CAIR: Causal Adaptive Information-based Reinforcement Learning for Multimodal Emotion Reasoning

Fengyu Zhang[1,2], Bin Liu[1,2] *, Jianhua Tao[3,4], Zhuofan Wen[2,1], Shun Chen[1,2], Hailiang Yao[2], Zhengqi Wen[3,4]

[1]Institute of Automation, Chinese Academy of Sciences,
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences,
[3]The Department of Automation, Tsinghua University,
[4]BNRIST, Tsinghua University

zhangfengyu2024@ia.ac.cn liubin@nlpr.ia.ac.cn

Abstract

Multimodal emotion reasoning requires both accurate identification and logical rationales to explain emotional triggers. However, current methods often suffer from *causal degeneracy*, where models produce linguistically fluent but superficial explanations that lack authentic logical derivation. To resolve this, we propose CAIR (Causal Adaptive Information-based Reinforcement Learning), a reinforcement learning framework that treats rationales as causal mediators between raw perceptual signals and emotional semantics. Our core contribution is the Causal Mediation Reward (CMR), which quantifies a rationale's interventional utility by measuring its marginal contribution to resolving predictive uncertainty. Additionally, we introduce an adaptive optimization mechanism based on the information bottleneck to balance perception and reasoning across varying cognitive loads. CAIR achieves state-of-the-art performance on MTMEUR with 73.80% accuracy and competitive results on the SCEA subset of EmoBench-M (68.5%), outperforming specialized SFT baselines by up to 14.4% while enhancing rationale faithfulness. Our findings underscore that principled reward design, rather than mere model scaling, is essential for building systems with authentic, human-like emotional understanding.

1 Introduction

Comprehending human emotion remains a primary goal for advanced AI. In complex multimodal settings like video dialogues, emotions are not isolated signals but outcomes of sophisticated causal logic. This requires models to move beyond mere recognition (*the what*) toward explaining underlying triggers (*the why*), raising a critical question: *can AI discern authentic causal threads or does it merely mirror surface correlations?*

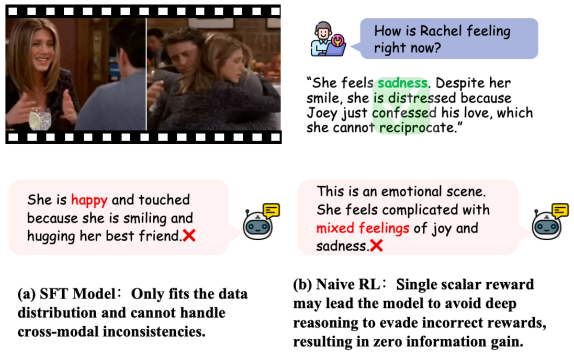

Figure 1: (a) Causal Degeneracy in SFT: Models rely on surface-level patterns and fail to resolve cross-modal contradictions, such as misinterpreting Rachel's defensive smiles by ignoring latent contextual tension. (b) Reasoning Stagnation in Naive RL: Scalar or likelihood-based rewards discourage deep exploration, resulting in "safe" yet vacuous rationales that provide no information gain for emotional decision-making.

While humans naturally navigate a cognitive trajectory from *event perception* to *mental state inference* and finally to *emotion attribution* (Camras, 1980), current computational paradigms struggle to bridge this gap. Existing approaches, predominantly centered on Supervised Fine-Tuning (SFT) (Yang et al., 2024; Zhang et al., 2023; Lian et al., 2024), treat multimodal reasoning as a distribution alignment task. By minimizing cross-entropy against reference explanations, these models learn to produce linguistically fluent text that frequently lacks logical faithfulness to the decision-making process. As illustrated in Figure 1(a), SFT models frequently default to such superficial mimicry, where a rationale might accurately describe a facial expression but fails to anchor the true emotional intent, leading to a disconnect between the explanation and the label. Recent explorations into reinforcement learning (RL) have attempted to move beyond SFT (Zhao et al., 2025; Wang et al., 2025b; Li et al., 2025), yet they often

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 25249–25264
July 2-7, 2026

optimize for final label accuracy or sequence likelihood in isolation. We observe a persistent causal degeneracy across these methods. Models receive rewards as long as their text satisfies superficial statistical patterns. However, these rationales often fail to genuinely facilitate the transition from raw perception to emotional insight.

To resolve this discrepancy, we propose **CAIR** (Causal Adaptive Information-based Reinforcement Learning), a framework that reconceptualizes the reasoning chain as a causal mediator within the decision process. Our core philosophy shifts the focus from *how an explanation looks* to *how much it contributes*. CAIR treats the rationale as a semantic intervention and quantifies its quality by how much it reduces predictive entropy. A rationale earns high reward only if it acts as critical evidence. It must link perception to the underlying emotional cause. This measurable interventional utility in uncertainty provides a direct signal to guide reinforcement learning.

However, we observe that the impact of this interventional guidance is not monolithic across the emotional landscape. While straightforward expressions allow for direct perception, complex or ambivalent emotions present a severe *information bottleneck* that demands high-order reasoning. Recognizing this heterogeneity, we introduce a difficulty-aware dynamic weighting mechanism. This approach does not merely implement curriculum learning; it adaptively reallocates credit based on the *reasoning signal-to-noise ratio* of each sample. By intensifying the guidance on cognitively demanding instances while maintaining basic alignment on simple ones, we ensure a robust optimization trajectory that honors the intrinsic complexity of human affect.

CAIR achieves competitive results on MTMEUR (73.80%) and EmoBench-M's SCEA subset (68.5%), while improving causal faithfulness and logical coherence of rationales. By rewarding interventional utility, CAIR shifts models from superficial pattern matching to logically coherent reasoning. This provides a definitive answer to our opening question: when rewards are grounded in causal mediation, models can finally learn not just the appearance of emotion, but the profound logic of *why people feel the way they do*.

Our contributions are summarized as follows: **(1) CAIR framework**—a novel RL approach that treats rationales as causal mediators between perception and emotion; **(2) Causal Mediation Reward (CMR)**—quantifying rationale utility via predictive uncertainty reduction; **(3) Cognitive-load Adaptive Optimization (CLO)**—dynamically balancing reasoning depth across task complexities; and **(4) comprehensive evaluation**—showing state-of-the-art results on MTMEUR and competitive performance on the EmoBench-M benchmark.

2 Related Work

Multimodal Emotion Analysis. Multimodal emotion analysis began as a supervised classification task, with benchmarks like CMU-MOSEI (Zadeh et al., 2018) and MULT (Tsai et al., 2019) providing aligned video-audio-text data and models using LSTMs, TFNs, or cross-modal attention to predict discrete labels. While effective for overt emotions, these approaches lack explanatory capability, and struggle with implicit expressions. The rise of large language models shifted focus toward explainable reasoning. New benchmarks such as MTMEUR (Hu et al., 2025b) require natural language explanations, spurring SFT-based methods like Emollm (Yang et al., 2024) and AffectGPT (Lian et al., 2024), which leverages LLMs for multi-modal empathetic understanding. Recent efforts like the ECR-Chain (Huang et al., 2024b) formalize structured reasoning, establishing causal chains as a new paradigm, yet they rely on handcrafted templates or strong supervision and do not address how to intrinsically learn which reasoning steps possess genuine causal mediation utility for the final prediction.

Limitations of Supervised Fine-Tuning. The core flaw of SFT lies in its misaligned objective: cross-entropy loss rewards outputs that "look like" references, regardless of whether the explanation logically justifies the answer (Zhang et al., 2025; Sun and Zhou, 2025). For instance, given a film clip labeled "angry because misunderstood," a model outputting "he frowns, so angry" may still incur low loss due to token-level similarity, despite missing the true cause (Yang et al., 2025a). This superficial matching fails catastrophically on unseen patterns like sarcasm or cultural metaphors (Fu et al., 2025). Alternative strategies include structured prompting (e.g., ECR-Chain (Huang et al., 2024b)) and reinforcement learning (RL). However, early RL attempts are inadequate: Cobbe et al. (Cobbe et al., 2021) used the confidence score from a verifier as the reward signal, optimizing only for the correctness of the final answer without

imposing any constraints on the logical coherence or completeness of the reasoning process, leading models to produce "answer-oriented" outputs that lack valid explanations. Critically, scaling up model size does not resolve this—without a proper learning signal, even 32B-parameter LLMs prioritize stylistic mimicry over causal reasoning, underscoring the need for an interventional reward that distinguishes evidentiary gain from vacuous surface description.

Reinforcement Learning for Emotion Reasoning. Recent RL approaches introduce more sophisticated designs (Yang et al., 2025b; Wang et al., 2025a; Keerthana and Gupta, 2025). RLVER (Wang et al., 2025b) employs a user simulator to optimize long-term emotional trajectories, while R1-Omni (Zhao et al., 2025) applies verifiable-reward RL to multimodal LLMs, rewarding correct answers and compliant reasoning formats. These show that task-aligned rewards improve performance. Closer to our goal, EMO-RL (Li et al., 2025) proposes emotion-similarity-weighted rewards and explicit reasoning templates for speech emotion recognition. Yet all share critical limitations: their rewards target the final output (label or full text) rather than directly evaluating the marginal contribution of a reasoning chain to the model's belief revision ; they apply uniform optimization strength, failing to account for the information bottleneck across varying sample complexities. In contrast, our work introduces a reward based on the interventional utility a rationale provides—measured by the predictive uncertainty reduction once the rationale is treated as a semantic intervention —and couples it with a cognitive-load adaptive mechanism, enabling end-to-end learning of context-sensitive explanations without external simulators or rigid templates.

3 Methodology

3.1 Problem Definition

The essence of multimodal emotion reasoning lies not merely in label identification but in the construction of a logically faithful pathway that bridges raw perceptual signals and high-level emotional semantics. Drawing inspiration from causal inference, we reconceptualize the generated rationale z as a causal mediator within the decision-making process $X \to z \to Y$. Here, $X = \{V, Q\}$ represents the multimodal context comprising video and textual queries, while $Y \in \mathcal{Y}$ denotes the discrete emotion categories.

Our objective is to optimize a policy $\pi_\theta(O|X)$ where the output O follows the structured format

$$[\texttt{<reason>}]z[\texttt{</reason>}][\texttt{<answer>}]y[\texttt{</answer>}] \quad (1)$$

Unlike standard Supervised Fine-Tuning (SFT) that minimizes the discrepancy between output distributions, our goal is to ensure that z possesses causal faithfulness. That is, the rationale must function as an evidentiary bridge: it should not only be linguistically fluent but also serve as a semantic intervention that significantly reduces the model's predictive uncertainty regarding the true emotion y^*, where y^* represents the ground-truth category in $\mathcal{Y}$. Such a formulation begs the question: *How can we quantitatively distinguish a rationale that genuinely explains the 'why' from one that merely describes the 'what'?* This necessity drives our design of an intrinsic, interventional reward system.

3.2 Dual-Path Intrinsic Verifier

To implement this interventional logic, we transition away from heuristic "if-else" reward checks toward a Dual-Path Intrinsic Verifier architecture. Specifically, this framework allows the model to evaluate its own reasoning utility without relying on external annotations.

As illustrated in Figure 2, the verifier processes each sample through two parallel reasoning streams:**Observation Stream (Path A):** The model M directly predicts the emotion based on X alone, yielding the baseline belief distribution $P_{obs} = P(Y|X)$. This represents the model's "intuitive" judgment driven by surface-level multimodal correlations. **Intervention Stream (Path B):** We treat the generated rationale z as a semantic intervention $\text{do}(Z = z)$ (i.e., injecting the generated rationale into the model's context to break the natural dependence of the intuitive path). Formally, this intervention removes the causal influence of X on z in the model's generative process, isolating the inferential utility of the rationale. The model M then revises its judgment by conditioning on both the context and the rationale, resulting in the interventional posterior $P_{int} = P(Y|X, \text{do}(z))$. By measuring the *divergence shift* between these two paths, we can isolate the unique contribution of the reasoning chain. This architecture essentially transforms the model into a self-consistent verifier, where the rationale is validated by its capacity to drive Belief Revision.

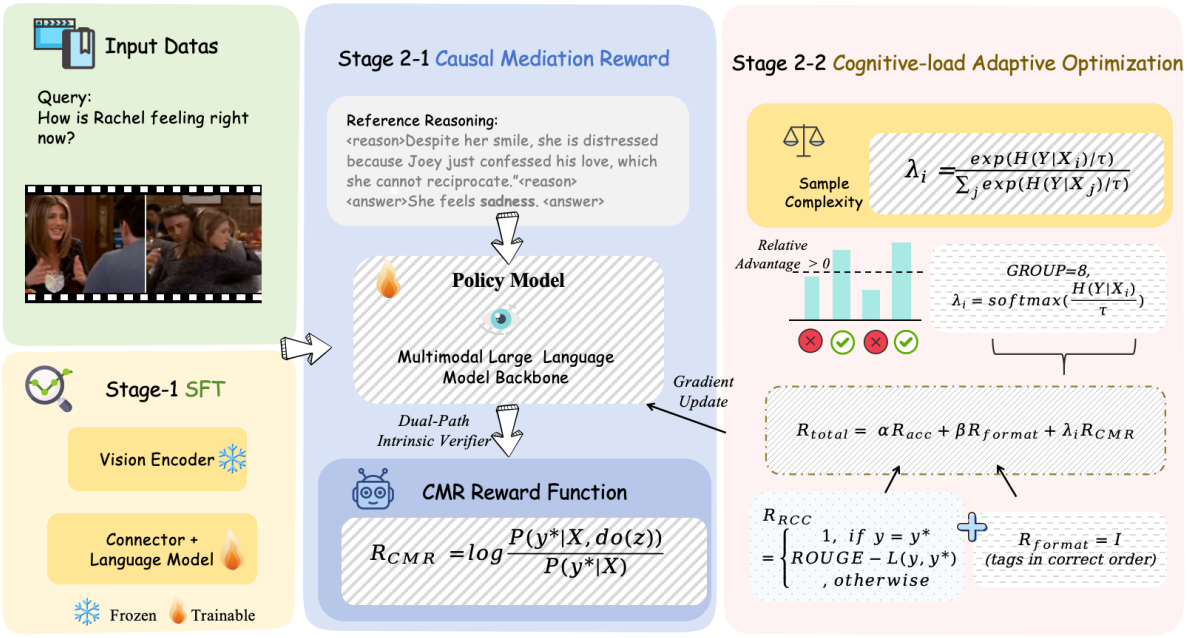

Figure 2: Overview of the CAIR framework. Stage 1 is the supervised fine-tuning (SFT) process. Stage 2-1 implements a dual-path intrinsic validator that evaluates the interventional utility of reasoning by comparing the observational stream with the interventional stream. Stage 2-2 jointly optimizes answer confidence and reasoning reliability through Cognitive Load-aware Optimization (CLO).

3.3 Causal Mediation Reward

To ensure basic task performance, we incorporate two auxiliary rewards: (1) Accuracy Reward (R_{acc}), which provides a dense signal by combining ground-truth matching with ROUGE-L similarity for incorrect labels:

$$R_{acc} = \begin{cases} 1, & \text{if } y = y^* \\ \text{ROUGE-L}(y, y^*), & \text{otherwise} \end{cases} \quad (2)$$

and (2) Format Compliance Reward (R_{format}), a binary indicator $I(\cdot)$ ensuring the output follows the required structured tags:

$$R_{format} = I(\text{tags in correct order and complete}) \quad (3)$$

Building upon the dual-path architecture, we define the Causal Mediation Reward (CMR) as the core signal for policy gradient updates. We quantify the Interventional Utility of a rationale in the log-likelihood space, a formulation that aligns with the fundamental principles of information gain and statistical physics:

$$R_{CMR} = \log \frac{P(y^*|X, \text{do}(z))}{P(y^*|X)} \quad (4)$$

This log-likelihood ratio, inspired by the principle of individual causal effect, functions as a rigorous filter for reasoning quality. If z is a redundant description (e.g., "the man is shouting"), the interventional posterior P_{int} remains close to P_{obs}, resulting in a near-zero reward. Conversely, if z identifies a latent causal mechanism (e.g., "he feels betrayed by the betrayal of trust"), it provides a substantial marginalized information boost, thereby significantly increasing the confidence in the correct label y^*. By optimizing for this log-ratio, we effectively distill the Latent Logic of the model, forcing it to generate rationales that are not just descriptive, but cognitively indispensable.

Parallel to the CMR, we maintain auxiliary rewards for *Accuracy* (R_{acc}) and *Format Compliance* (R_{format}) to ensure basic task performance, resulting in a composite reward

$$R_{total} = \alpha R_{acc} + \beta R_{format} + \lambda_i R_{CMR} \quad (5)$$

Here, α and β are static hyperparameters. However, a uniform application of this reward across all samples would ignore the intrinsic heterogeneity of human emotions.

3.4 Cognitive-load Adaptive Optimization

A key observation in our study is that the necessity of causal reasoning is not constant; rather, it is tied to the Information Bottleneck of the task. For straightforward samples with overt emotional cues,

the observation stream already provides high confidence, leaving little room for reasoning gain. In contrast, ambiguous or implicit scenarios present a high Predictive Entropy, where the "intuitive" judgment is muddled.

To honor this complexity, we introduce the **Cognitive-load Adaptive Optimization (CLO)** mechanism. Instead of heuristic difficulty scaling, we dynamically allocate reward weights based on the model's predictive uncertainty, which we conceptually term as the Reasoning Signal-to-Noise Ratio (R-SNR) to quantify the task complexity of each sample. We define the adaptive weight λ_i for each sample as:

$$\lambda_i = \text{Softmax}\left(\frac{H(Y \mid X_i)}{\tau}\right) = \frac{\exp\big(H(Y \mid X_i)/\tau\big)}{\sum_j \exp\big(H(Y \mid X_j)/\tau\big)} \quad (6)$$

where $H(Y|X_i)$ is the predictive entropy of the observation stream and τ is a temperature hyperparameter. The predictive entropy $H(Y|X_i)$ is calculated as $-\sum_{y\in\mathcal{Y}} P_{obs}(y|X_i)\log P_{obs}(y|X_i)$. This mechanism ensures that the optimization focus is adaptively shifted: **Low-Entropy Samples (Low Cognitive Load):** The weight λ_i is suppressed, preventing the model from over-optimizing on trivial rationales. **High-Entropy Samples (High Cognitive Load):** The weight is intensified, compelling the model to break through the information bottleneck via high-order causal reasoning. This strategy transforms training into a journey of Curriculum Belief Revision, where the model progressively masters complex emotional logic by identifying where its intuition fails.

3.5 Training Strategy

We optimize π_θ using Group Relative Policy Optimization (GRPO). For each input X, we sample rationales $\{z_1, \ldots, z_k\}$ to compute advantages from $\mathcal{R}_{total}$. Intra-group reward normalization reframes the objective as a Relative Preference Ranking, which stabilizes gradient updates for lightweight models. To prevent policy collapse into linguistically degenerate sequences, we impose a KL-divergence constraint against the reference policy π_{ref}. Optimizing the expectation of the advantage ensures CAIR evolves into a logically self-consistent reasoner that transitions from superficial pattern matching to authentic, intervention-driven emotional intelligence.

4 Experiments

4.1 Experimental Setup

Datasets Preparation For training, we integrate data from multiple public sources—including MELD (Poria et al., 2019), Social-IQ (Zadeh et al., 2019), and IEMOCAP (Busso et al., 2008)—and augment them with human-annotated structured reasoning chains to construct a unified supervised fine-tuning dataset named EmoReason-SFT, which contains approximately 31,000 high-quality multimodal emotion reasoning samples. Building upon this foundation, we further design a curriculum learning strategy based on model uncertainty to identify difficult examples, thereby constructing a difficulty-aware dataset EmoReason-RL tailored for reinforcement learning to prioritize performance improvement on complex cases.

Evaluation Metrics We evaluate on EmoBench-M (Hu et al., 2025a) for real-world scenarios and MTMEUR (Hu et al., 2025b) for implicit emotions, using the CA-MER (Han et al., 2025) subset for cross-modal contradiction stress tests. We benchmark CAIR against various state-of-the-art models like Gemini-2.0-Flash and Emotion-LLaMA.To capture causal reasoning, we complement accuracy with interventional metrics: Information Gain $\Delta H = H(Y|X) - H(Y|X, z)$, where H denotes Shannon entropy. This measures the reduction in predictive uncertainty regarding emotion Y once rationale z is provided. Faith-F1 is the F1-score of causal keywords in the rationale z against keywords extracted from the reference reasoning chains prepared for the benchmark, measuring logical consistency between reasoning and output.

4.2 Implementation Details

We adopt the Qwen2.5-VL-7B architecture (Xu et al., 2025) , training in two stages to transition from imitation to causal reasoning. **Stage 1: SFT for Causal Initialization.** To instill causal templates and prevent cold-start collapse, we fine-tune on 31K structured instances. We use a learning rate of 2×10^{-5}, a global batch size of 32, and a cosine annealing schedule over 2 epochs. FlashAttention-2 and a 4096-token sequence limit are enabled to handle long videos. **Stage 2: Reinforcement Learning via GRPO.** We employ GRPO (Shao et al., 2024) to enhance stability and reduce memory overhead. Training utilizes 3K high-entropy samples from the SFT model, these samples repre-

Table 1: Emotion-specific accuracy (%) on MTMEUR. Best results in **bold**.

Method	Disgust	Sadness	Happy	Surprise	Excited	Angry	Fear	Overall
VideoChatGPT (Maaz et al., 2024)	27.11	24.30	28.20	23.80	36.20	26.60	27.81	29.10
ShareGPT4Video (Chen et al., 2024)	27.24	25.60	27.90	24.30	36.10	27.20	27.68	29.39
Chat-UniVi (Jin et al., 2024)	29.84	24.50	30.08	25.70	35.70	24.80	29.59	30.09
VTimeLLM (Huang et al., 2024a)	33.51	21.30	34.40	34.00	46.10	28.20	33.81	34.26
VideoLLaVA (Lin et al., 2024)	38.02	32.30	39.20	31.70	45.90	29.90	38.52	38.72
Emotion-LLaMA (Cheng et al., 2024a)	45.52	47.10	46.60	44.30	53.50	35.40	46.13	48.65
Qwen-VL-Chat (Bai et al., 2023)	52.40	46.50	56.70	54.40	64.10	44.70	53.29	56.45
MiniCPM-V-2.6 (Hu et al., 2024)	59.15	50.90	62.75	63.80	70.80	53.60	59.42	63.21
VideoLLaMA2 (Cheng et al., 2024b)	59.61	58.20	65.75	64.40	72.40	53.32	59.26	66.14
Qwen2-VL (Wang et al., 2024)	66.97	59.67	68.84	67.00	85.30	56.34	68.46	71.19
MTMEUR-base (Hu et al., 2025b)	68.32	60.08	71.90	69.20	87.20	56.50	69.31	72.93
CAIR (Ours)	**69.28**	**62.35**	**72.51**	**72.90**	**89.90**	**56.62**	**70.11**	**73.80**

sent "high-value regions" for training causal reasoning.The model generates $G = 8$ parallel responses per query. We set the learning rate to 1×10^{-6} and initialize η at 0.04 with dynamic adjustment, the temperature τ is set to 1.0, static hyperparameters $\alpha = 1.0$ and $\beta = 1.0$. Training was completed in 42 hours on 8 NVIDIA A800 GPUs using DeepSpeed ZeRO-3.

4.3 Main Results

We evaluate the performance of CAIR by benchmarking it against an extensive array of state-of-the-art (SOTA) multimodal large language models (MLLMs) and specialized reinforcement learning (RL) baselines. The comparative results, summarized in Table 1 and Table 2, expose a fundamental limitation in existing paradigms. While generalist models achieve respectable accuracy through surface-level pattern matching, they often struggle with deep emotional reasoning.

On the MTMEUR benchmark (Table 1), CAIR establishes a new state-of-the-art with an overall accuracy of 73.80%. This performance surpasses the specialized MTMEUR-base. A granular analysis across emotion categories shows that CAIR excels in identifying complex and implicit states. For instance, it achieves 72.90% in *Surprise* and 62.35% in *Sadness*. These categories are particularly challenging because they often involve subtle contextual triggers rather than overt facial expressions. The substantial improvement over models that rely on direct perception suggests that our Causal Mediation Reward (CMR) is effective. By quantifying the interventional utility of rationales, CMR guides the model to resolve the information bottleneck inherent in nuanced emotional scenarios.

Does this superior performance hold when models face different cognitive demands? We explore this by examining the EmoBench-M results in Table 2. We observe that while CAIR slightly underperforms Gemini-2.0-Flash, it achieves the highest performance among all specialized and reinforcement learning-based models on the SCEA subset. CAIR reaches an accuracy of 68.5% on SCEA, which requires deep psychological attribution. This significantly outperforms both the SFT baseline (54.1%) and the Naive RL approach (62.5%).

This shift in performance across task complexities highlights the impact of our Cognitive-load Adaptive Optimization (CLO). In low-entropy tasks like FER, the model can rely on perceptual intuition. However, in high-entropy SCEA tasks, the reasoning signal becomes critical. CAIR demonstrates a superior ability to generate "evidence anchors" in these difficult cases. This is further evidenced by the Information Gain (ΔH) metrics. CAIR achieves a ΔH of 0.19 on high-load tasks, nearly doubling that of the GPT-4o baseline. These results confirm that CAIR does not just mimic the appearance of reasoning. Instead, it learns to utilize rationales as essential causal mediators for accurate emotion attribution.

4.4 Ablation Study

To further dissect the contribution of each component within the CAIR framework, we perform a systematic ablation study on the MTMEUR dataset. The results, as detailed in Table 3, illustrate how different reward terms influence the model's interventional utility. Starting from the SFT base, which achieves a modest ΔH of 0.05 and a Faith-F1 score of 37.5%, we observe that adding standard accu-

Table 2: Performance comparison on EmoBench-M. We specifically report results on FER (Low Cognitive Load) and SCEA (High Cognitive Load) to evaluate the effectiveness of Causal Mediation. ΔH measures the interventional information gain in bits; Faith-F1 evaluates logical consistency. Naive RL baseline utilizes the standard GRPO framework but optimizes only for label accuracy and format compliance

Method	Accuracy (%) ↑			Causal Faithfulness ↑		Interventional Utility ↑	
	FER	SCEA	Avg.	Faith-F1 (%)	ROUGE-L	ΔH (Low)	ΔH (High)
Human Benchmark (Hu et al., 2025a)	62.0	72.7	73.0	-	-	-	-
Generalist MLLMs (Zero-shot)							
GPT-4o (Hurst et al., 2024)	58.4	62.1	60.5	46.2	38.5	0.04	0.09
Qwen2.5-VL-72B-Instruct (Xu et al., 2025)	53.0	72.5	57.8	44.5	37.2	0.05	0.11
Gemini-2.0-Flash (Comanici et al., 2025)	61.4	72.0	62.3	48.1	40.4	0.07	0.12
Specialized SFT Models							
Emotion-LLaMA (Cheng et al., 2024a)	36.9	54.1	40.6	37.5	31.2	0.05	0.06
Video-LLaMA2-7B (Cheng et al., 2024b)	45.4	61.3	47.1	38.9	32.5	0.06	0.07
Reinforcement Learning Baselines							
Naive RL	46.2	62.5	51.4	40.1	33.8	0.08	0.12
R1-Omni (Zhao et al., 2025)	52.1	66.8	54.9	42.4	35.6	0.10	0.15
CAIR (Ours)	**52.8**	**68.5**	**56.4**	**51.2**	**41.8**	**0.12**	**0.19**

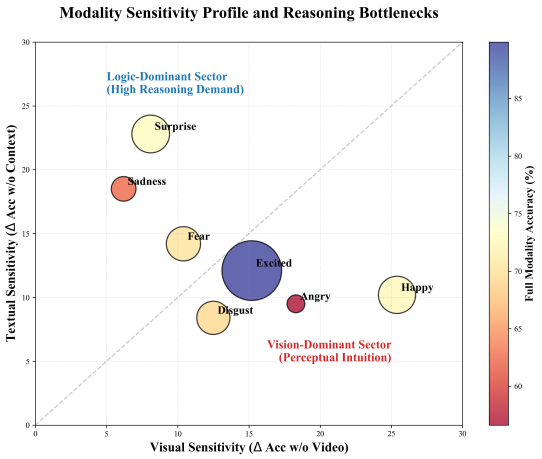

Figure 3: Cross-modal sensitivity profile across emotional categories. The bubble size correlates with the full-modality accuracy. The distribution reveals two distinct clusters: "Vision-Dominant" states (e.g., *Happy*, *Angry*) that rely on perceptual cues, and "Logic-Dominant" states (e.g., *Surprise*, *Sadness*) where the rationale z provides the primary interventional utility.

racy and format rewards yields marginal gains in reasoning quality. This confirms our intuition that optimizing for surface-level correctness does not inherently foster logical depth.

Upon Causal Mediation Reward (CMR) inclusion, the information gain ΔH surges from 0.12 to 0.32, while the Faith-F1 score jumps from 40.1% to 52.7%. This substantial improvement validates our core hypothesis: treating the reasoning chain as a causal mediator, rather than a mere linguistic sequence, is essential for faithful emotion reasoning. By forcing the model to generate rationales that actively reduce predictive uncertainty, CMR effectively filters out "vacuous" descriptions that look fluent but lack evidentiary value. Our observations further confirm that this efficacy is strictly content-dependent; for instance, replacing meaningful rationales with shuffled tokens or irrelevant descriptions leads to significant degradation in information gain ΔH. This ensures that the observed performance leap is driven by authentic causal discovery rather than generic context augmentation from intermediate text.

Table 3: Ablation of reward components on MTMEUR. **Faith-F1** measures logical consistency; ΔH is measured in bits.

Configuration	Acc	Faith-F1	ΔH
SFT Base	60.2	37.5	0.05
+ Acc Reward	63.8	38.2	0.09
+ Format Reward	65.5	40.1	0.12
+ CMR (CAIR)	**73.8**	**52.7**	**0.32**

Beyond reward design, we also examine the role of the CLO mechanism in handling task complexity. As shown in Table 4, our approach outperforms two alternative strategies: (1) *Uniform RL*, which treats all samples equally and exhibits significant instability in high-conflict scenarios (IQR = 8.2), and (2) *Curriculum RL*, which follows a heuristic difficulty schedule and improves stability over *Uniform RL* but remains suboptimal. In contrast, CAIR—powered by CLO—dynamically intensifies optimization on high-entropy samples to

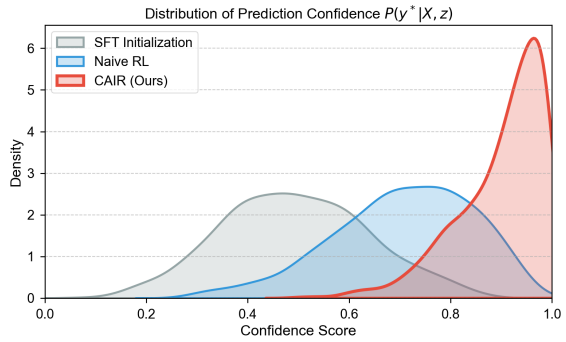

Figure 4: KDE analysis reveals CAIR concentrates prediction confidence in the >0.85 range with 42% narrower variance than SFT, confirming causal rationales as evidence anchors that stabilize decision boundaries.

Table 4: CLO performance on CA-MER. **Cons. Acc** and **Conf. Acc** denote accuracy on samples with synergistic and contradictory multimodal cues, respectively. **Interquartile Range (IQR)** ↓ measures stability (lower = more robust).

Strategy	Cons. Acc	Conf. Acc	IQR ↓
Uniform RL	72.1	58.4	8.2
Curriculum RL	73.5	61.2	6.5
CAIR (CLO)	**74.8**	**68.7**	**3.8**

alleviate the information bottleneck in ambiguous cases. This yields a much lower IQR of 3.8 and boosts conflict accuracy to 68.7%, demonstrating that CLO enables the model to master complex emotional reasoning without compromising stability on simpler tasks.

The Modality Sensitivity Profile (Figure 3) analyzes emotional dependencies by mapping performance degradation from ablating visual vs. textual streams. A diagonal divergence emerges: *Happy* and *Angry* occupy the Vision-Dominant Sector, indicating that surface-level perceptual signals are sufficient for these intuitive judgments. Conversely, complex states like *Surprise* and *Sadness* gravitate toward the Logic-Dominant Sector, where rationale z provides essential "causal threads" to resolve cross-modal inconsistencies. Large bubble sizes in this zone confirm that CAIR mitigates the "Information Bottleneck" by leveraging rationales as anchors when visual intuition is ambiguous.

4.5 Visual and Statistical Analysis

Confidence Distribution As illustrated in Figure4, KDE analysis reveals a distinct "Right Shift" in CAIR's prediction confidence compared to SFT and Naive RL, with probability mass concentrated above 0.85. CAIR achieves a 42% narrower variance in its posterior distribution than SFT(quantified by the reduction in standard deviation σ of posterior distribution), suggesting rationales function as "evidence anchors". By utilizing reasoning chains to solidify belief states rather than relying on surface correlations, the model achieves more decisive and stable emotion attributions.

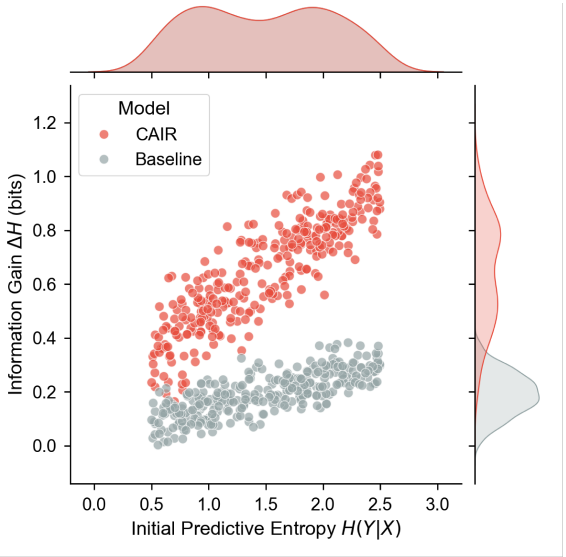

Figure 5: CAIR generates rationales with higher information density in high-entropy samples, demonstrating causal mediation's role in belief revision for complex emotional logic. The Baseline denotes the SFT-initialized model, representing the model's intuitive perception without causal reinforcement.

Mediation Effect across Cognitive Loads To evaluate reasoning reliance, we plot Initial Predictive Entropy against Information Gain (ΔH) in Figure5. While info gain remains low for low-entropy samples with overt cues, CAIR provides significantly higher ΔH as emotional ambiguity increases. This validates a "Belief Revision" process. In scenarios where raw visual signals are muddled or contradictory, CAIR generates high-density rationales, resolving the information bottleneck.

5 Summary

In this work, we address the challenge of equipping multimodal models with *causally grounded* emotional reasoning, moving beyond simple labeling to event-driven explanations. We propose CAIR, a multi-dimensional reinforcement learning framework grounded in information theory and cognitive plausibility. CAIR utilizes the Causal Mediation Reward (CMR) to quantify the utility

of rationales in driving belief revision, alongside a Cognitive-load Adaptive Optimization (CLO) mechanism that reweights samples by predictive entropy. Stabilized by GRPO and a two-phase curriculum, CAIR achieves competitive accuracy and Faith-F1 on EmoBench-M and MTMEUR while exhibiting robustness against reward hacking.

Limitations

CAIR currently assumes access to ground-truth emotion labels during reward computation, which may limit its direct application in fully unsupervised settings. While the Causal Mediation Reward (CMR) and Cognitive-load Adaptive Optimization (CLO) successfully mitigate reward hacking and promote logical faithfulness, the reasoning chains are generated autoregressively without explicit structural constraints, which might affect consistency in extremely long-form or multi-agent scenarios. Additionally, our evaluation primarily focuses on English-language video-text benchmarks; the generalizability to low-resource languages or varied cultural contexts remains to be further explored. Finally, the interventional utility measured via log-likelihood ratios can be sensitive to the base model's initial calibration, suggesting that future work could incorporate adversarial reward validation to further enhance robustness.

Acknowledgments

This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB0500103, the National Natural Science Foundation of China(NSFC) (No.62276259,No.U2436210,No.62271083).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

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.

Linda Camras. 1980. Emotion: a psychoevolutionary synthesis.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, and 1 others. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495.

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

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024b. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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

Zhiyuan Han, Beier Zhu, Yanlong Xu, Peipei Song, and Xun Yang. 2025. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5528–5537.

He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. 2025a. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*.

Jinpeng Hu, Hongchang Shi, Chongyuan Dai, Zhuo Li, Peipei Song, and Meng Wang. 2025b. Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5814–5823.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and 1 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024a. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280.

Zhaopei Huang, Jinming Zhao, and Qin Jin. 2024b. Ecr-chain: Advancing generative language models to better emotion-cause reasoners through reasoning chains. *arXiv preprint arXiv:2405.10860*.

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

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Garapati Keerthana and Manik Gupta. 2025. Towards emotionally intelligent and responsible reinforcement learning. *arXiv preprint arXiv:2511.10573*.

Pengcheng Li, Botao Zhao, Zuheng Kang, Junqing Peng, Xiaoyang Qu, Yayun He, and Jianzong Wang. 2025. Emo-rl: Emotion-rule-based reinforcement learning enhanced audio-language model for generalized speech emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18744–18754.

Zheng Lian, Haiyang Sun, Licai Sun, Jiangyan Yi, Bin Liu, and Jianhua Tao. 2024. Affectgpt: Dataset and framework for explainable multimodal emotion recognition. *arXiv preprint arXiv:2407.07653*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602.

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

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*.

Yuanyuan Sun and Ting Zhou. 2025. Dialoguemllm: transforming multimodal emotion recognition in conversation through instruction-tuned mllm. *IEEE Access*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

Cong Wang, Changfeng Gao, Yang Xiang, Zhihao Du, Keyu An, Han Zhao, Qian Chen, Xiangang Li, Yingming Gao, and Ya Li. 2025a. Rrpo: Robust reward policy optimization for llm-based emotional tts. *arXiv preprint arXiv:2512.04552*.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, and 1 others. 2025b. Rlver: Reinforcement learning with verifiable emotion rewards for empathetic agents. *arXiv preprint arXiv:2507.03112*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Qize Yang, Detao Bai, Yi-Xing Peng, and Xihan Wei. 2025a. Omni-emotion: Extending video mllm with detailed face and audio modeling for multimodal emotion analysis. *arXiv preprint arXiv:2501.09502*.

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

Ting Yang, Li Chen, and Huimin Wang. 2025b. Towards open-ended emotional support conversations in llms via reinforcement learning with future-oriented rewards. *arXiv preprint arXiv:2508.12935*.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, pages 2236–2246.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Ren Zhang, Guoliang Xu, Tianyu Zhou, Junjun Zhao, Bo Yu, Zhicheng Zhang, and Jianqin Yin. 2025. Emosynergy: Synergizing task-specific and general experts for multimodal emotion recognition. In *Proceedings of the 3rd International Workshop on Multimodal and Responsible Affective Computing*, pages 30–34.

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

A Appendix

A.1 Additional Implementation Details

To ensure the reproducibility of our results and provide transparency regarding the training process, we elaborate on the computational infrastructure and the two-stage optimization pipeline used for CAIR.

Hardware Infrastructure and Memory Management All experiments were conducted on a high-performance compute cluster featuring 8 NVIDIA A800 (80GB) GPUs. These units are interconnected via NVLink to facilitate low-latency gradient synchronization during the full-parameter fine-tuning phase. Given the memory-intensive nature of processing long-form video data alongside a 7B-parameter language model, we utilized DeepSpeed ZeRO-3. This strategy partitions model states, gradients, and optimizer states across all distributed GPUs, significantly reducing the memory footprint per device. To further optimize throughput, we integrated FlashAttention-2, enabling the model to handle multimodal sequences up to 4,096 tokens without the quadratic memory growth typical of standard attention mechanisms.

Stage 1: Full-Parameter Supervised Fine-Tuning (SFT) Unlike traditional methods that rely on parameter-efficient techniques like LoRA, we performed a full-parameter fine-tuning of the Qwen2.5-VL-7B backbone. We believe this allows the model to more deeply internalize the complex cross-modal alignments required for emotional reasoning. We utilized the EmoReason-SFT dataset, comprising 31,000 samples. Training was conducted over 2 full epochs with a global batch size of 32, resulting in approximately 1,940 optimization steps. The learning rate was set to 2×10^{-5} with a cosine decay schedule and a 5% warmup period to stabilize the initial weight updates.

The initial performance shift observed during the SFT phase (Stage 1) reflects a standard behavioral recalibration (often termed the 'alignment tax') required to strictly internalize structured reasoning templates . During this stage, the model learns to prioritize rigorous logical scaffolding over intuitive pattern-matching shortcuts . Although this calibration may cause a temporary accuracy fluctuation, it establishes the indispensable causal threads that enable the subsequent reinforcement learning phase to reach a breakthrough performance beyond the original zero-shot capabilities .

Stage 2: Reinforcement Learning via GRPO Following the initialization from the SFT stage, the model underwent reinforcement learning using the Group Relative Policy Optimization (GRPO) algorithm. We specifically curated 3,000 "high-entropy" samples(hereafter referred to as EmoReason-RL) where the SFT model exhibited predictive uncertainty, targeting the model's capacity for belief revision. For each input, we sampled a group size of $G = 8$ to compute relative advantages within the cohort. The RL phase spanned 500 steps with a learning rate of 1×10^{-6}. The entire training pipeline was completed in 42 hours.

To evaluate the logical consistency of rationales, the ground-truth keywords for Faith-F1 were derived from human-annotated reasoning chains specifically prepared for MTMEUR and EmoBench-M. We utilized KeyBERT to extract the top-5 nouns and verbs representing causal triggers (e.g., "confession," "betrayal") from these reference rationales. These keywords were fixed prior to evaluation and remained independent of model outputs, ensuring the metric objectively reflects the model's adherence to human-like causal logic during the inference phase. Following the inherent task hierarchy of MTMEUR, we treat the ground-truth of the emotion recognition sub-task in each session as the authoritative category anchor. This allows us to map subsequent reasoning questions to their respective official emotion categories for the granular analysis.

A.2 Training Stability and Convergence Analysis

To provide a deeper understanding of the CAIR optimization trajectory, we visualize the progression of key metrics across both training phases in Figure

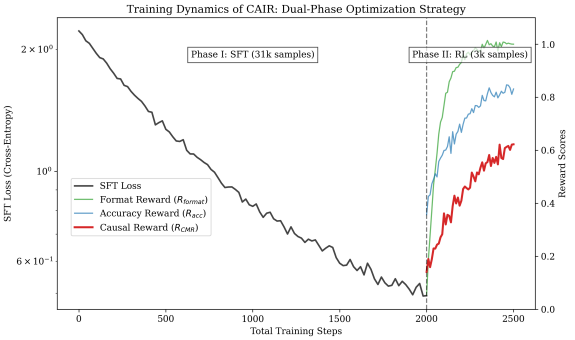

Figure 6: Optimization dynamics during the two-phase training process of CAIR. The primary axis (left) denotes the cross-entropy loss during the full-parameter supervised fine-tuning (SFT) phase, while the secondary axis (right) illustrates the progression of diverse reward components—including format compliance (R_{format}), accuracy (R_{acc}), and causal mediation (R_{CMR})—during the reinforcement learning phase. The transition point indicates the shift from distributional alignment on 31k samples to causality-driven optimization on 3k high-entropy samples.

6. This visualization confirms the stability of our two-stage approach and highlights how different reward signals interact during the policy refinement stage.

Stage 1: Distributional Alignment via SFT. As illustrated in the left region of the figure, the cross-entropy loss during the full-parameter supervised fine-tuning phase initiates at approximately 2.2. This starting point reflects the initial semantic gap between the general-purpose pre-trained weights of Qwen2.5-VL-7B and the specialized multimodal reasoning patterns in the EmoReason-SFT dataset. Over the course of 1,940 steps (representing 2 epochs on 31,000 samples with a batch size of 32), the loss exhibits a smooth exponential decay, eventually stabilizing near 0.45. This successful convergence indicates that the model has effectively internalized the basic templates of emotional explanation and cross-modal mapping before entering the more complex reinforcement learning phase.

Stage 2: Multi-objective Optimization via RL. Upon transitioning to the reinforcement learning phase, we observe a distinct divergence in the learning rates of the three reward components, with the format reward (R_{format}) reaching near-maximal values within the first 100 steps. This rapid saturation suggests that the preceding SFT phase provided a sufficiently strong structural bias, allowing the GRPO algorithm to quickly enforce tag compliance with minimal exploration cost.

Following the stabilization of output formats, the accuracy reward (R_{acc}) shows a steady upward trend, while the Causal Mediation Reward (R_{CMR}) exhibits a more gradual but consistent improvement. Since R_{CMR} measures interventional utility in resolving predictive uncertainty, its late-stage growth signifies that the model is moving beyond surface-level pattern replication to master the "latent logic" of emotion via reasoning chains that function as genuine evidence anchors. The overall stability of these curves, maintained under a dynamic KL-divergence constraint, demonstrates that CAIR effectively balances the exploitation of high-reward causal paths with the preservation of linguistic fluency, ensuring the final model is a logically consistent reasoner capable of handling complex information bottlenecks.

A.3 Formal Derivation of Reward Components

The objective of CAIR is to optimize a composite reward function $R_{total} = \alpha R_{acc} + \beta R_{format} + \lambda_i R_{CMR}$. The design of this reward system is grounded in the principles of Reinforcement Learning from Verifiable Feedback (RLVF) and Causal Mediation Analysis. Here, we formally define each component and the underlying theoretical rationale.

Accuracy Reward (R_{acc}): Dense Signal for Correct Attribution Standard reinforcement learning for complex reasoning tasks often suffers from the "sparse reward" problem. In a multi-class emotion classification setting, a binary 0/1 reward provides no gradient information for incorrect trajectories, leading to inefficient exploration. To mitigate this, we employ the principle of Reward Shaping. We define R_{acc} as:

$$R_{acc} = \begin{cases} 1, & \text{if } y = y^* \\ \text{ROUGE-L}(y, y^*), & \text{otherwise} \end{cases} \quad (7)$$

where y is the predicted label and y^* is the ground truth. The integration of ROUGE-L serves as a "soft" distance metric. By rewarding the model for generating labels that are semantically or linguistically proximate to the target category, we provide a dense and continuous feedback signal. This ensures that the policy π_θ remains oriented toward the correct region of the label space even during early-stage exploration.

Format Compliance Reward (R_{format}): Structural Integrity for Interventional Parsing To ensure that the model output O follows the structured format $[< reason >]z[< /reason >][< answer >]y[< /answer >]$, we define a binary indicator reward:

$$R_{format} = \mathbf{I}(\text{tags in correct order and complete}) \tag{8}$$

This reward is crucial not only for task compliance but also as a prerequisite for the Dual-Path Intrinsic Verifier. Since our framework requires the automated extraction of the rationale z to perform the semantic intervention $do(Z = z)$, any structural failure in the output string would render the causal evaluation impossible. Thus, R_{format} acts as a hard constraint to stabilize the parsing interface between the policy model and the reward function.

Causal Mediation Reward (R_{CMR}): Quantifying Interventional Utility The most distinctive component of CAIR is the R_{CMR}, which evaluates the quality of rationale z through the lens of Interventional Influence. Drawing from Judea Pearl's structural causal models (SCM), we treat z as a mediator between the perceptual input X and the emotional decision Y. The reward is formulated as a log-likelihood ratio:

$$R_{CMR} = \log \frac{P(y^*|X, do(z))}{P(y^*|X)} \tag{9}$$

where $P(y^*|X)$ represents the observational "intuitive" belief, and $P(y^*|X, do(z))$ represents the interventional belief revision. Theoretically, this ratio measures the Individual Causal Effect of the generated rationale. If z contains redundant or superficial information (e.g., "the person is smiling"), the posterior belief remains stagnant ($R_{CMR} \approx 0$). Conversely, if z identifies a critical latent trigger (e.g., "perceived betrayal of trust"), it significantly boosts the model's confidence in the correct label y^*, yielding a high positive reward. This mechanism forces the model to move past vacuous surface descriptions and instead prioritize rationales that function as indispensable "evidence anchors".

A.4 The CAIR Optimization Algorithm and Training Stability

The optimization of a 7B-parameter multimodal model through reinforcement learning presents unique challenges, particularly regarding computational overhead and gradient stability. To address the inherent limitations of standard Proximal Policy Optimization (PPO)—which necessitates the maintenance of an auxiliary Value Function (Critic) and often doubles the GPU memory footprint—we adoptGroup Relative Policy Optimization (GRPO) as our primary training framework.

Algorithm 1 CAIR Optimization via GRPO

Require: SFT-initialized policy π_θ, reference policy π_{ref}, dataset $\mathcal{D}$, group size $G = 8$.

while not converged **do**
  Sample a batch of multimodal contexts $\{X_i\}_{i=1}^{B}$ from $\mathcal{D}$.
  for each X_i **do**
    Generate G diverse outputs $\{O_{i,j}\}_{j=1}^{G} \sim \pi_\theta(O|X_i)$.
    Parse $\{O_{i,j}\}$ into rationales $\{z_{i,j}\}$ and answers $\{y_{i,j}\}$.
    for each response $j \in \{1, \ldots, G\}$ **do**
      Compute R_{acc} and R_{format} based on ground truth y_i^*.
      Compute observational belief $P(y_i^* \mid X_i)$ via the verifier.
      Compute interventional belief $P(y_i^* \mid X_i, \text{do}(z_{i,j}))$ via path B.
      Calculate
      $$R_{\text{CMR}} = \log \frac{P(y_i^* \mid X_i, \text{do}(z_{i,j}))}{P(y_i^* \mid X_i)}.$$
      Aggregate
      $$r_{i,j} = \alpha R_{\text{acc}} + \beta R_{\text{format}} + \lambda_i R_{\text{CMR}}.$$
    end for
    Compute group advantage:
    $$A_{i,j} = \frac{r_{i,j} - \text{mean}(\{r_{i,k}\}_{k=1}^{G})}{\text{std}(\{r_{i,k}\}_{k=1}^{G})}.$$
  end for
  Update θ by maximizing the GRPO objective with KL constraint:
  $$\mathcal{J}(\theta) = \mathbf{E}\left[\frac{\pi_\theta(O|X)}{\pi_{\theta_{\text{old}}}(O|X)} A - \eta\, \text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})\right]$$
end while

The core philosophy behind selecting GRPO for CAIR is two-fold. First, from a resource perspective, eliminating the Critic network allows us to allocate the saved memory to handle high-resolution visual tokens from datasets. This ensures that the full-parameter fine-tuning of the Qwen2.5-VL-7B backbone remains feasible on a standard cluster of 8 NVIDIA A800 GPUs. Second, from a learning perspective, GRPO facilitates Relative Preference Learning. In complex tasks like causal emotion

reasoning, assigning an absolute scalar reward to a specific rationale z is often noisy. By sampling a group of $G = 8$ diverse outputs for each multimodal context X_i, the model can effectively perform a cohort-based comparison. The advantage $A_{i,j}$ is then derived by normalizing the reward $r_{i,j}$ against the group mean, which naturally filters out sample-specific difficulty variance and focuses the gradient update on the relative causal faithfulness of the generated explanations.

The integration of the Causal Mediation Reward (CMR) into the GRPO loop is a critical feature of our implementation. Unlike traditional RLVR (Reinforcement Learning from Verifiable Rewards) which targets the final answer alone, our algorithm requires each generated candidate $j \in \{1, \ldots, G\}$ to pass through the Dual-Path Intrinsic Verifier. For every sample in the group, we compute the observational belief $P(y^*|X)$ (Path A) and the interventional belief $P(y^*|X, do(z))$ (Path B). The resulting log-likelihood ratio R_{CMR} serves as a direct measure of the "individual causal effect" of the rationale z. By aggregating this with R_{acc} and R_{format}, the model is intrinsically incentivized to favor rationales that not only look fluent but actively resolve the model's predictive uncertainty.

To prevent "policy collapse"—a phenomenon where the model generates high-reward but linguistically degenerate or repetitive sequences—we impose a strict KL-divergence constraint relative to the SFT-initialized reference policy π_{ref}. The coefficient β is dynamically adjusted to ensure that the model explores new causal reasoning pathways while remaining anchored to human-understandable language. This two-stage transition, starting from a 31k-sample SFT initialization and concluding with a targeted 3k-sample RL refinement, ensures that CAIR achieves a robust balance between cross-modal perception and high-level logical derivation. The detailed procedural flow for this optimization is formalized in Algorithm 1.

A.5 Entropy-Based Sample Curation

The efficiency of reinforcement learning, particularly within the GRPO framework, is heavily contingent upon the quality of the training trajectories. Rather than performing RL on the entire 31,000 samples—which would incur prohibitive computational costs and potentially lead to gradient dilution—we implement a "difficulty-aware" curation strategy. We specifically identify 3,000 high-entropy samples where the SFT-initialized model exhibits the greatest predictive uncertainty.

The Principle of Information Bottleneck Our selection process is grounded in the Information Bottleneck principle. We hypothesize that samples with low predictive entropy represent scenarios where the model's intuitive "pattern matching" is already sufficient. In contrast, samples with high entropy indicate a failure of surface-level multimodal alignment, necessitating the construction of a logical rationale to resolve ambiguity. To quantify this, we pass all 31,000 samples through the SFT model and compute the Shannon entropy of the observational belief distribution $P(Y|X)$:

$$H(Y|X) = -\sum_{y \in \mathcal{Y}} P(y|X) \log P(y|X) \quad (10)$$

where $\mathcal{Y}$ denotes the set of emotion categories. As shown in Figure 7, the entropy distribution across the dataset is non-uniform, with a significant "long tail" of complex instances.

Curation Process and Thresholding We rank the 31,000 samples by their entropy values and select the top 3,000 instances to form the RL training subset. This partitioning effectively splits the dataset into two functional groups: a stable subset of 28,000 samples and a high-entropy target subset of 3,000 samples. Empirically, these curated samples predominantly consist of socially complex emotions, such as Surprise and Sadness, where visual cues often contradict contextual behaviors.

Why 3,000 Samples Choosing 3,000 samples (approximately 10% of the original dataset) serves a dual purpose. First, it ensures that the GRPO algorithm focuses its group-relative advantage computation on cases where the "reasoning signal-to-noise ratio" is highest. Second, it facilitates a more stable optimization of the Causal Mediation Reward (CMR). By exposing the model to scenarios where its intuitive judgment is most muddled, we maximize the potential for "Belief Revision" via the rationale z. This strategy transforms the reinforcement learning phase into a targeted curriculum, where the model masters complex emotional logic without being distracted by trivial instances that it has already solved during the SFT stage.

A.6 Detailed Performance Breakdown on MTMEUR

As demonstrated in the experimental results in Section 4.3, CAIR establishes a new state-of-the-art

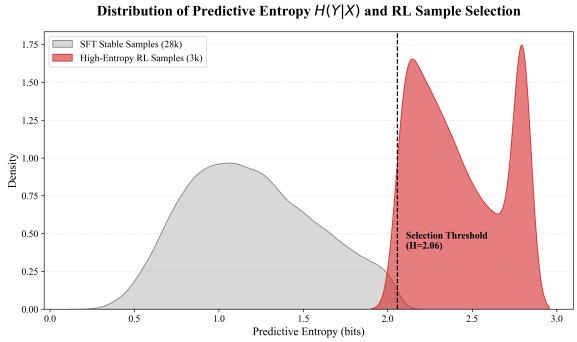

Figure 7: Probability density function (PDF) of the predictive entropy $H(Y|X)$ for the entire 31,000 SFT samples. The red-shaded region represents the 3,000 high-entropy samples selected for the reinforcement learning phase, while the grey region represents the 28,000 stable samples. The thresholding identifies the information bottleneck where reasoning becomes indispensable.

across nearly all emotional dimensions on the MTMEUR dataset. To provide a more intuitive understanding of the model's localized strengths, we present a granular performance breakdown in Figure 8. This visualization compares our framework against two formidable baselines: MTMEUR-base, a specialized architecture fine-tuned specifically for multimodal emotion understanding, and VideoLLaMA2, a state-of-the-art generalist large vision-language model.

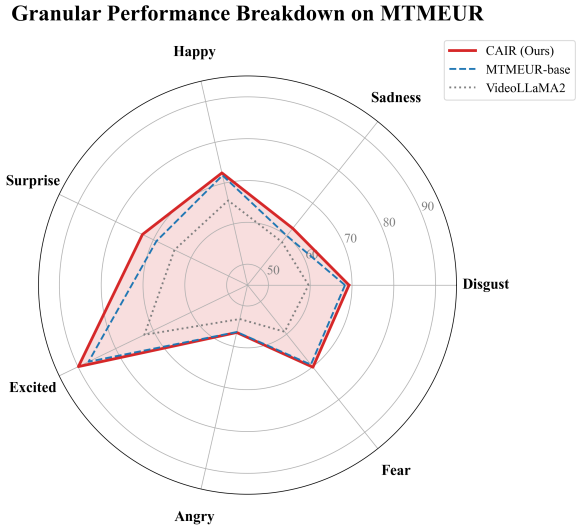

Figure 8: Radar chart illustrating the reasoning gap across 7 emotion categories on MTMEUR. Radar chart analysis of emotion-specific accuracy (%) on the MTMEUR benchmark. CAIR (red line) consistently outperforms both specialized and generalist baselines, particularly in complex emotional categories.

The radar chart reveals a significant "reasoning gap" in high-entropy emotional categories. For instance, in the Surprise category, CAIR achieves an accuracy of 72.90%, which represents a substantial margin over MTMEUR-base (69.20%) and VideoLLaMA2 (64.40%). Such categories often exhibit fleeting micro-expressions and high contextual ambiguity, which typical SFT-based models fail to resolve. While generalist models often default to the most frequent distribution (e.g., misinterpreting surprise as generic happiness), CAIR utilizes the Causal Mediation Reward (CMR) to anchor its prediction in the underlying causal triggers identified within the rationale z.

A similar trend is observed in the Sadness dimension, where CAIR attains 62.35% accuracy. The logic behind this improvement is rooted in the model's ability to distinguish between overt distress and subtle, implicit sorrow. By rewarding rationales that provide a measurable "interventional utility," CAIR effectively filters out vacuous descriptive markers—such as "the person looks down"—and instead captures indispensable causal mechanisms, such as "internalized grief following a social loss".

The visual "envelopment" of the baselines by CAIR's performance boundary in Figure 8 confirms that our approach does not achieve its gains through a few outliers. Instead, the combination of CMR and Cognitive-load Adaptive Optimization (CLO) provides a systematic uplift across the entire emotional spectrum. This consistent superiority across diverse emotions like Fear (70.11%) and Exitement (89.90%) proves that grounding reinforcement learning in causal mediation is a robust strategy for mastering the profound logic of human emotion.

A.7 Cognitive Load and Stability Analysis

The intrinsic heterogeneity of human emotions dictates that not all emotional states require the same depth of logical derivation. To further investigate how CAIR handles varying task complexities, we conduct a comparative analysis against Naive RL and SFT baselines across different cognitive loads, specifically focusing on the FER and SCEA subsets of EmoBench-M. We evaluate CLO against two baselines: (1) Uniform RL, a vanilla GRPO implementation where the optimization weight λ_i is fixed at 1.0, treating all samples equally ; and (2) Curriculum RL, which employs a static difficulty schedule based on initial predictive entropy

$H(Y|X)$, training on simpler samples before progressing to complex ones . Unlike these static approaches, CLO dynamically reallocates credit based on real-time uncertainty.

Performance Dynamics across Information Bottlenecks As summarized in Table 2, CAIR achieves a specialized accuracy of 52.8% on FER and a remarkably high accuracy of 68.5% on the SCEA subset. This results in an overall average accuracy of 56.4%, representing a 15.8% absolute improvement over the Emotion-LLaMA SFT baseline (40.6%). From the perspective of the Information Bottleneck principle, FER tasks possess a wide "perceptual gate" where the observation stream $P(Y|X)$ already provides sufficient confidence. In contrast, the SCEA subset—which involves deep psychological attribution—presents a severe bottleneck where raw visual signals are often ambiguous. Our Cognitive-load Adaptive Optimization (CLO) successfully identifies these high-entropy scenarios and intensifies the reward signal. This compels the model to seek "evidentiary anchors" to resolve predictive uncertainty, leading to a substantial 14.4% gain on SCEA compared to SFT (54.1%).

Optimization Stability and Policy Robustness Beyond raw accuracy, a critical contribution of CLO is the stabilization of the reinforcement learning trajectory. Traditional RL strategies often suffer from gradient noise when forced to optimize complex reasoning on "trivial" samples, leading to performance fluctuations. As evidenced by our stress tests on the CA-MER subset (Table 4), CAIR significantly reduces the Interquartile Range (IQR) from 8.2 to 3.8. This reduction in variance indicates that by adaptively reallocating credit based on the model's Reasoning Signal-to-Noise Ratio (R-SNR), CLO prevents the policy from over-fitting to superficial patterns. Instead, the optimization focus is dynamically shifted toward high-value regions of the emotional landscape.

In summary, the empirical evidence presented across Table 2 and Figure 9 suggests that the synergy between the Causal Mediation Reward (CMR) and Cognitive-load Adaptive Optimization (CLO) facilitates a fundamental shift in model behavior. CAIR no longer functions as a mere statistical pattern matcher that mimics frequent labels found in the training distribution. Instead, it evolves into a resource-aware reasoner.

By dynamically intensifying its "cognitive effort"—quantified as the optimization weight

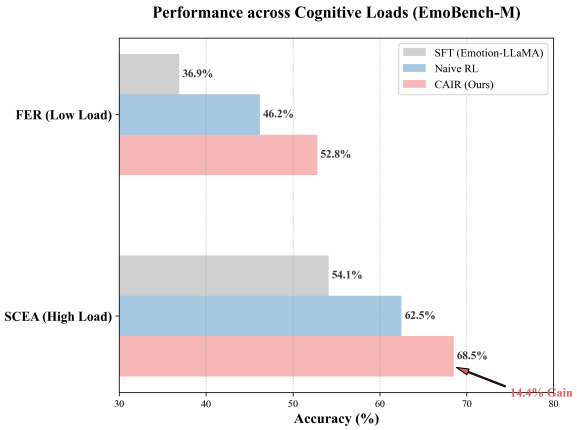

Figure 9: Comparative analysis of accuracy across low cognitive load (FER) and high cognitive load (SCEA) subsets on the EmoBench-M benchmark. The results demonstrate that while performance improves across the board, the sustained lead of CAIR in the SCEA subset—surpassing the SFT baseline by 14.4% absolute—highlights the efficacy of the CLO mechanism in bridging the reasoning gap. This confirms that CAIR successfully prioritizes high-entropy scenarios where surface-level perceptual intuition typically fails.

λ_i—on instances where the initial predictive entropy $H(Y|X)$ is high, the model achieves a robust transition from superficial mimicry to authentic, logically grounded emotional intelligence. This is particularly evident in the SCEA results, where the model must navigate complex social attributions. The substantial information gain ($\Delta H = 0.19$) observed in high-load tasks further confirms that the generated rationales z serve as essential causal mediators, enabling the model to perform a successful Belief Revision when raw perceptual signals are muddled or contradictory. This dual-path verification ensures that the model's decisions are not only accurate but also logically faithful to the underlying causal mechanisms of human affect.