41.14**	**41.90**	**68.26**	59.00	**39.02**	**39.88**	**67.54**	60.45
SFT w/ Direct Answer	38.81	40.61	65.14	**61.43**	38.11	39.71	65.01	**61.32**
SFT w/ Free-form CoT	40.26	41.01	67.34	58.78	38.41	39.19	66.87	59.10

objective. This behavior is expected, as weakening this term relaxes the strongest supervision signal for exact triplet correctness, suggesting that a relatively larger weight is beneficial for stabilizing task-level performance.

Figure 2 (b) further shows that structural rewards mainly affect their corresponding structural metrics, but again within a narrow range. For example, rescaling R_{core} leads to modest changes in Core-F1, while R_{segment} exhibits a similar localized effect on Seg-F1. These trends indicate that structural rewards function as refinement signals that shape conversation-level consistency, rather than dominating overall optimization.

Taken together, this analysis demonstrates that ECFlow exhibits strong robustness to reward weight perturbations. The performance gains are therefore attributable to the proposed graph-based modeling and verifiable reward design, rather than reliance on a finely tuned set of reward coefficients.

H Training Dynamics of ECFlow

Figure 3 compares actor entropy and response length for ECFlow (SFT+RL/Zero-RL), GRPO w/ R_{triplet}, and DAPO w/ R_{triplet} over training steps. We observe entropy collapse in GRPO w/ R_{triplet}: its actor entropy monotonically drops from around 0.4 to near 0, indicating vanishing exploration. In contrast, ECFlow (SFT+RL/Zero-RL) and DAPO w/ R_{triplet} maintain stable entropy at roughly 0.5, suggesting sustained exploration during optimization. This behavior is consistent with the design choice in both ECFlow and DAPO w/ R_{triplet} to use a higher clipping ratio, which explicitly encourages exploration and helps prevent premature policy collapse.

For response length, GRPO w/ R_{triplet}, DAPO w/ R_{triplet}, and ECFlow (Zero-RL) remain stable at about 1,700 tokens across training. ECFlow (SFT+RL) starts with longer responses due to SFT initialization and then slightly decreases, stabilizing around 2,500 tokens. Overall, these curves indicate that ECFlow maintains stable training dynamics with sustained exploration and controlled generation length, supporting robust optimization and reliable final performance.

I Case Study

To further assess the effectiveness of our approach beyond aggregate metrics, we conduct a detailed case analysis on the ECF dataset. Specifically, we select Conversation 172 as a representative example, as it contains compact yet salient emotional triggers and causal dependencies within a short conversation, making it well-suited for examining fine-grained emotion–cause reasoning behavior (Figures 4–6). Overall, ECFlow with SFT+RL achieves the highest triplet extraction F1 on this example. In addition, its generated reasoning traces align more closely with the structured reasoning style encour-

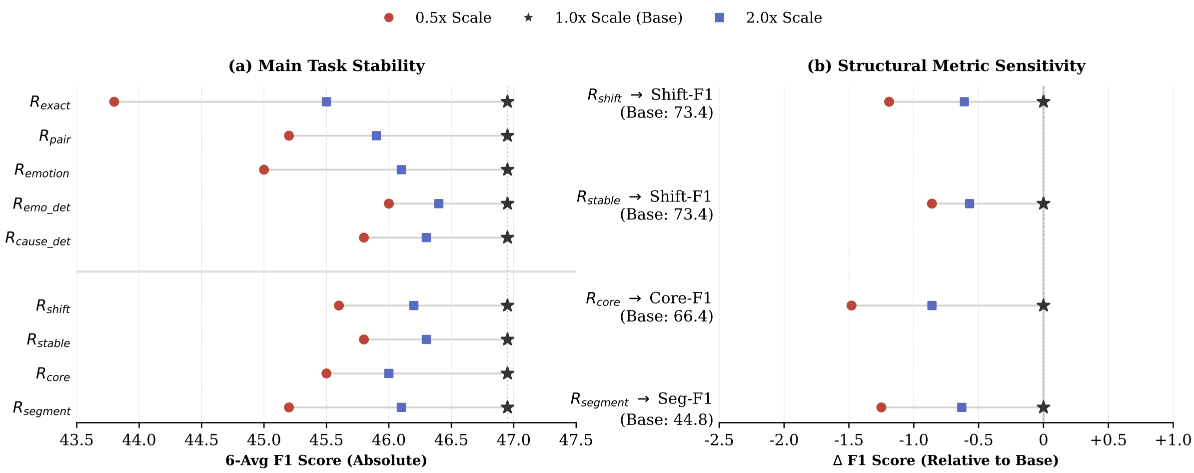

Figure 2: Sensitivity analysis of reward weight scaling in ECFlow. (a) Stability of the main task performance (6-Avg F1) when individually scaling each reward term to $0.5\times$ and $2.0\times$ while keeping all other weights fixed. (b) Relative changes (ΔF1) of flow-based structural metrics with respect to the $1.0\times$ baseline under the same weight scaling.

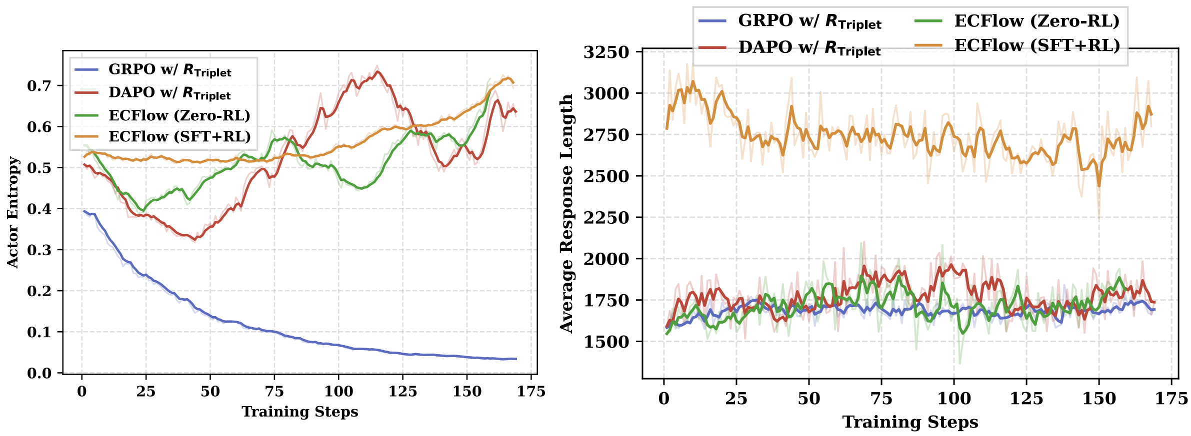

Figure 3: Training curves comparing actor entropy and response length. Actor entropy and average response length are reported over training steps for ECFlow (SFT+RL/Zero-RL), GRPO w/ $R_{triplet}$, and DAPO w/ $R_{triplet}$.

aged by UMECS, following a coherent progression from contextual appraisal to state transition and causal attribution, rather than relying on isolated or locally inconsistent predictions.

A closer inspection reveals that this advantage stems from both accurate localization of key triggers and consistent causal linking across turns. For instance, in the strong rejection scenario, the SFT+RL model correctly associates the expressed emotion with the preceding request and explicitly captures the shift from a neutral prior state to a negative reaction, yielding a stable and interpretable triplet structure. In contrast, models without reinforcement alignment are more prone to emotion label drift around the same trigger, which subsequently affects downstream cause attribution. Similarly, when handling self-evaluative utterances, SFT+RL treats the internal appraisal and emotion generation as an embedded causal process within the same turn, resulting in a reasoning pattern that is both structurally coherent and semantically grounded. These observations suggest that supervised fine-tuning provides a necessary knowledge-aware initialization, while reinforcement learning further consolidates this inductive bias into globally consistent decision-making, leading to improved extraction accuracy and more reliable reasoning behavior overall.

ECF Dataset - Conversation 172 (UMECS-SFT Reasoning Trace)

DIALOGUE INPUT

[1] Joey: You know, I think I was sixteen.
[2] Monica: Please, just a little bit off the back.
[3] Phoebe: I am still on "no."
[4] Rachel: Uh, morning. Do you guys think you could close your eyes for just a sec?
[5] Joey: No... no... no... no... no, I am not fallin for that again.
[6] Phoebe: What going on?
[7] Rachel: Well, I sorta did a stupid thing last night.
[8] Chandler: What stupid thing did you do?
[9] Paolo: Bon giorno tutti!
[10] Phoebe: Ewww!

GROUND TRUTH TRIPLETS [5, 4, "anger"], [5, 5, "anger"], [7, 7, "sadness"], [9, 9, "joy"], [10, 9, "disgust"]

UMECS-SFT REASONING PROCESS

EXPLANATION

Macro-Contextual Appraisal (Global)

Cultural Background: The interaction reflects a Western, informal, sitcom-style context (similar to "Friends"): close friends, casual teasing, and open emotional expression. Mild sarcasm and exaggerated reactions are socially acceptable and often used for humor.
Scene Context: Friends are in a shared living space, waking up and interacting. Rachel is about to do something she knows others will find silly; Paolo arrives speaking another language; Phoebe reacts strongly to something (likely a prank or unexpected behavior).
Social Dynamics: They are peers with equal power, comfortable enough to joke and react openly. Emotional expressions (surprise, disgust, mild anger) are expressed directly but usually in a playful, not hostile, way.

Sequential Reappraisal (Utterance-by-Utterance)

[1] Joey (Neutral): **Situation Perception**: recalling an age. **Previous State & Shift**: calm, **MAINTENANCE**. **Appraisal & Trigger**: neutral content; steady prosody/expression.
[2] Monica (Neutral): **Situation Perception**: asking for a small trim. **Previous State & Shift**: goal-focused, **MAINTENANCE**. **Appraisal & Trigger**: polite negotiation; low intensity.
[3] Phoebe (Neutral): **Situation Perception**: declining a suggestion. **Previous State & Shift**: controlled, **MAINTENANCE**. **Appraisal & Trigger**: boundary setting without escalation.
[4] Rachel (Neutral): **Situation Perception**: asks others to close eyes, likely playful. **Previous State & Shift**: tentative but fine, **MAINTENANCE**. **Appraisal & Trigger**: low-intensity mixed affect → neutral.
[5] Joey (Disgust): **Situation Perception**: reads [4] as renewed prank; recalls being tricked. **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: "no way" / distancing from being fooled again → disgust. **Justification**: repeated "no," emphatic pointing/head shake.
[6] Phoebe (Surprise): **Situation Perception**: odd request + strong refusal. **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: expectation mismatch → surprise. **Justification**: "What going on?"
[7] Rachel (Sadness): **Situation Perception**: admits a "stupid thing." **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: self-critique/regret → sadness. **Justification**: subdued tone + self-deprecation.
[8] Chandler (Neutral): **Situation Perception**: asks for details. **Previous State & Shift**: **MAINTENANCE**. **Appraisal & Trigger**: curiosity, not harsh.
[9] Paolo (Neutral): **Situation Perception**: greeting in Italian. **Previous State & Shift**: no A/V cues → keep neutral.
[10] Phoebe (Disgust): **Situation Perception**: "Ewww!" reaction. **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: stimulus appraised as gross/inappropriate → disgust. **Justification**: prototypical disgust facial/vocal burst.

Causal Attribution (Cause Linking)

[5] Disgust: Emotion at [5] is caused by [4] because Rachel's request is appraised as a renewed prank/scheme, which Joey strongly rejects.
[6] Surprise: Emotion at [6] is caused by [4] because the request is unusual and prompts information-seeking.
[7] Sadness: Emotion at [7] is caused by [7] because recollection + negative self-evaluation co-occur in the same utterance (embedded cause).
[10] Disgust: Emotion at [10] is caused by [9] because Paolo's greeting/presence is appraised as gross/inappropriate, triggering "Ewww!".

ANSWER (Green: Correct, Red: False Positive)

[[5,4,"disgust"], [6,4,"surprise"], [7,7,"sadness"], [10,9,"disgust"]]

Figure 4: An illustrative reasoning example of UMECS-SFT on ECF (Conversation 172).

ECF Dataset - Conversation 172 (ECFlow (Zero-RL) Reasoning Trace)

DIALOGUE INPUT

[1] Joey: You know, I think I was sixteen.
[2] Monica: Please, just a little bit off the back.
[3] Phoebe: I am still on "no."
[4] Rachel: Uh, morning. Do you guys think you could close your eyes for just a sec?
[5] Joey: No... no... no... no... no, I am not fallin for that again.
[6] Phoebe: What going on?
[7] Rachel: Well, I sorta did a stupid thing last night.
[8] Chandler: What stupid thing did you do?
[9] Paolo: Bon giorno tutti!
[10] Phoebe: Ewww!

MODEL OUTPUT TRIPLETS [5, 5, "anger"], [6, 5, "surprise"], [7, 7, "sadness"], [8, 7, "surprise"], [9, 9, "joy"], [10, 10, "disgust"]

ECFlow (Zero-RL) REASONING PROCESS

EXPLANATION

Macro-Contextual Appraisal (Global)

Cultural Background: The dialogue appears to take place in a casual setting where people are familiar with each other.
Scene Context: Joey, Monica, Phoebe, Rachel, and Chandler are engaged in a conversation.
Social Dynamics: The speakers seem to have a friendly relationship, with some level of familiarity and rapport.

Sequential Reappraisal (Utterance-by-Utterance)

[1] Joey (Neutral): **Situation Perception**: sharing a personal detail about his age. **Previous State & Shift**: no emotional shift, **MAINTENANCE**. **Appraisal & Trigger**: simply stating a fact → neutral.
[2] Monica (Neutral): **Situation Perception**: asking for a slight adjustment. **Previous State & Shift**: stable, **MAINTENANCE**. **Appraisal & Trigger**: making a request → neutral.
[3] Phoebe (Neutral): **Situation Perception**: repeating a negative response. **Previous State & Shift**: firm but stable, **MAINTENANCE**. **Appraisal & Trigger**: repeating refusal → neutral.
[4] Rachel (Neutral): **Situation Perception**: asking the group to close their eyes. **Previous State & Shift**: polite/gentle, **MAINTENANCE**. **Appraisal & Trigger**: request with no clear emotional shift → neutral.
[5] Joey (Anger): **Situation Perception**: responding to a suggestion he perceives as a trick. **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: frustration/defiance at being tricked again → anger.
[6] Phoebe (Surprise): **Situation Perception**: asking for information after Joey's response. **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: reacting to unexpected negativity and seeking clarification → surprise.
[7] Rachel (Sadness): **Situation Perception**: admitting she did something "stupid." **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: regret/shame from admitting a mistake → sadness.
[8] Chandler (Surprise): **Situation Perception**: reacting to Rachel's admission. **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: surprised/concerned reaction to the admission → surprise.
[9] Paolo (Joy): **Situation Perception**: greeting everyone. **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: friendly/cheerful greeting → joy.
[10] Phoebe (Disgust): **Situation Perception**: reacting to something unpleasant. **Previous State & Shift**: **SHIFT**. **Appraisal & Trigger**: strong negative reaction to an unpleasant stimulus → disgust.

Causal Attribution (Cause Linking)

Emotion at **[5]** is caused by **[5]** because Joey is expressing frustration and defiance.
Emotion at **[6]** is caused by **[5]** because Phoebe is reacting to Joey's negative statement.
Emotion at **[7]** is caused by **[7]** because Rachel is admitting to a mistake.
Emotion at **[8]** is caused by **[7]** because Chandler is reacting to Rachel's admission of a mistake.
Emotion at **[9]** is caused by **[9]** because Paolo is using a cheerful and friendly greeting.
Emotion at **[10]** is caused by **[10]** because Phoebe is reacting to something unpleasant.

ANSWER (Green: Correct, Red: False Positive)

[[5,5,"anger"], [6,5,"surprise"], [7,7,"sadness"], [8,7,"surprise"], [9,9,"joy"], [10,10,"disgust"]]

Figure 5: An illustrative reasoning example of ECFlow (Zero-RL) on ECF (Conversation 172)

ECF Dataset - Conversation 172 (ECFlow (SFT+RL Reasoning Trace)

DIALOGUE INPUT

[1] Joey: You know, I think I was sixteen.
[2] Monica: Please, just a little bit off the back.
[3] Phoebe: I am still on "no."
[4] Rachel: Uh, morning. Do you guys think you could close your eyes for just a sec?
[5] Joey: No... no... no... no... no, I am not fallin for that again.
[6] Phoebe: What going on?
[7] Rachel: Well, I sorta did a stupid thing last night.
[8] Chandler: What stupid thing did you do?
[9] Paolo: Bon giorno tutti!
[10] Phoebe: Ewww!

GROUND TRUTH TRIPLETS [5, 4, "anger"], [5, 5, "anger"], [7, 7, "sadness"], [9, 9, "joy"], [10, 9, "disgust"]

ECFlow (SFT+RL) REASONING PROCESS

EXPLANATION

Macro-Contextual Appraisal (Global)

Cultural Background: The interaction reflects a Western, informal, sitcom-style context (e.g., close friends, casual joking, open expression of emotions). Emotional restraint is moderate; friends can show irritation, surprise, or disgust directly without it threatening the relationship.

Scene Context: Friends are in a shared living space. Rachel is about to do something that others have done before (e.g., a prank or trick), and Phoebe reacts strongly to something Paolo says or does. The scene mixes light humor with genuine self-disclosure.

Social Dynamics: These are close friends with equal social power. They feel comfortable challenging each other, expressing irritation (e.g., Joey refusing a trick again), surprise (e.g., Phoebe's reaction to Paolo), and self-critique (Rachel admitting she did a "stupid thing").

Sequential Reappraisal (Utterance-by-Utterance)

[1] Joey (Neutral): Situation: Recalling an age-related memory. Previous State & Shift: Prev.=neutral, **MAINTENANCE**. Appraisal & Trigger: No strong appraisal; factual statement.

[2] Monica (Neutral): Situation: Requesting a small trim. Previous State & Shift: Prev.=neutral, **MAINTENANCE**. Appraisal & Trigger: Polite, task-focused request.

[3] Phoebe (Neutral): Situation: Rejecting a proposal. Previous State & Shift: Prev.=neutral, **MAINTENANCE**. Appraisal & Trigger: Undesirable proposal → boundary assertion.

[4] Rachel (Neutral): Situation: Asking others to close their eyes. Previous State & Shift: Prev.=neutral, **MAINTENANCE**. Appraisal & Trigger: Mild hesitation, controlled request (likely prank setup).

[5] Joey (Anger): Situation: Recognizes Rachel's trick. Previous State & Shift: Prev.=neutral, **SHIFT**. Appraisal & Trigger: Unfair/annoying to be fooled again → irritation + firm refusal. Justification: Repeated "no", loudness, head shake, pointing.

[6] Phoebe (Surprise): Situation: Notices odd request + strong refusal. Previous State & Shift: Prev.=neutral, **SHIFT**. Appraisal & Trigger: Expectation-reality mismatch → surprise/curiosity. Justification: Raised brows, question "What going on?".

[7] Rachel (Sadness): Situation: Admits she did a "stupid thing". Previous State & Shift: Prev.=neutral, **SHIFT**. Appraisal & Trigger: Self-critique / violation of standards → self-directed negative affect. Justification: Soft/breathy voice, downward pitch, closed eyes.

[8] Chandler (Neutral): Situation: Asks for details. Previous State & Shift: Prev.=neutral, **MAINTENANCE**. Appraisal & Trigger: Inquisitive, steady tone.

[9] Paolo (Joy): Situation: Greets the group. Previous State & Shift: Prev.=neutral, **SHIFT**. Appraisal & Trigger: Friendly greeting → positive social engagement.

[10] Phoebe (Disgust): Situation: Reacts "Ewww!". Previous State & Shift: Prev.=neutral, **SHIFT**. Appraisal & Trigger: Stimulus is physically/socially off-putting → avoidance. Justification: Scrunched face, abrupt push-away gesture.

Causal Attribution (Cause Linking)

Emotion at **[5]** is caused by **[4]** because Rachel's request is appraised by Joey as a trick setup, triggering anger; and also linked to **[5]** (self-appraisal: "not fallin for that again").

Emotion at **[6]** is caused by **[4]** (unexpected request) and **[5]** (strong refusal), creating an expectation-reality mismatch.

Emotion at **[7]** is caused by **[7]** because the admission + negative self-evaluation co-occur in the same utterance.

Emotion at **[9]** is caused by **[9]** because greeting expresses positive affect.

Emotion at **[10]** is caused by **[9]** because Paolo's greeting/presence is appraised as unpleasant, eliciting disgust.

ANSWER (Green: Correct, Red: False Positive)

[[5,4,"anger"], [5,5,"anger"], [6,4,"surprise"], [6,5,"surprise"], [7,7,"sadness"], [9,9,"joy"], [10,9,"disgust"]]

Figure 6: An illustrative reasoning example of ECFlow (SFT+RL) on ECF (Conversation 172)

A Reference Prompt for the Appraisal-Guided Reasoning Paradigm.

````
You are an expert in emotion-cause reasoning and cognitive psychology, specializing in Emotion-Cause
Triplet Extraction in Conversations (ECTEC).
Your goal is to produce a reasoning-based analysis and extract emotion-cause triplets from a
multi-speaker conversation. You must ensure that emotion labels and utterance indices are
strictly valid.

## TASK DESCRIPTION
Given a conversation, identify:
- which utterances express emotions,
- which utterances serve as their causes,
- and the corresponding emotion category.
You must reason psychologically about how emotions arise from contextual events, others' behaviors,
and internal appraisals.

## ALLOWED EMOTION LABELS
You are strictly limited to the following seven emotion labels:
`surprise`, `joy`, `sadness`, `anger`, `disgust`, `fear`, `neutral`.

## DEFINITIONS & RULES
1. emotion_utterance_id: The ID of the utterance expressing an emotion.
2. cause_utterance_id: The ID of the utterance that expresses the cause.
   - The cause usually appears before the emotion.
   - If it appears after, you must justify it.
3. One emotion can have multiple causes; one cause can trigger multiple emotions.
4. Final Constraints:
   - Include only non-neutral emotions in the final triplets.
   - Never use an invalid or nonexistent utterance index.

## OUTPUT FORMAT (Integrated Markdown Reasoning + Extraction)

```markdown
### EXPLANATION
- Macro-Contextual Appraisal (Global)
  - Cultural Background: <Describe general cultural or social norms.>
  - Scene Context: <Summarize where/why the conversation happens.>
  - Social Dynamics: <Explain power dynamics, tension, or alignment.>

- Sequential Reappraisal (Utterance-by-Utterance)
  For each utterance (starting from [1] and matching dialogue order):
  - [1] <speaker>: "<text>"
    - Situation Perception: <What event is described or reacted to?>
    - Previous State & Shift: <Prev. emotion. Does this turn mark a SHIFT or MAINTENANCE?>
    - Appraisal & Trigger: <How does the event + internal state lead to the new emotion?>
    - Emotion Label: `<one of: surprise, joy, sadness, anger, disgust, fear, neutral>`
    - Justification: <Why this label fits.>
  - [2] ...

- Causal Attribution (Cause Linking)
  - Based on the appraisal above, match non-neutral emotions to their causes.
  - Format: "Emotion at [i] is caused by [j] because..."

### ANSWER
\ boxed{
[
  [emotion_utterance_id, cause_utterance_id, "emotion_label"],
  ...
]
}
````

Table 8: A Reference Prompt for the Appraisal-Guided Reasoning Paradigm.
````

A Reference Prompt for Extracting Audio Cues.

```
You are an audio understanding assistant for Multimodal Emotion--Cause Triplet Extraction in
Conversations (MECTEC).

Goal
Given (1) the dialogue history, (2) the current utterance, and (3) the aligned audio segment, write
a short description of audible paralinguistic + interaction evidence that supports inferring the
speaker's emotion-relevant cues, possible trigger evidence (as cues), and any emotion shift vs.
just before.

Constraints
- Audio-only (acoustic cues only): describe only what is clearly audible. Do NOT mention or infer
any visual information (e.g., facial expressions, gaze, gestures, scene objects) or narrative events
not present in the audio.
- No quoting speech content: do not quote or paraphrase the spoken words. Treat the transcript as
  separate; here you describe only how it is said, not what is said.
- Evidence, not interpretation: avoid mental-state/intent/narrative language (e.g., "thinking",
  "lying", "mocking", "sarcastic", "dismissive", "triggered by", "trying to"). Describe only
  acoustic phenomena and turn-taking patterns.
- Cue-first, minimal labels: prefer cue-only descriptions without emotion terms. If you use an
  emotion term, keep it minimal and back it with concrete acoustic cues (pitch/energy/rate/pauses)
  in the same sentence.
- No trigger conclusions: do not state that a trigger is present or absent. If no clear trigger cue
is audible, omit trigger language and describe only the acoustic interaction context (e.g., overlap,
  interruption, silence).
- No naming: refer to people as the speaker, the listener, other speaker(s). Do not guess identities.
- Group speaker handling: if the current utterance speaker is a group (e.g., "All") or not clearly
  separable in the audio, describe group-level audible reactions without assigning a single person.
- Unclear audio: if the signal is noisy/overlapped/unclear, explicitly say it is not clearly audible.
- Primary subject lock: the analysis subject is always the speaker of the current utterance; describe
  others only as context (e.g., overlap, background speech), and never as the emotional subject.

Focus (in order)
- Prosody: pitch level/range, pitch movement, loudness/energy, speech rate, rhythm, emphasis.
- Voice quality: breathiness, creakiness, trembling, strain, laughter/crying/sobbing, sighs, sniffles.
- Disfluencies and pauses: hesitations, fillers, long silences, abrupt cut-offs.
- Interaction cues: interruptions, overlap, turn-taking changes, response latency.
- Before -> after changes within the segment (or explicitly say no clear shift).

Output
ONE paragraph, 2--5 sentences, each sentence includes >=1 concrete acoustic cue.

Inputs
Dialogue history: {DIALOGUE_HISTORY}
Current utterance: {CURRENT_UTTERANCE}

Now analyze the audio segment and write the description.
```

Table 9: A Reference Prompt for extracting audio cues.

A Reference Prompt for Extracting Video Cues.

```
You are a video understanding assistant for Multimodal Emotion--Cause Triplet Extraction in
Conversations (MECTEC).

Goal
Given (1) the dialogue history, (2) the current utterance, and (3) the aligned video clip, write
a short description of visually observable nonverbal + interaction evidence that supports
inferring the speaker's emotion-relevant cues, possible trigger evidence (as cues), and any
emotion shift compared to just before.

Constraints
- Video-only (visual cues only): describe only what is clearly visible on screen. Do NOT mention
  or infer any audio-related information (e.g., tone, loudness, shouting, sighing, laughter,
  speech rate, emphasis).
- No quoting speech content: do not quote or paraphrase spoken words. Treat the transcript as
  separate; here you describe only what is visually observable, not what is said.
- Evidence, not interpretation: avoid mental-state, intent, or narrative language (e.g.,
  "thinking", "remembering", "mocking", "sarcastic", "dismissive", "triggered by", "trying to").
  Describe only visible actions, reactions, and interaction patterns.
- Cue-first, minimal labels: prefer cue-only descriptions without emotion terms. If an emotion
  term is used, keep it minimal and support it with concrete visual cues (e.g., facial action,
  gaze change, posture shift, gesture) in the same sentence.
- No trigger conclusions: do not state that a trigger is present or absent. If no clear trigger
  cue is visible, omit trigger language and describe only the visible interaction context
  (e.g., who looks at whom, who reacts, timing of reactions).
- No naming: refer to people as the speaker, the listener, other speaker(s), person on the left/
  right/background. Do not guess identities.
- Group speaker handling: if the current utterance speaker is a group (e.g., "All") or not
  clearly identifiable in the frame, describe group-level visible reactions without assigning
  a single individual as the speaker.
- Unclear visibility: if the video is blurred, occluded, off-screen, or the relevant cue is not
  clearly visible, explicitly state that it is not clearly observable.
- Primary subject lock: the analysis subject is always the speaker of the current utterance;
  describe others only as visual context or potential interaction targets, and never as the
  emotional subject.

Focus (in order)
- Facial actions: eyebrow movement, eye openness, gaze direction, blinking, mouth shape.
- Head and body movement: nodding, shaking, leaning, posture changes, tension or relaxation.
- Gestures and interaction cues: hand movements, self-touching, pointing, distancing,
  approach/avoidance, mirroring.
- Turn-taking and interaction timing visible in the clip (e.g., reaction onset, mutual gaze,
  interruption cues).
- Before -> after changes within the clip (or explicitly say no clear shift is observed).

Output
ONE paragraph, 2--5 sentences, each sentence includes >=1 concrete visual cue.

Inputs
Dialogue history: {DIALOGUE_HISTORY}
Current utterance: {CURRENT_UTTERANCE}

Now analyze the video clip and write the description.
```

Table 10: A Reference Prompt for extracting video cues.

A Reference prompt for LLM Verification of Synthesized Appraisal-guided Traces.

```
You are a strict verifier for synthesized appraisal-guided reasoning traces in Multimodal Emotion-
Cause Triplet Extraction in Conversations (MECTEC).

Your job is to AUDIT (not rewrite) a generated reasoning trace for:
(1) Grounding: every claim about triggers, speaker states, or events must be supported by the
provided dialogue text and multimodal cue descriptions. Do NOT allow invented facts, hidden
intentions, or unstated events.
(2) Cross-turn coherence: appraisals should be internally consistent across turns for each speaker
(no contradictory emotions/goals without evidence of change).
(3) Constraint compliance: utterance indices must be valid; emotion labels must be within the allowed
set; the final boxed triplets must EXACTLY match the provided ground truth (order can differ).

ALLOWED EMOTION LABELS:
surprise, joy, sadness, anger, disgust, fear, neutral

INPUTS
[Dialogue with multimodal cues]
{DIALOGUE_WITH_CUES}

[Locked ground truth]  (order may vary)
{GOLD_TRIPLETS}

[Generated Trace to Verify]
{GENERATED_TRACE}

TASK
1) Check whether the boxed triplets in the trace exactly match ground truth (ignoring order). If not,
mark as FAIL and report mismatch.
2) Check for invalid utterance indices anywhere in the trace (including in explanations and cause
linking).
3) Check for ungrounded or hallucinated reasoning:
   - Claims about events not present in the dialogue/cues
   - Claims of intent/mental state not supported by explicit evidence
   - Trigger explanations that contradict the dialogue/cues
4) Check cross-turn coherence within each speaker:
   - Sudden emotion changes without any supporting cue
   - Inconsistent goals/appraisals across adjacent turns without evidence
5) Localize problems: list the utterance IDs whose analysis should be revised, and briefly state why.

OUTPUT (JSON ONLY)
Return exactly one JSON object with the following fields:
{
  "verdict": "PASS" | "FAIL",
  "summary": "<1-3 sentences summarizing the main issues or confirming validity>",
  "checks": {
    "triplets_match_gold": true | false,
    "invalid_indices_found": true | false,
    "label_violations_found": true | false,
    "ungrounded_claims_found": true | false,
    "cross_turn_inconsistency_found": true | false
  },
  "errors": [
    {
      "type": "<one of: TRIPLET_MISMATCH | INVALID_INDEX | INVALID_LABEL | UNGROUNDED_CLAIM |
      COHERENCE_CONTRADICTION | OTHER>",
      "utterance_ids": [<list of integer ids, can be empty if global>],
      "evidence": "<quote or point to the relevant dialogue/cue snippet briefly>",
      "reason": "<why it violates grounding/coherence/constraints>",
      "fix_instruction": "<how the repair model should revise this part without changing the locked
      triplets>"
    }
  ],
  "spans_to_fix": [<sorted unique utterance ids that require revision>]
}
```

Table 11: A Reference prompt for LLM verification of synthesized appraisal-guided traces.

A Reference prompt for LLM-based refinement of synthesized appraisal-guided traces.

```
You are a trace refiner for synthesized appraisal-guided reasoning in Multimodal Emotion-Cause
Triplet Extraction in Conversations (MECTEC).

Your goal is to REPAIR a previously generated reasoning trace using the verifier's diagnostic report.
You must fix only the problematic spans while keeping the final boxed triplets LOCKED to the provided
gold set.

CRITICAL CONSTRAINTS (MUST FOLLOW)
1) DO NOT change the ground truth. In the final ANSWER, copy the ground truth verbatim.
2) DO NOT introduce any new utterance indices. Use only indices that already exist in the dialogue.
3) DO NOT invent events, intentions, or triggers not supported by the dialogue text or multimodal cues.
4) Preserve the overall checklist structure and headings of the original trace.
5) Edit ONLY the spans listed in "spans_to_fix" (and, if absolutely necessary, the minimal surrounding
sentences to maintain coherence). All other parts should remain unchanged.

ALLOWED EMOTION LABELS:
surprise, joy, sadness, anger, disgust, fear, neutral
Note: the final boxed triplets must contain ONLY non-neutral emotions, consistent with ground truth.

INPUTS
[Dialogue with multimodal cues]
{DIALOGUE_WITH_CUES}

[Locked ground truth]
{GOLD_TRIPLETS}

[Original Generated Trace]
{GENERATED_TRACE}

[Verifier Diagnostic Report (JSON)]
{VERIFIER_REPORT_JSON}

INSTRUCTIONS
Step 1: Read the verifier report. Identify the utterance-level spans to fix from "spans_to_fix"
and the corresponding issues from "errors".
Step 2: In the trace, revise ONLY those spans to:
  - remove or rewrite ungrounded claims,
  - resolve cross-turn contradictions,
  - ensure every trigger/appraisal statement is backed by explicit dialogue/cue evidence,
  - keep speaker attribution correct.
Step 3: Keep the fixed checklist headings and formatting intact.
Step 4: Output the FULL corrected trace in the same Markdown format as the original, and end with:

### ANSWER
\boxed{
<GOLD_TRIPLETS copied verbatim here>
}

OUTPUT REQUIREMENTS
- Output Markdown only (no JSON).
- Keep the content concise and evidence-grounded.
- Do not add new sections or change section titles.
- Do not mention the verifier report explicitly in the final text.
- Ensure the final output is parsable and indices/labels are valid.
```

Table 12: A Reference prompt for LLM-based refinement of synthesized appraisal-guided traces.