PROCRIT: SELF-ELICITED MULTI-PERSPECTIVE REASONING WITH CRITIC-GUIDED REVISION FOR MULTIMODAL SARCASM DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sarcasm detection requires reasoning over cross-modal incongruities between literal expression and intended meaning, yet the analytical perspectives needed to uncover these incongruities vary across samples. While recent methods make this analytical process explicit, they still rely on fixed, predefined perspectives that operate independently under hand-crafted routing rules. We argue that multimodal sarcasm detection instead calls for *self-elicited multi-perspective reasoning*, where a model autonomously generates the perspectives needed for each sample and progressively integrates them into a coherent analysis. To realize this goal, we propose **ProCrit**, a **Pro**posal–**Crit**ic two-agent framework with a proposal agent for multi-perspective reasoning and a critic agent for external evaluation and targeted revision guidance. First, to overcome the lack of process-level supervision in existing sarcasm datasets, we synthesize reasoning annotations through a *dynamic-role agentic rollout*, in which a strong vision-language model sequentially spawns analytical roles within a shared context; the resulting trajectories are flattened into single sequences that preserve cross-perspective dependencies while enabling efficient autoregressive generation. Second, ProCrit adopts a *draft–critique–revise* paradigm in which the proposal agent dynamically elicits and progressively integrates sample-specific perspectives into a coherent analysis, while the critic agent evaluates the resulting reasoning and provides targeted natural-language feedback for directed revision. Finally, we develop a *reciprocal* training framework that jointly optimizes the proposal agent's drafting and feedback-guided revision via dual-stage reinforcement learning, while refining the critic agent according to the actual effectiveness of its feedback. Extensive experiments on three benchmarks demonstrate that ProCrit consistently outperforms reasoning-based methods, and human evaluation confirms substantially better-grounded rationales.

1 INTRODUCTION

Multimodal sarcasm detection, determining whether an image-text pair conveys sarcastic intent, has attracted growing interest due to its importance in content moderation, sentiment analysis, and emotionally aware dialogue systems (Joshi et al., 2017; Abulaish et al., 2020; Chen et al., 2024; Gao et al., 2025). Sarcasm fundamentally relies on incongruities between literal expression and intended meaning, and in the multimodal setting these incongruities often span modalities, *e.g.,* a caption reading "what a perfect day" paired with a photo of a traffic jam in a downpour.

Detecting such incongruities is challenging because sarcasm arises from diverse forms—verbal irony, hyperbole, rhetorical questions, cultural allusion, among others. Exposing the underlying incongruity requires examining the sample from multiple *analytical perspectives*—distinct angles of inspection such as assessing attitude contrasts, probing a cultural reference, or evaluating literal claims against visual evidence (examples in Fig. 4, Appendix B). The optimal perspective combination is intrinsically linked to the varied presentations of sarcasm and cue distributions inherent in each text-image sample.

To address this challenge, early works (Cai et al., 2019; Liang et al., 2022; 2024a; Guo et al., 2025b) have explored advanced cross-modal fusion strategies through attention, graph-based, and contrastive learning mechanisms, in an effort to capture the diverse signals that convey sarcastic intent. Recent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

methods Wang et al. (2025a); Zhang et al. (2025c); Ramakrishnan et al. (2025) explicitly model the analytical reasoning process by prompting Large Language Models (LLMs) or Vision–Language Models (VLMs) with a fixed, predefined set of analytical perspectives, each operating independently under hand-crafted routing rules. Although promising, these methods overlook that the analytical perspectives needed are inherently sample-specific, and thus the fixed, predefined set can hardly cover the diverse reasoning demands across samples; and that effective multi-perspective analysis should be progressive: the exploration of each subsequent perspective should be informed by insights gleaned from prior analytical steps, as illustrated in Figure 1. This calls for *self-elicited multi-perspective reasoning* in which the model autonomously explores and integrates different analytical perspectives and progressively builds a sample-adaptive, coherent analysis.

In this work, we introduce self-elicited multi-perspective reasoning to multimodal sarcasm detection. This paradigm empowers the model to dynamically generate sample-specific perspectives and progressively synthesize them into a coherent analysis, where each perspective is grounded in evidence from preceding steps. Nevertheless, achieving this goal raises two considerations. ❶ **Absence of process-level supervision.** Internalizing the self-elicited multi-perspective reasoning capability in a single model requires training with explicit supervision over the reasoning process, yet existing sarcasm datasets (Qin et al., 2023; Desai et al., 2022) provide only binary labels or brief explanation of the intended meaning, with no annotations capturing the underlying analytical process. ❷ **Reasoning reliability under subtle cues.** Given that sarcastic intent heavily relies on subtle, implicit, and cross-modal discrepancies, models employing self-elicited multi-perspective reasoning remain vulnerable to overlooking critical nuances, misaligning subtle cross-modal cues, or failing to resolve underlying incongruities, ultimately leading to incomplete or flawed reasoning chains.

In view of these, we present **ProCrit**, a **Pro**posal–**Crit**ic two-agent framework comprising a *proposal agent* that performs multi-perspective reasoning and a *critic agent* that provides external evaluation and targeted revision guidance, building on three components. ❶ **Process-level reasoning annotation synthesis**. To provide process-level training supervision, we synthesize reasoning annotations through a *dynamic-role agentic rollout*, where the strong vision-language model sequentially spawns analytical roles, each contributing a distinct perspective within a shared context. Each role has full visibility of all preceding analyses and autonomously determines what perspective is still needed and when analysis is sufficient. The resulting trajectories are flattened into unified reasoning sequences that preserve cross-perspective dependencies, providing process-level supervision for a single proposal agent. ❷ **Draft–critique–revise paradigm**. To improve the reliability of self-elicited multi-perspective reasoning, we introduce a *draft–critique–revise* paradigm: the proposal agent *drafts* a multi-perspective analysis, an independent *critic* provides targeted natural-language feedback on overlooked perspectives or misinterpreted cues, and the proposal agent revises its analysis under this guidance. This paradigm closes the reliability loop by anchoring self-elicited reasoning in independent critique, transforming open-ended generation into an externally evaluated analytical process refined through targeted natural-language guidance. ❸ **Reciprocal training strategy**. We further develop a *reciprocal* training strategy where the two agents are optimized in turn through reciprocal training signals: the proposal agent's drafting and revision capabilities are jointly optimized via dual-stage reinforcement learning using the critic's feedback as the training signal, establishing reciprocal supervision where the critic agent is in turn refined based on the effectiveness of its feedback in guiding revisions.

Our contributions are summarized as follows:

- We introduce **ProCrit**, a proposal–critic framework that reformulates multimodal sarcasm detection from manually predefined perspectives to *self-elicited multi-perspective reasoning*, in which analytical perspectives are generated adaptively and integrated progressively.
- We design a *dynamic-role agentic rollout* that generates sample-adaptive,unified reasoning trajectories in which each analytical perspective builds on preceding analyses, providing process-level supervision that enables the model to learn coherent multi-perspective reasoning.
- We propose a *draft–critique–revise* paradigm with *reciprocal training*, jointly optimizing drafting and feedback-guided revision and refining the critic through revision outcomes.

2 Related Work

Multimodal sarcasm detection. Early approaches model cross-modal incongruity through feature fusion (Cai et al., 2019; Liang et al., 2022; 2024a; Yue et al., 2023), while later methods incorporate

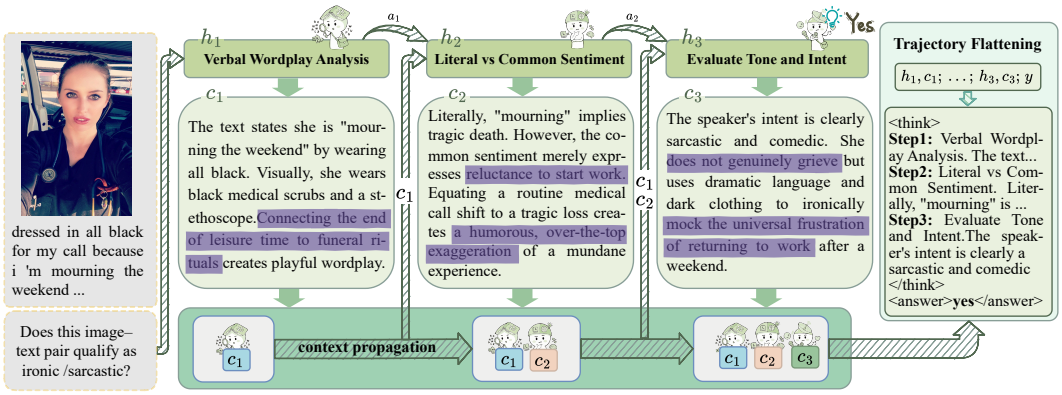


Figure 1: Illustration of the reasoning annotation synthesis process.

LLM-generated information and demonstration retrieval (Tang et al., 2024; Zhou et al., 2025). CaRGo-T (Nandy et al., 2026) prompts VLMs to generate causal reasoning graphs as code before answering multimodal queries. Perspective-based approaches make the analysis explicit: S3 Agent (Wang et al., 2025a) uses three predefined perspectives, Commander-GPT (Zhang et al., 2025c) dynamically selects from predefined subtasks, and IRONIC (Ramakrishnan et al., 2025) conditions interpretation on predefined discourse relations. Although these methods organize reasoning around complementary cues, their analytical categories remain specified in advance. ProCrit uses synthesized process-level supervision to learn *self-elicited multi-perspective reasoning*: perspectives are generated from each sample and preceding analyses, then progressively integrated into a single reasoning sequence.

Reasoning and refinement. Chain-of-thought prompting (Wei et al., 2022) and Self-Refine (Madaan et al., 2023) make reasoning and iterative refinement explicit, but self-correction can be unreliable without external feedback (Huang et al., 2023; Kamoi et al., 2024). Critic-V (Zhang et al., 2025a) trains a multimodal critic through preference optimization to provide natural-language guidance, while RL Tango (Zha et al., 2025) interleaves generator and verifier training, rewarding the verifier for correct judgments. ProCrit extends GRPO (Shao et al., 2024) to jointly optimize drafting and feedback-conditioned revision, and rewards the critic for corrections achieved by its feedback while penalizing harmful revisions. This ties feedback optimization to its downstream corrective effect. Appendix A provides an extended review.

3 Method

We develop **ProCrit**, a proposal–critic two-agent framework for multi-perspective reasoning in multimodal sarcasm detection, with three components: (1) *process-level reasoning annotation synthesis*, which uses a dynamic-role agentic rollout to generate analytical roles based on preceding analyses and remaining reasoning needs; (2) a *draft–critique–revise paradigm*, in which the proposal agent drafts an analysis, the critic evaluates it and provides targeted feedback, and the proposal agent revises accordingly; (3) a *mutual-refinement training strategy*, which alternates between jointly optimizing the proposal agent's drafting and revision using critic feedback and refining the critic using revision outcomes.

3.1 Process-Level Reasoning Annotation Synthesis

While existing sarcasm corpora (Qin et al., 2023; Desai et al., 2022) offer binary labels or brief explanations of the intended meaning, they lack process-level annotations that capture the reasoning steps or critical cues through which sarcastic intent is identified, as illustrated by the comparison in Figure 5. To obtain process-level supervision for coherent, sample-adaptive multi-perspective reasoning, we design a *dynamic-role agentic rollout* in which a strong vision-language model, denoted as $\mathcal{T}$, sequentially spawns analytical roles within a shared context, as illustrated in Figure 1.

Dynamic-role agentic rollout. Given an image-text pair (I, Q), the model $\mathcal{T}$ generates a reasoning trajectory by iteratively producing analytical roles, each contributing a distinct perspective. The t-th role yields a structured output $\mathbf{r}_t = (h_t, c_t, a_t)$, where h_t is its analytical perspective, c_t is

evidence-focused analysis, and a_t is a sufficiency judgment indicating whether the accumulated analyses suffice for a final prediction. Each role is conditioned on the full history of preceding roles:

$$(h_t, c_t, a_t) = \mathcal{T}\big(I,\, Q,\, [(h_1, c_1), \ldots, (h_{t-1}, c_{t-1})],\, a_{t-1}\big). \tag{1}$$

With full visibility of prior analyses, each successive role builds on established findings and identifies unaddressed aspects, progressively extending the reasoning trajectory by introducing a distinct perspective. When a_t indicates that the analyses are sufficient, the rollout terminates and $\mathcal{T}$ produces the final prediction $\hat{y}$, forming a trajectory $\tau = (\mathbf{r}_1, \ldots, \mathbf{r}_T, \hat{y})$. As a result, both the number of roles T and the perspectives h_t $(t = 1, ..., T)$ they cover are fully adaptive, governed by the sufficiency judgment a_t rather than manually designed role taxonomies or routing rules.

Trajectory flattening. The rollout yields a multi-turn trajectory τ in which every role is elicited by a separate prompt. Reproducing this protocol at inference would require an external controller to prompt one role at a time. We therefore flatten τ into a single sequence by concatenating all role outputs in order, followed by the final prediction:

$$\mathbf{s} = [\mathbf{r}_1;\, \mathbf{r}_2;\, \ldots;\, \mathbf{r}_T;\, \hat{y}] = [h_1, c_1;\, h_2, c_2;\, \ldots;\, h_T, c_T;\, \hat{y}], \tag{2}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes sequence concatenation. The flattened sequence $\mathbf{s}$ preserves the analytical progression and cross-perspective dependencies of trajectory τ, providing process-level supervision for coherent multi-perspective reasoning. This enables the full analysis to be learned and generated as a single autoregressive sequence, without requiring a separate prompt for each analytical step.

3.2 Draft–Critique–Revise Paradigm

With the process-level supervision above, the proposal agent is trained to generate sample-adaptive multi-perspective reasoning for a given sample (details in §3.3). However, since sarcastic intent often hinges on subtle and implicit cues, such single-pass reasoning may still overlook critical perspectives or misinterpret weak signals, and growing evidence shows that models cannot reliably detect flaws in their own reasoning (Huang et al., 2023). To improve reasoning reliability, we introduce an independent critic agent to decouple reasoning generation from reasoning evaluation, forming a *draft–critique–revise* paradigm in which the critic agent identifies specific reasoning deficiencies and provides targeted natural-language feedback for directed revision, as illustrated in Figure 2.

Draft. Given an image-text pair (I, Q), the proposal agent π_θ operates in *draft* mode and generates an initial multi-perspective reasoning $\bar{d}$ together with a binary prediction $\hat{y}_{\text{draft}}$ under a drafting prompt p_{draft}:

$$(\bar{d},\, \hat{y}_{\text{draft}}) = \pi_\theta(I,\, Q;\, p_{\text{draft}}). \tag{3}$$

The draft reasoning $\bar{d}$ follows the flattened dynamic-role format learned from the synthesized trajectories: a coherent sequence of sample-specific analytical perspectives within `<think>···</think>` tags, followed by the predicted sarcasm label in `<answer>···</answer>` tags.

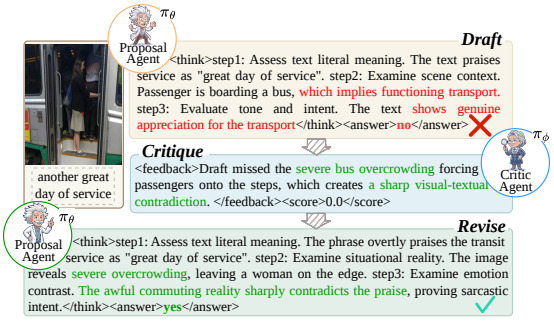


Figure 2: Illustration of the draft–critique–revise paradigm.

Critique. To further refine the initial draft reasoning $\bar{d}$, the critic agent π_ϕ evaluates it against the original multimodal evidence under a critic prompt p_{critic}:

$$(f,\, s) = \pi_\phi(I,\, Q,\, \bar{d};\, p_{\text{critic}}), \tag{4}$$

where f is a natural-language feedback within `<feedback>···</feedback>` tags that identifies specific reasoning deficiencies and provide targeted revision guidance, and $s \in \{0, 1, 2\}$ is a quality score within `<score>···</score>` tags, providing an estimate of overall reasoning quality.

Revise. Guided by the feedback f, the same proposal agent π_θ then switches to *revise* mode and regenerates the full reasoning conditioned on the draft and the feedback, rather than editing the draft locally, under a revision prompt p_{revise} (see Appendix C.3):

$$(d,\, \hat{y}_{\text{revise}}) = \pi_\theta(I,\, Q,\, \bar{d},\, f;\, p_{\text{revise}}). \tag{5}$$

The revised reasoning d follows the same structured format as the draft reasoning $\bar{d}$, and the revised output $(d, \hat{y}_{\text{revise}})$ is taken as the final result of the draft–critique–revise pipeline, while the draft output $(\bar{d}, \hat{y}_{\text{draft}})$ remains available as the first-pass result for training and analysis.

3.3 RECIPROCAL TRAINING

We train the two agents in turn, each against a frozen counterpart, so that each serves as the training signal for the other (*reciprocal training*): the critic supplies quality scores and targeted feedback as training signals for the proposal agent's dual-stage reinforcement learning (§3.3.2), and the proposal agent's revision outcomes reward the critic (§3.3.3).

3.3.1 INITIALIZATION

Both agents are initialized via Supervised Fine-Tuning (SFT) to establish basic capabilities, with training data constructed through an interleaved process. The proposal agent is first trained on the synthesized reasoning trajectories (§3.1) for sample-adaptive multi-perspective reasoning; the trained proposal agent then generates drafts on the training set, which are evaluated by the model $\mathcal{T}$ to produce training examples for the critic agent; the resulting feedback from the trained critic agent is in turn used to prompt $\mathcal{T}$ for revision, which serve as the proposal agent's SFT data for the revise mode. Implementation details are provided in §4.1.

3.3.2 PROPOSAL AGENT: DUAL-STAGE REINFORCEMENT LEARNING

The proposal agent is optimized through a dual-stage Group Relative Policy Optimization (GRPO) objective that jointly trains its drafting and revising capabilities. During training, the critic agent is frozen and deployed as an inference service.

Stage 1: Draft optimization. For each image-text pair (I, Q), the proposal agent generates G draft completions $\{d_1, \ldots, d_G\}$. The critic agent evaluates each draft d_i, producing the corresponding quality score s_i and natural-language feedback f_i. Each draft is assigned a composite reward:

$$r_{\text{draft}}(d_i) = r_{\text{acc}}(d_i) + r_{\text{fmt}}(d_i) + r_{\text{eval}}(d_i), \tag{6}$$

where $r_{\text{acc}} \in \{0, 1\}$ is a binary accuracy reward indicating whether the predicted label matches the ground truth, $r_{\text{fmt}} \in \{0, 1\}$ is a format reward verifying adherence to the required output format, $r_{\text{eval}}(d_i) = \tilde{s}_i$ is a critic evaluation reward given by the normalized quality score.

Stage 2: Feedback-guided revision optimization. From the G drafts, K parents are selected via balanced sampling to include both correct and incorrect drafts. For each selected parent d_k with critic feedback f_k, the proposal agent generates M revised completions $\{o_{k,1}, \ldots, o_{k,M}\}$ by rewriting the draft reasoning from scratch under the feedback. The revision reward is defined as:

$$r_{\text{revise}}(o_{k,j}) = r_{\text{acc}}(o_{k,j}) + r_{\text{fmt}}(o_{k,j}) + r_{\text{imp}}(o_{k,j},\, d_k), \tag{7}$$

where the improvement reward r_{imp} is designed to capture the marginal effect of revision relative to the draft and y^* denotes the gold label:

$$r_{\text{imp}}(o, d) = \begin{cases} +1 & \text{if } \hat{y}_{\text{draft}} \neq y^* \text{ and } \hat{y}_{\text{revise}} = y^* \quad \text{(fix)} \\ -1 & \text{if } \hat{y}_{\text{revise}} \neq \hat{y}_{\text{draft}} \text{ and } \hat{y}_{\text{revise}} \neq y^* \quad \text{(damage)} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Since correctness is already captured by r_{acc}, r_{imp} gives additional credit only to genuine corrections, penalizes harmful changes, and remains neutral for rewrites without a correctness gain.

Dual-stage GRPO. Rather than optimizing drafting and revision separately, we couple them into a single training rollout that produces a joint gradient update. Let $\mathcal{D} = \{d_i\}_{i=1}^{G}$ denote the draft candidate set and $\mathcal{O}_k = \{o_{k,j}\}_{j=1}^{M}$ denote the revision candidate set for parent draft d_k. Group-relative advantages $\hat{A}_{\text{draft}}$ and $\hat{A}_{\text{revise}}$ are computed within each set by standardizing rewards as in GRPO (Appendix C.2). Let $\Phi(\mathcal{C}, \hat{A})$ denote the GRPO clipped policy-ratio objective evaluated over a candidate set $\mathcal{C}$ with its group-relative advantages $\hat{A}$. The proposal agent is trained with a single dual-stage objective:

$$\mathcal{L}_{\text{proposal}} = (1 - \lambda)\Phi(\mathcal{D}, \hat{A}_{\text{draft}}) + \lambda \frac{1}{K} \sum_{k=1}^{K} \Phi(\mathcal{O}_k, \hat{A}_{\text{revise}}), \tag{9}$$

where $\lambda \in [0, 1]$ controls the relative emphasis on feedback-guided revision. Since both stages share parameters, revision training simultaneously strengthens drafting: learning to correct flaws under feedback helps the model avoid similar flaws in its initial drafts.

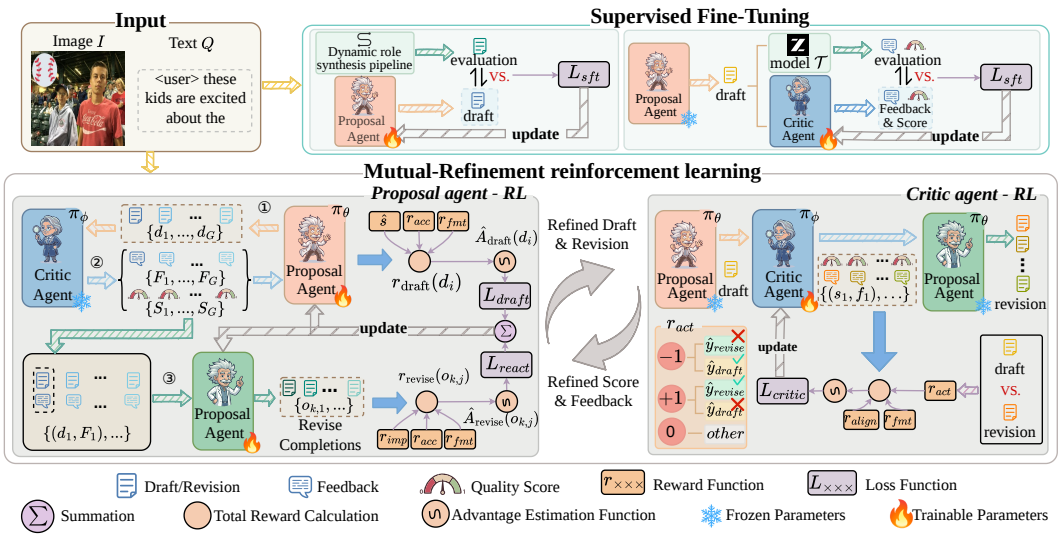


Figure 3: Overview of reciprocal training. **Top:** Both agents are initialized via SFT. **Bottom left:** The proposal is optimized in dual stages (draft and revision) while the critic is frozen. **Bottom right:** The critic is optimized while the proposal is frozen, using revision outcomes as rewards.

3.3.3 Critic agent: Reinforcement Learning

After supervised fine-tuning, the critic agent is further optimized via GRPO to improve its scoring calibration and feedback utility. For each training sample containing an image-text pair (I, Q) and a draft reasoning, G completions are sampled and rewarded by the composite reward $r_{\text{critic}} = r_{\text{fmt}} + r_{\text{align}} + r_{\text{act}}$. The three components are defined as follows.

Format reward r_{fmt}. A binary reward verifying that the output contains valid `<feedback>···</feedback>` and `<score>···</score>` tags with non-empty content.

Score alignment reward r_{align}. Let s denote the critic's predicted score, $\hat{y}$ the proposal's answer:

$$r_{\text{align}} = \begin{cases} s - 1 & \text{if } \hat{y} = y^* \\ 1 - s & \text{if } \hat{y} \neq y^*. \end{cases} \tag{10}$$

This rewards the critic for assigning high scores to correct proposals and low scores to incorrect ones. r_{align} grounds the score in verifiable correctness, a prerequisite for directing revision, while the process-level criteria of the rubric (Prompt 6) are instilled through SFT on teacher critiques.

Feedback actionability reward r_{act}. The critic's feedback is sent to a frozen proposal agent deployed as a separate inference service, which generates a corresponding revision.

$$r_{\text{act}} = \begin{cases} +1 & \text{if } \hat{y}_{\text{draft}} \neq y^* \text{ and } \hat{y}_{\text{revise}} = y^* \quad \text{(fix)} \\ -1 & \text{if } \hat{y}_{\text{revise}} \neq \hat{y}_{\text{draft}} \text{ and } \hat{y}_{\text{revise}} \neq y^* \quad \text{(damage)} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

We design r_{act} to reward feedback only when it provides verifiable corrective value. Feedback without observable improvement is therefore treated as neutral, preventing the critic agent from learning generic or unnecessary feedback patterns.

3.3.4 Training Pipeline

Both agents are first initialized via SFT (§3.3.1), then refined through alternating reinforcement learning stages. We first optimize the critic agent with the proposal agent frozen after SFT, and then optimize the proposal agent with the critic agent frozen after its RL stage. At each stage, one agent is frozen while the other is optimized using training signals produced by its counterpart, avoiding the co-adaptation instabilities of simultaneous training.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on two in-domain benchmarks and one out-of-domain benchmark: (1) **MMSD** (Cai et al., 2019), the first large-scale image-text sarcasm detection dataset collected from Twitter, containing $24,317$ samples. Its test set comprises 969 sarcastic and $1,404$ non-sarcastic examples, respectively; (2) **MMSD2.0** (Qin et al., 2023), a cleaned version of MMSD that removes the spurious cues such as hashtags and re-annotates the unreasonable samples, serving as the current standard benchmark with the same scale but more reliable labels. Its test set contains $1,037$ sarcastic and $1,372$ non-sarcastic examples, respectively; (3) **RedEval** (Tang et al., 2024), constructed for out-of-domain evaluation of models trained on Twitter-based datasets, and samples are sourced from Reddit, with 395 sarcastic and 609 non-sarcastic examples, respectively.

Evaluation metrics. Following prior works (Qin et al., 2023; Wang et al., 2025a), we report F1 score (**F1**), accuracy (**Acc**), precision (**P**), and recall (**R**), with F1 as the primary metric.

Implementation details. Both the proposal and critic agents are initialized from Qwen2.5-VL-7B-Instruct (Bai et al., 2025), with the vision encoder and multimodal projector frozen and only the language model parameters updated. For *reasoning annotations generation*, we use GLM-4.6V (Glm et al., 2024) as the strong vision-language model $\mathcal{T}$. For *supervised fine-tuning*, the proposal agent is fine-tuned on 9K generated reasoning sequences with a global batch size of 64 and a learning rate of 1×10^{-5} for 3 epochs; the critic agent is fine-tuned on 20K examples sourced from the proposal agent's rollouts, annotated by $\mathcal{T}$. For *reinforcement learning*, the two agents are optimized in turn, with the frozen counterpart deployed via vLLM (Kwon et al., 2023) to supply training signals at each stage. The proposal agent is optimized using 2K instances with G=8, M=4, λ=0.5, and learning rate 3×10^{-6} for 4 epochs. The critic agent is optimized using 5K instances with G=8 completions and learning rate 1×10^{-6} for 4 epochs. All training is conducted on 8 NVIDIA A800 GPUs with DeepSpeed ZeRO-3 (Rajbhandari et al., 2020) to reduce memory consumption. More hyperparameter details are provided in Appendix C.

Evaluation Settings. We compare methods along two axes: training regime (*zero-shot* vs. *fine-tuned*) and reasoning perspective setting (*None*, *Predefined*, and *Adaptive*). (1) *None* denotes reasoning without explicitly specified analytical perspectives, including Zero-shot CoT (Kojima et al., 2022), Automatic Prompt Engineer (Zhou et al., 2022), Plan-and-Solve (Wang et al., 2023), Generated Knowledge Prompting (Liu et al., 2022), and CaRGo-T (Nandy et al., 2026), which generates a code-based causal reasoning graph before predicting the label. We also implement Plain CoT Reasoner, a fine-tuned baseline trained to produce generic chain-of-thought reasoning. (2) *Predefined* denotes reasoning with a fixed set of manually specified perspectives, including IRONIC (Ramakrishnan et al., 2025), S3Agent (Wang et al., 2025a), and Commander-GPT (Zhang et al., 2025c). We additionally implement Fixed-role Reasoner, a fine-tuned baseline using S3Agent's three predefined perspectives. (3) *Adaptive* denotes sample-adaptive perspective generation, where analytical perspectives are self-elicited per instance, as used by ProCrit in both zero-shot and fine-tuned settings. All zero-shot baselines are reproduced with the same Qwen2.5-VL-7B backbone. All fine-tuned variants employ the same backbone and training protocol to ensure a fair comparison.

4.2 MAIN RESULTS

As shown in Table 1, ProCrit achieves the strongest overall performance across all three benchmarks under both zero-shot and fine-tuned settings. First, adaptive perspective generation consistently outperforms both none and predefined alternatives. In the zero-shot regime, ProCrit attains the highest F1 score on all three benchmarks, surpassing methods that use no perspectives or predefined ones. This gain stems from its ability to select sample-specific analytical perspectives, leading to more balanced precision–recall trade-offs. The same trend holds under fine-tuning: the Fixed-Role Reasoner (79.8 F1) slightly outperforms the Plain CoT Reasoner (79.4 F1) on MMSD2.0, confirming that even predefined multi-perspective reasoning offers a modest advantage over unstructured chain-of-thought. However, ProCrit†, which employs adaptive perspectives, further improves F1 to 81.1, demonstrating that sample-adaptive perspective generation yields superior reasoning compared to both plain CoT and fixed-role decomposition. Second, critic-guided revision further improves performance beyond adaptive reasoning alone. ProCrit consistently outperforms ProCrit†, indicating that targeted feedback from the critic agent (ProCrit) enhances overall reasoning quality—not merely by adjusting

Table 1: Main results on MMSD2.0, MMSD, and redEval. The *Perspective* column indicates the reasoning paradigm: no explicit perspectives (None), predefined perspectives (Predefined), or sample-adaptive perspectives (Adaptive). The best and second-best results are in **bold** and underlined, respectively. ★ marks the best result within the zero-shot group. † denotes the draft-only output.

Method	Perspective	MMSD2.0				MMSD				redEval			
		F1	Acc	P	R	F1	Acc	P	R	F1	Acc	P	R
Zero-shot methods													
Zero-shot CoT (Kojima et al., 2022)	None	61.6	71.1	71.9	53.9	63.2	73.2	68.1	59.0	70.6	78.2★	75.1	66.6
Automatic Prompt Engineer (Zhou et al., 2022)	None	61.7	71.4	72.9	53.4	61.6	72.9	68.9	55.8	70.2	78.2★	75.9	65.3
Plan-and-Solve (Wang et al., 2023)	None	54.1	69.0	74.4	42.5	57.1	71.9	70.6	48.0	58.9	72.0	77.4	47.6
Generated Knowledge Prompting (Liu et al., 2022)	None	56.4	70.2	76.1★	44.7	57.8	72.6	**72.2**★	48.1	61.5	75.4	80.1	49.9
CaRGo-T (Nandy et al., 2026)	None	51.2	61.6	65.8	56.2	52.6	64.5	65.5	56.3	63.0	70.2	73.5	63.7
IRONIC (Ramakrishnan et al., 2025)	Predefined	40.8	64.5	72.5	28.4	41.8	66.8	71.4	29.6	59.1	76.0	**89.7**★	44.1
S3Agent (Wang et al., 2025a)	Predefined	61.2	70.7	71.2	53.6	63.7	73.9★	69.6	58.7	61.0	73.8	73.6	52.2
Commander-GPT (Zhang et al., 2025c)	Predefined	63.3	65.8	58.8	68.6★	61.6	66.0	55.0	70.1★	70.1	72.3	60.9	82.5★
ProCrit	Adaptive	67.4★	71.6★	66.7	68.2	66.7★	70.9	64.5	69.1	71.4★	76.8	69.2	73.6
Fine-tuned methods													
Plain CoT Reasoner (Kojima et al., 2022)	None	79.4	80.3	72.3	88.0	77.5	79.4	67.5	90.8	82.6	84.8	74.9	92.2
Fixed-role Reasoner (Wang et al., 2025a)	Predefined	79.8	80.8	73.1	87.9	78.0	80.1	68.4	90.8	81.6	83.9	74.3	90.6
ProCrit†	Adaptive	81.1	82.5	75.7	87.4	78.4	81.0	70.4	88.4	**84.8**	**87.2**	79.2	91.4
ProCrit	Adaptive	**83.1**	**84.0**	**76.1**	**91.5**	**80.7**	**82.6**	71.2	**93.2**	83.7	85.7	75.7	**93.7**

the precision–recall balance; notably, recall improves substantially: on MMSD2.0, recall rises from 87.4 to 91.5, and on MMSD from 88.4 to 93.2, suggesting that the critic in (ProCrit) helps recover sarcastic instances whose decisive cues are missed or under-analyzed in the initial proposal (ProCrit†). On redEval, ProCrit† achieves the highest F1 (84.8), while ProCrit attains the best recall (93.7). Given that redEval contains a higher proportion of non-sarcastic examples (class ratio $\approx$ 3:2), the additional true positives recovered through revision are partially offset by a drop in precision, resulting in a marginally lower F1. Nevertheless, the consistent recall improvement across all three benchmarks confirms that critic-guided revision (ProCrit) enables the model to capture subtle sarcastic signals overlooked by draft-only reasoning (ProCrit†). Additional comparisons with strong task-specific methods in Appendix D.3 highlight ProCrit's out-of-domain advantage: its Draft mode outperforms the compared methods on RedEval while providing explicit multi-perspective rationales.

4.3 Ablation Study

We conduct systematic ablations on MMSD2.0 and MMSD to isolate the contribution of each component; Appendix D.5 reports the complete suite on the OOD RedEval benchmark. All ablations use Qwen2.5-VL-7B as the backbone.

Ablation on Training Stages. Table 2 analyzes the contribution of each training stage in the proposal agent. The full pipeline combining SFT and RL achieves the highest F1 in both Draft and Revise modes on MMSD2.0 and MMSD. Removing either stage degrades performance in both generation modes. The largest drop without SFT underscores its critical role in establishing the proposal agent's sample-adaptive multi-perspective reasoning, while RL plays a complementary role in refining reasoning quality through feedback-driven optimization.

Table 2: Ablation on training stages of the proposal agent. Draft and Revise denote the two generation modes.

SFT	RL	MMSD2.0 (F1)		MMSD (F1)	
		Draft	Revise	Draft	Revise
✗	✓	77.3	81.7	77.0	80.4
✓	✗	78.1	82.3	75.9	80.2
✓	✓	**81.1**	**83.1**	**78.4**	**80.7**

Table 3: Ablation on the Revise mode. The proposal agent is held fixed across all experiments.

Revision Setting	MMSD2.0		MMSD	
	F1	Acc	F1	Acc
None	81.1	82.5	78.4	81.0
Self-revision	75.9	79.0	77.1	80.2
Initialized critic	81.5	82.9	79.4	81.9
Proposal-refined critic	**83.1**	**84.0**	**80.7**	**82.6**

Ablation on Rewards and Balance Weight. We ablate the reward components and the draft–revise balance weight λ in Appendix D.4. For the proposal agent, removing r_{eval} reduces Draft F1 from

Table 4: Human evaluation of rationale quality (left) and perspective quality (right). EG: evidence grounding; LC: logical coherence; RLC: rationale–label consistency; Cue: core-cue coverage; Hall.: hallucination severity; Rel.: relevance; Unique: unique contribution; Redund.: redundancy; Unsupp.: unsupported content; Neces.: necessity for conclusion.

Method	EG	LC	RLC	Cue	Hall.↓	Overall
Plain CoT	1.84	1.93	3.72	1.84	2.56	1.56
Fixed-role	1.78	1.96	3.86	1.80	2.88	1.50
ProCrit Draft	**2.96**	**2.89**	4.09	2.62	0.95	2.47
ProCrit Revise	2.89	2.72	**4.16**	**2.66**	**0.84**	**2.63**
Teacher rationale	4.06	4.03	4.72	3.73	0.42	3.72

Method	Rel.	Unique	Redund.↓	Unsupp.↓	Neces.
Plain CoT	–	–	–	–	–
Fixed-role	3.34	4.00	0.91	2.82	3.77
ProCrit Draft	4.04	4.37	**0.52**	**1.51**	4.07
ProCrit Revise	**4.17**	**4.52**	**0.52**	1.80	**4.42**
Teacher rationale	4.53	4.61	0.55	0.65	4.52

81.1 to 79.9 on MMSD2.0, while removing r_{imp} reduces Revise F1 from 83.1 to 82.6, supporting their contributions to drafting and revision, respectively. For the critic agent, ablating either r_{align} or r_{act} also lowers Revise F1. For the balance weight, joint optimization (λ=0.5) outperforms training either mode alone on both in-domain benchmarks.

Ablation on Revise Mode. Table 3 compares four revise modes with the proposal agent fixed. Self-revision re-generates the proposal without external feedback and underperforms the initial draft (None), indicating that a second pass alone does not ensure improvement. In contrast, incorporating external critique from an independent critic agent consistently improves performance over both baselines. The initialized critic is trained only by supervised distillation from teacher critiques and never receives a label-based reward, yet it already improves over the draft (81.5 vs. 81.1; 79.4 vs. 78.4); since an uninformed second pass hurts, the gain stems from the content of the feedback. Further improvement is achieved through reciprocal training, where the proposal-refined critic attains the best results. This supports revision-outcome-grounded critic optimization: outcomes from the frozen proposal provide signals that move the critic beyond supervised imitation toward more actionable guidance. A single revision round captures the dominant improvement (Appendix D.8).

4.4 Qualitative Analysis

Qualitative examples in Appendix E illustrate the effects of RL on cue interpretation and the benefits of critic-guided revision over self-revision. They further show how feedback corrects prediction errors and prompts rationale revision even when the initial labels are already correct.

4.5 Reasoning-Level Evaluation

Five trained graduate annotators score anonymized outputs on a 0–5 scale. We evaluate four systems on 100 label-balanced MMSD2.0 test examples and separately audit 100 label-balanced trajectories from the GLM-4.6V synthesized supervision. Table 4 reports mean scores; Appendix D.1 provides the full protocol and sample standard deviations. Both ProCrit modes outperform the baselines across all rated dimensions, particularly in evidence grounding and hallucination, while their perspectives are more relevant, less redundant, and more necessary than predefined roles. Revision improves overall ratings but slightly reduces grounding and coherence, consistent with its recall gains (Table 1). The GLM-4.6V audit supports annotation reliability, with consistent agreement across dimensions (ordinal Krippendorff's $\alpha \geq 0.68$; within-one-point agreement $\geq 82\%$).

5 Conclusion

This paper presents ProCrit, a two-agent framework that shifts multimodal sarcasm detection from relying on manually prescribed analytical perspectives to self-elicited multi-perspective reasoning. To obtain process-level supervision for this capability, we design a dynamic-role agentic rollout that synthesizes sample-adaptive reasoning annotations where each analytical perspective progressively builds on preceding analyses. To improve reasoning reliability, ProCrit executes external critique through feedback-conditioned revision and grounds critic optimization in the observed revision outcomes. Extensive experiments on three benchmarks, together with human evaluation and cross-backbone and cross-task studies, demonstrate the effectiveness of ProCrit and the quality of its rationales. We provide further discussion on limitations and broader impact in the Appendix F and G.

AI USE STATEMENT

We used GLM-4.6V to generate synthetic training annotations for ProCrit, including reasoning trajectories, teacher critiques, and feedback-guided revisions, as described in Section 3 and Appendix B. A human audit of sampled synthetic reasoning trajectories is reported in Appendix D.1. We also used generative AI tools to improve the clarity and fluency of the manuscript. The authors take responsibility for all content of this manuscript.

REFERENCES

Muhammad Abulaish, Ashraf Kamal, and Mohammed J Zaki. A survey of figurative language and its computational detection in online social networks. *ACM Transactions on the Web (TWEB)*, 14 (1):1–52, 2020.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2506–2515, 2019.

Wangqun Chen, Fuqiang Lin, Guowei Li, and Bo Liu. A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. *Neurocomputing*, 578:127428, 2024.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10563–10571, 2022.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

Xiyuan Gao, Shekhar Nayak, and Matt Coler. Spoken in jest, detected in earnest: A systematic review of sarcasm recognition-multimodal fusion, challenges, and future prospects. *IEEE Transactions on Affective Computing*, 2025.

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

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

Diandian Guo, Cong Cao, Fangfang Yuan, Yanbing Liu, Guangjie Zeng, Xiaoyan Yu, Hao Peng, and Philip S Yu. Multi-view incongruity learning for multimodal sarcasm detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1754–1766, 2025b.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Soumyadeep Jana and Ranbir Singh Sanasam. Teaching sarcasm: Few-shot multimodal sarcasm detection via distillation to a parameter-efficient student. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1338–1349, 2025.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22, 2017.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1767–1777, 2022.

Bin Liang, Lin Gui, Yulan He, Erik Cambria, and Ruifeng Xu. Fusion and discrimination: A multimodal graph contrastive learning framework for multimodal sarcasm detection. *IEEE Transactions on Affective Computing*, 15(4):1874–1888, 2024a.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3154–3169, 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Abhilash Nandy, Rahul Seetharaman, Aman Bansal, Rounak Saha, Manav Nitin Kapadnis, Millon Madhur Das, Pawan Goyal, and Niloy Ganguly. Cargo-t: Causal reasoning graph-of-thought improves multimodal humor comprehension. *arXiv preprint arXiv:2608.23172*, 2026.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. Mmsd2. 0: Towards a reliable multi-modal sarcasm detection system. In *Findings of the association for computational linguistics: ACL 2023*, pp. 10834–10845, 2023.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

Aashish Anantha Ramakrishnan, Aadarsh Anantha Ramakrishnan, and Dongwon Lee. Ironic: Coherence-aware reasoning chains for multi-modal sarcasm detection. *arXiv preprint arXiv:2505.16258*, 2025.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1732–1742, 2024.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 2609–2634, 2023.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Peng Wang, Yongheng Zhang, Hao Fei, Qiguang Chen, Yukai Wang, Jiasheng Si, Wenpeng Lu, Min Li, and Libo Qin. S3 agent: Unlocking the power of vllm for zero-shot multi-modal sarcasm detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21 (11):1–16, 2025a.

Xinyu Wang, Yue Zhang, and Liqiang Jing. Can large vision-language models understand multimodal sarcasm? In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5340–5345, 2025b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. Knowlenet: Knowledge fusion network for multimodal sarcasm detection. *Information Fusion*, 100:101921, 2023.

Kaiwen Zha, Zhengqi Gao, Maohao Shen, Zhang-Wei Hong, Duane Boning, and Dina Katabi. Rl tango: Reinforcing generator and verifier together for language reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38,

Main Conference, pp. 119283–119313. Curran Associates, Inc., 2025. doi: 10.52202/085713-3981. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ad06e23fe0c39b9de6e0cefe3b701f45-Paper-Conference.pdf.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9050–9061, 2025a.

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *IEEE Transactions on Affective Computing*, 2025b.

Yazhou Zhang, Chunwang Zou, Bo Wang, Jing Qin, and Prayag Tiwari. Commander-gpt: Dividing and routing for multimodal sarcasm detection. *arXiv preprint arXiv:2506.19420*, 2025c.

Hengyang Zhou, Jinwu Yan, Yaqing Chen, Rongman Hong, Wenbo Zuo, and Keyan Jin. Ldgnet: Llms debate-guided network for multimodal sarcasm detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022.

APPENDIX CONTENTS

A Related Work **15**

A.1 Multimodal Sarcasm Detection 15

A.2 Reasoning and Refinement 15

A.3 Reinforcement Learning for LLM Reasoning 15

B Process-Level Reasoning Annotation Synthesis **15**

B.1 Dynamic-Role Rollout Details 15

B.2 Synthesized Data Analysis 16

B.3 Synthesis Examples 17

C Training Details **17**

C.1 Hyperparameters 17

C.2 Advantage Estimation 17

C.3 Prompts 18

D Additional Experiments **20**

D.1 Human Evaluation Protocol and Full Results 21

D.2 Generalization Across Backbones and Prompting Regimes 21

D.3 Task-Specific and Proprietary Baselines 22

D.4 Reward and Balance-Weight Ablations 22

D.5 Complete RedEval Ablations 23

D.6 Optimization Order and Training Stability 24

D.7 Cross-Task Transfer to Hateful Memes 24

D.8 Additional Ablations 25

D.9 Efficiency Analysis 25

E Qualitative Analysis **26**

F Limitations **26**

G Broader Impact **31**

A Related Work

A.1 Multimodal Sarcasm Detection

Early multimodal sarcasm detection methods learn cross-modal representations through attention (Cai et al., 2019), graph convolution (Liang et al., 2022), contrastive learning (Liang et al., 2024a), and external knowledge (Yue et al., 2023). Recent approaches incorporate LLMs through demonstration retrieval (Tang et al., 2024) and debate-guided text generation (Zhou et al., 2025). CaRGo-T (Nandy et al., 2026) prompts VLMs to generate causal reasoning graphs as code before producing answers for multimodal humor understanding and detection, including sarcasm detection. It supports zero-shot prompting and in-context learning without updating model parameters. Perspective-based methods organize analysis around predefined categories: S3 Agent (Wang et al., 2025a) examines expression, semantics, and sentiment; Commander-GPT (Zhang et al., 2025c) dynamically selects from six specialized subtasks; and IRONIC (Ramakrishnan et al., 2025) selects discourse relations to guide interpretation. ProCrit instead learns *self-elicited multi-perspective reasoning* from synthesized process-level supervision, generating sample-specific perspectives conditioned on preceding analyses and integrating them into a single reasoning sequence.

A.2 Reasoning and Refinement

Chain-of-thought prompting (Wei et al., 2022), Tree-of-Thoughts (Yao et al., 2023), self-consistency (Wang et al., 2022), and multi-agent debate (Du et al., 2024; Liang et al., 2024b) organize reasoning through intermediate steps, alternative paths, or exchanges among agents. Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) use feedback to refine outputs or guide subsequent attempts, although self-correction from self-generated feedback can be unreliable (Huang et al., 2023; Kamoi et al., 2024). Process reward models (Lightman et al., 2023) and Math-Shepherd (Wang et al., 2024) assess intermediate-step correctness, while Critic-V (Zhang et al., 2025a) uses a preference-trained critic to provide natural-language revision guidance. ProCrit couples external evaluation with targeted feedback to address overlooked perspectives and misinterpreted cross-modal cues in multi-perspective sarcasm reasoning.

A.3 Reinforcement Learning for LLM Reasoning

Reinforcement learning has advanced reasoning in language models (Guo et al., 2025a; Hu et al., 2025) and vision-language models (Shen et al., 2025; Huang et al., 2025). SCoRe (Kumar et al., 2024) trains multi-turn self-correction, while RL Tango (Zha et al., 2025) alternates generator and verifier optimization, rewarding the verifier for correct judgments. ProCrit extends GRPO (Shao et al., 2024) to jointly optimize drafting and critic-guided revision (§3.3.2), while rewarding its critic for feedback that corrects predictions and penalizing feedback that damages them (§3.3.3).

B Process-Level Reasoning Annotation Synthesis

This section details the process-level reasoning annotation synthesis pipeline underlying ProCrit. As different sarcastic samples rely on distinct forms and require different analytical perspectives to expose the underlying incongruity (see Fig. 4), our synthesis pipeline generates sample-adaptive, multi-perspective reasoning annotations. We describe the dynamic-role rollout protocol (§B.1), present quantitative analyses of the synthesized data (§B.2), and provide concrete synthesis examples (§B.3).

B.1 Dynamic-Role Rollout Details

While existing sarcasm corpora (Qin et al., 2023; Desai et al., 2022) offer binary labels or brief explanations of the intended meaning, they lack process-level annotations that capture the reasoning steps or critical cues through which sarcastic intent is identified. As illustrated in Figure 5, the dynamic-role agentic rollout (§3.1) addresses this by synthesizing multi-perspective, context-aware reasoning annotations from raw image-text samples. Specifically, we use GLM-4.6V (Glm et al., 2024) as the teacher model $\mathcal{T}$ and draw source samples from the MMSD2.0 training set. Trajectories whose final answers are correct are directly flattened into single-turn draft training sequences. For

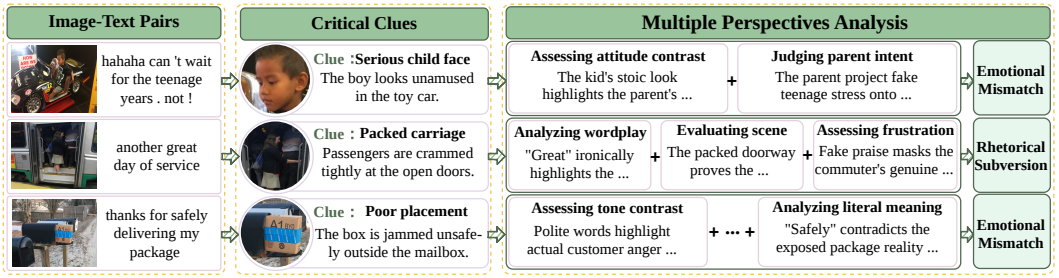


Figure 4: Examples illustrating that different sarcastic samples involve distinct forms and demand different combinations of analytical perspectives to expose the text–image incongruity.

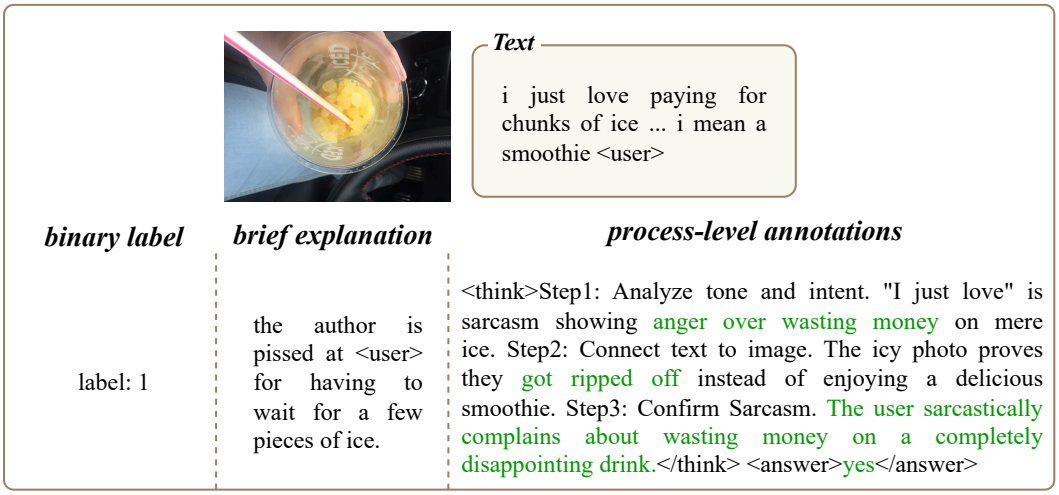


Figure 5: Comparison between existing dataset annotations and the process-level reasoning annotations synthesized by ProCrit.

trajectories that produce incorrect answers, we collect the flawed reasoning. After the critic agent is initialized by SFT using teacher-annotated draft evaluations, we use this SFT critic to generate diagnostic feedback (Prompt 6) for trajectories with incorrect final answers. The feedback is then provided to the teacher model $\mathcal{T}$ to regenerate corrected analyses (Prompt 7). This produces draft–feedback–revision triples that serve as multi-turn training data for the proposal agent's revision capability. The resulting training dataset contains 9,000 sequences: 6,780 single-turn draft sequences (from initially correct trajectories) and 2,220 draft–revision pairs (from critic-guided correction of initially incorrect trajectories).

B.2 Synthesized Data Analysis

We analyze the raw synthesized reasoning trajectories (15,042 valid samples before filtering) to characterize the data produced by the dynamic-role rollout.

Step count distribution. Table 5 reports the distribution of reasoning steps across all synthesized trajectories. The majority of samples (73.9%) use 3 analytical steps, while 14.8% require 4 steps and 10.9% use 2 steps. Only 0.3% of samples use 5 steps. Notably, sarcastic samples tend to require more reasoning steps on average (3.12 vs. 2.95 for non-sarcastic), reflecting the greater analytical complexity involved in identifying cross-modal incongruities. Sarcastic samples also exhibit a higher proportion of 4-step trajectories (18.3% vs. 10.6%), suggesting that uncovering sarcastic intent often demands additional analytical perspectives beyond the baseline.

Table 5: Reasoning step count distribution across all synthesized trajectories, broken down by sarcasm label.

	2-step	3-step	4-step	5-step	Avg.
Sarcastic	578 (7.1%)	6,027 (74.1%)	1,491 (18.3%)	38 (0.5%)	3.12
Non-sarcastic	1,069 (15.5%)	5,096 (73.8%)	733 (10.6%)	10 (0.1%)	2.95
All	1,647 (10.9%)	11,123 (73.9%)	2,224 (14.8%)	48 (0.3%)	3.05

Label distribution. Among the 15,042 raw synthesized trajectories, 8,134 (54.1%) are labeled as sarcastic and 6,908 (45.9%) as non-sarcastic, reflecting a roughly balanced distribution consistent with the source MMSD2.0 training set.

Human audit of synthesis quality. To complement these structural statistics, we conduct a separate human audit of the synthesized supervision. The results indicate that the trajectories are generally grounded, coherent, and label-consistent with low hallucination, while their constituent perspectives are relevant and non-redundant; Appendix D.1 reports the full protocol, dimension-level results, and agreement statistics.

B.3 Synthesis Examples

Figures 6 present concrete examples of the process-level reasoning annotations produced by the dynamic-role agentic rollout, demonstrating the sample-adaptive selection of perspectives based on sample characteristics.

C Training Details

This section reports the full implementation and training configuration for ProCrit.

C.1 Hyperparameters

We have illustrated the key hyper-parameters in §4.1. In this section, we provide more information about the hyper-parameters used in our experiment. Tables 6 and 7 summarize the specific hyperparameter settings for supervised fine-tuning and reinforcement learning, respectively. For the dual-stage GRPO, we utilize the OpenR1 framework for training.

Table 6: Hyperparameters for supervised fine-tuning.

Hyperparameter	Proposal Agent	Critic Agent
Epochs	3	3
Learning rate	1×10^{-5}	1×10^{-5}
Warmup ratio	0.1	0.1
Per-device batch size	4	4
Gradient accumulation	2	2
Max sequence length	2048	2048
Precision	bf16	bf16

Table 7: Hyperparameters for GRPO reinforcement learning.

Hyperparameter	Proposal Agent	Critic Agent
Epochs	4	4
Learning rate	3×10^{-6}	1×10^{-6}
Per-device batch size	8	8
Gradient accumulation	4	4
Max completion length	512	768
KL coefficient β	0.02	0.0
Clip ratio ϵ	0.2	0.2
Precision	bf16	bf16

C.2 Advantage Estimation

Following GRPO, advantages are standardized within each candidate set of the dual-stage rollout (§3.3.2):

$$\hat{A}_{\text{draft}}(d_i) = \frac{r_{\text{draft}}(d_i) - \mu}{\sigma + \epsilon}, \qquad \mu = \text{mean}_{d \in \mathcal{D}}\, r_{\text{draft}}(d), \qquad \sigma = \text{std}_{d \in \mathcal{D}}\, r_{\text{draft}}(d),$$
$$\hat{A}_{\text{revise}}(o_{k,j}) = \frac{r_{\text{revise}}(o_{k,j}) - \mu_k}{\sigma_k + \epsilon}, \quad \mu_k = \text{mean}_{o \in \mathcal{O}_k}\, r_{\text{revise}}(o), \quad \sigma_k = \text{std}_{o \in \mathcal{O}_k}\, r_{\text{revise}}(o), \tag{12}$$

where $\epsilon = 10^{-4}$ is a small constant for numerical stability.

C.3 PROMPTS

This subsection provides the complete prompts used in reasoning annotation synthesis (§3.1), proposal agent inference (drafting and revision), and critic evaluation, as referenced throughout §3.2.

Reasoning annotation synthesis. During the dynamic-role agentic rollout (§3.1), the teacher model $\mathcal{T}$ receives an initialization prompt (Prompt 1) that specifies the structured JSON output format for each reasoning step. At every subsequent turn, a follow-up prompt (Prompt 2) elicits a new analytical perspective conditioned on the accumulated analyses, following the rollout procedure of Eq. 1.

Proposal drafting. At inference time, the proposal agent uses the *dynamic-perspective* prompt (Prompt 3), in which the model autonomously determines the analytical perspectives needed for each sample. For controlled comparison, we additionally design two ablation baselines: a *fixed-perspective* prompt (Prompt 4) that prescribes three predefined perspectives, and a *generic* prompt (Prompt 5) that provides a minimal chain-of-thought instruction without perspective guidance. These two variants correspond to the Fixed-role Reasoner and Plain CoT Reasoner baselines in Table 1, respectively. The phrase "three steps" in Prompt 5 is vestigial; the prompt prescribes neither perspectives nor a step structure.

Prompt 1: Data Synthesis System Prompt

You are an expert AI assistant that explains your reasoning step by step. For each step, provide a title that describes what you're doing in that step, along with the content. Decide if you need another step or if you're ready to give the final answer. Respond in JSON format with 'title', 'content', and 'next_action' (either 'continue' or 'final_answer') keys. USE AT LEAST 2 STEPS OF REASONING. Decide confidently and never respond with "not enough context", "cannot determine", or any form of hedging. YOU CAN BE WRONG. WHEN YOU SAY YOU ARE RE-EXAMINING, ACTUALLY RE-EXAMINE, AND USE ANOTHER APPROACH TO DO SO. DO NOT JUST SAY YOU ARE RE-EXAMINING. USE AT LEAST 3 METHODS TO DERIVE THE ANSWER. USE BEST PRACTICES. Example of a valid JSON response: `{"title": "Identifying Key Information", "content": "To begin solving this problem, we need to carefully examine the given information and identify the crucial elements that will guide our solution process. This involves...", "next_action": "continue"}` Strictly follow the JSON response format. The response should start with `{`. DO NOT START WITH "“‘json" OR ANYTHING ELSE.

Prompt 2: Data Synthesis Follow-up Prompt

Continue with exactly one new reasoning step. Return a valid JSON object with keys: `title`, `content`, `next_action`. Please provide an one-more different step analysis here that you believe is likely to answer the question correctly. Do not provide the final answer yet unless you have already produced at least 2 reasoning steps. If you are not done, set `next_action` to `continue`.

Prompt 3: Dynamic-Perspective Proposal Drafting Prompt

Question
`<image>`
Text: {text}
Does the composite message of this image-text pair qualify as ironic/sarcastic?
Instruction
1. Output step-by-step analysis strictly within `<think>...</think>` tags.
Requirements:
- Decide the necessary number of steps (recommended: 3–5)
- Assign each step a clear, incisive title (e.g., "Step2: Pragmatic intent decoding.")
- For each step, explicitly select an analytical perspective that is:
  - most relevant to the characteristics of the specific image-text pair

Prompt 3 (continued)

  - logically builds upon / responds to the findings and open questions from previous steps
- Provide concise, critical, evidence-focused analysis under each step (avoid speculation, prioritize observable multimodal cues)

2. After completing all steps, output the final answer within `<answer>...</answer>` tags using one of the following options only:
- yes (sarcasm or irony is present)
- no (sarcasm or irony is not present)

Prompt 4: Fixed-Perspective Proposal Drafting Prompt

Question
`<image>`
Text: {text}
Does the composite message of this image-text pair qualify as ironic/sarcastic?
Instruction
1. Output the analysis strictly within `<think>...</think>` tags.
Use exactly only the following three fixed perspectives, in this exact order and with these exact step titles:
Step1: Surface-Level Discrepancy Analysis. Analyze whether there is an obvious mismatch, exaggeration, reversal, or unexpected contrast between the image and the text at the surface level.
Step2: Semantic Relation Analysis. Analyze how the text semantically relates to the image: whether it reinforces the image, contradicts it, reframes it, mocks it, or creates an implied meaning beyond the literal content.
Step3: Sentiment & Tone Analysis. Analyze the tone, attitude, emotional polarity, and pragmatic intent of the text in relation to the image, focusing on whether the speaker's apparent attitude signals sarcasm or irony.
2. After completing the three steps, output the final answer within `<answer>...</answer>` tags using one of the following options only:
- yes (sarcasm or irony is present)
- no (sarcasm or irony is not present)

Prompt 5: Generic Proposal Drafting Prompt

Question
`<image>`
Text: {text}
Does the composite message of this image-text pair qualify as ironic/sarcastic?
Instruction
Analyze the image-text pair with a chain of thought, then determine whether it is sarcastic.
1. Output the analysis strictly within `<think>...</think>` tags. Detailed analysis and reasoning steps supporting the conclusion.
2. After completing the three steps, output the final answer within `<answer>...</answer>` tags using one of the following options only:
- yes (sarcasm or irony is present)
- no (sarcasm or irony is not present)

Critic evaluation. The critic agent uses the following prompt (Prompt 6) to evaluate the quality of the proposal's reasoning process. The scoring rubric defines a 0–2 scale that evaluates interpretation accuracy, cross-modal reasoning, and reasoning coherence.

Prompt 6: Critic Evaluation Prompt

You are an expert evaluator for multimodal sarcasm/irony detection reasoning.
You are given an image-text pair along with a reasoning chain and concluded answer that analyzes whether this pair conveys sarcasm/irony.
Input
Text: {text}

Prompt 6 (continued)

Reasoning and answer:
{reasoning}
Your task
Evaluate the quality of this reasoning process — whether it correctly understood what is happening in this image-text pair and built a sound argument.
A correct conclusion does not guarantee a high score — evaluate the reasoning process itself, not just whether the answer matches.
Focus on:

1. **Interpretation accuracy (primary)** — Does the reasoning correctly interpret the combined meaning of the image-text pair and explain WHY it is or isn't sarcastic?
2. **Cross-modal reasoning** — Does it connect image and text into joint reasoning? For sarcastic pairs: does it identify how they contradict or recontextualize each other to create irony? For non-sarcastic pairs: does it show how they reinforce the same tone and real meaning?
3. **Reasoning coherence and efficiency** — Does the evidence chain build logically toward the conclusion, with each step contributing a concrete cue or reasoning move? Penalize unsupported leaps, contradictions, and filler steps that do not serve the final judgment.

Rate the reasoning on a 0–2 scale.
Scoring rubric
0 = Misunderstanding — the reasoning does not correctly understand this image-text pair. Common patterns:

- Treating any image-text contrast as automatic sarcasm
- Confusing playful banter, humor, or jokes with sarcasm
- Treating sarcasm-related words as proof of sarcastic intent
- Assuming positive wording must be sincere, ignoring image contradiction
- Missing subtle sarcasm due to a narrow definition
- Analyzing text in isolation without considering the image
- Treating absence of exaggeration as evidence of sincerity

1 = Partial understanding — the reasoning is on the right track but the analysis is weak. Common patterns:

- Identifies some relevant cues but misses others
- Stays surface-level without explaining cross-modal interaction
- Includes filler, detours, or redundant steps
- Has logical gaps or weak connections between steps

2 = Strong understanding — The reasoning clearly interprets what this pair is trying to express, precisely pinpoints the specific elements that create the sarcastic/non-sarcastic effect, and builds on joint image-text reasoning. Key cues are not missed, and the conclusion follows naturally.
Feedback (2–4 sentences): Briefly note what was done well or poorly. If the reasoning is weak or wrong, point out what was missed, misread, or insufficiently supported and indicate what the reasoning should have focused on instead.
Output format
`<feedback>`your feedback`</feedback>`
`<score>`[0–2]`</score>`

Proposal revision. After receiving the critic's feedback, the proposal agent uses the following revision prompt (Prompt 7) to generate a completely new analysis from scratch, rather than editing or extending the previous draft.

Prompt 7: Proposal Revision Prompt

Re-answer the original question from scratch. Using the reviewer's feedback below as reference, write a completely new analysis instead of editing or extending your previous one. Use the same output format.
Feedback: {feedback}

D Additional Experiments

This section collects supplementary experimental results and analyses that complement the main paper.

D.1 Human Evaluation Protocol and Full Results

We conduct two separate human evaluations. The first audits 100 label-balanced GLM-4.6V trajectories sampled from the synthesized training supervision; the second evaluates outputs from four systems on 100 label-balanced MMSD2.0 test examples. Five trained graduate annotators follow written definitions, score anchors, and worked examples. Gold labels are hidden, and system outputs are anonymized and randomly ordered. All dimensions use a 0–5 scale. Tables 8 and 9 report means and sample standard deviations across examples. Because the teacher and system outputs are evaluated on separate sample sets, the teacher row is an audit reference rather than a paired system comparison.

Table 8: Human evaluation of complete-rationale quality (mean ± sample standard deviation). Hallucination is lower-is-better; all other dimensions are higher-is-better.

Source / Method	Evidence grounding	Logical coherence	Rationale–label consistency	Core-cue coverage	Hallucination↓	Overall
Plain CoT	1.84 ± 2.03	1.93 ± 1.94	3.72 ± 2.10	1.84 ± 1.96	2.56 ± 2.10	1.56 ± 1.88
Fixed-role	1.78 ± 1.96	1.96 ± 1.96	3.86 ± 1.98	1.80 ± 1.97	2.88 ± 2.07	1.50 ± 1.84
ProCrit Draft	$\mathbf{2.96 \pm 2.25}$	$\mathbf{2.89 \pm 2.17}$	4.09 ± 1.90	2.62 ± 2.36	0.95 ± 1.68	2.47 ± 2.17
ProCrit Revise	2.89 ± 2.14	2.72 ± 2.15	$\mathbf{4.16 \pm 1.84}$	$\mathbf{2.66 \pm 2.23}$	$\mathbf{0.84 \pm 1.87}$	$\mathbf{2.63 \pm 2.27}$
GLM-4.6V training rationale	4.06 ± 1.78	4.03 ± 1.75	4.72 ± 1.01	3.73 ± 2.02	0.42 ± 1.18	3.72 ± 2.00

Table 9: Human evaluation of perspective quality (mean ± sample standard deviation). Redundancy and unsupported content are lower-is-better; all other dimensions are higher-is-better.

Source / Strategy	Relevant	Unique contribution	Redundant↓	Unsupported↓	Necessary for conclusion
Fixed predefined roles	3.34 ± 1.88	4.00 ± 1.46	0.91 ± 1.33	2.82 ± 2.08	3.77 ± 1.66
Self-elicited roles, Draft	4.04 ± 1.85	4.37 ± 1.33	$\mathbf{0.52 \pm 1.00}$	$\mathbf{1.51 \pm 2.15}$	4.07 ± 1.72
Self-elicited roles, Revise	$\mathbf{4.17 \pm 1.66}$	$\mathbf{4.52 \pm 1.19}$	$\mathbf{0.52 \pm 0.97}$	1.80 ± 2.19	$\mathbf{4.42 \pm 1.30}$
GLM-4.6V training supervision	4.53 ± 1.31	4.61 ± 0.86	0.55 ± 1.05	0.65 ± 1.60	4.52 ± 1.18

Dimension-wise ordinal Krippendorff's α ranges from 0.68 to 0.75 for complete-rationale quality and from 0.70 to 0.78 for perspective quality. The corresponding within-one-point agreement rates range from 82% to 90% and from 86% to 93%, respectively. These statistics indicate substantial consistency while preserving the expected subjectivity of fine-grained rationale assessment.

D.2 Generalization Across Backbones and Prompting Regimes

We evaluate unstructured CoT, predefined-perspective prompting, and adaptive-perspective prompting with Qwen2.5-VL-3B/7B/32B and InternVL3-14B under zero-shot and few-shot prompting. The few-shot demonstrations are fixed training-split examples shared across methods. These experiments test the portability of the perspective-generation procedure; they are single-pass prompting evaluations and do not include ProCrit's critic or post-training.

Table 10: MMSD2.0 results across model sizes, families, and prompting regimes. ZS and FS denote zero-shot and few-shot, respectively. Best values are bold; second-best values are underlined.

Method	Prompt	Qwen2.5-VL-3B		Qwen2.5-VL-7B		Qwen2.5-VL-32B		InternVL3-14B	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc
CoT	ZS	41.5	**64.8**	61.7	71.2	**72.2**	73.7	**73.8**	**75.1**
CoT	FS	21.8	59.9	<u>66.2</u>	71.3	69.0	73.4	70.7	73.2
Predefined perspectives	ZS	53.2	<u>62.3</u>	61.6	71.3	<u>71.8</u>	73.0	69.0	63.6
Predefined perspectives	FS	33.2	61.4	59.2	<u>71.5</u>	71.3	73.2	70.4	68.2
Adaptive perspectives	ZS	**57.1**	<u>62.3</u>	**67.4**	**71.7**	71.0	**74.0**	<u>73.6</u>	73.5
Adaptive perspectives	FS	<u>53.9</u>	<u>62.3</u>	63.0	69.4	70.1	72.4	72.7	<u>74.7</u>

Adaptive-perspective prompting remains competitive across model families on both MMSD2.0 and RedEval. On RedEval, it is the strongest zero-shot strategy for the Qwen backbones and the second-best for InternVL3-14B, outperforming predefined perspectives across all four backbones. Under few-shot prompting, adaptive perspectives remain competitive with predefined roles across both datasets, while the relative differences depend on the backbone and dataset. The post-trained Draft/Revise models further improve over these prompting-based results, showing that the gains extend beyond prompting alone.

Table 11: RedEval results across model sizes, families, and prompting regimes. Best values are bold; second-best values are underlined.

Method	Prompt	Qwen2.5-VL-3B		Qwen2.5-VL-7B		Qwen2.5-VL-32B		InternVL3-14B	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc
CoT	ZS	49.6	69.4	68.1	76.3	71.7	74.3	**74.0**	**75.9**
CoT	FS	32.1	65.8	62.8	71.1	69.0	74.2	69.7	73.2
Predefined perspectives	ZS	59.9	**69.8**	62.2	74.1	72.5	75.4	69.4	66.5
Predefined perspectives	FS	32.7	66.4	64.4	76.5	69.4	74.2	70.2	70.6
Adaptive perspectives	ZS	**60.0**	65.4	**71.4**	**76.8**	**75.9**	**78.8**	72.6	72.7
Adaptive perspectives	FS	57.8	66.3	70.1	76.1	69.4	73.0	72.1	75.2

D.3 Task-Specific and Proprietary Baselines

Table 12 compares against task-specific classifiers and retrieval/distillation methods using Macro-F1 throughout. This metric differs from the positive-class F1 used in Table 1. Prior results are taken from the corresponding papers or published benchmark evaluations; ProCrit Macro-F1 is recomputed from per-example predictions.

Table 12: Macro-F1 comparison with task-specific and retrieval-based methods.

Method	MMSD	MMSD2.0	RedEval	Explicit rationale
Multi-view CLIP (Qin et al., 2023)	87.70	85.50	80.73	✗
Tang et al. (2024)	**89.42**	**86.34**	82.83	✗
PEKD (Jana & Sanasam, 2025)	84.30	81.10	69.80	✗
ProCrit Draft	80.72	82.39	**86.85**	✓
ProCrit Revise	82.43	83.93	85.45	✓

Tang et al. obtains the strongest in-domain Macro-F1 on MMSD and MMSD2.0, whereas ProCrit Draft obtains the highest Macro-F1 on the OOD RedEval benchmark. Thus, the retrieval-based classifier is stronger on in-domain classification, while ProCrit shows stronger cross-domain performance on RedEval and additionally exposes a human-evaluated rationale.

We also evaluate two larger proprietary VLMs with the same simple zero-shot multimodal instruction, asking each model to predict a binary sarcasm label. Table 13 reports positive-class F1, matching the main table's metric.

Table 13: Positive-class F1 comparison with proprietary zero-shot VLMs.

Model / Mode	MMSD	MMSD2.0	RedEval
GPT-5.4	75.23	76.85	82.52
Gemini-3.5-Flash	75.10	75.90	76.67
ProCrit Draft	78.40	81.10	**84.80**
ProCrit Revise	**80.70**	**83.10**	83.70

Both ProCrit modes exceed the two proprietary zero-shot baselines on all three datasets. This performance pattern is consistent with prior findings that general-purpose LLMs do not reliably capture the implicit pragmatic and cross-modal semantics underlying sarcasm (Zhang et al., 2025b; Wang et al., 2025b).

D.4 Reward and Balance-Weight Ablations

This subsection collects the ablations of the reward terms and of the draft–revise balance weight that are summarized in §4.3.

Draft–Revise balance weight λ. Table 14 reports the sensitivity to λ in Eq. 9. The best F1 is attained at $\lambda = 0.5$, and both extremes ($\lambda = 0$, Draft only; $\lambda = 1$, Revise only) are weaker. Joint

Table 14: Ablation on the Draft–Revise balance weight λ in the proposal agent objective. Larger λ assigns more weight to the Revise objective.

	MMSD2.0 (F1)		MMSD (F1)	
λ	Draft	Revise	Draft	Revise
0	80.5	82.7	75.5	80.2
0.25	80.4	82.8	78.3	80.6
0.50	**81.1**	**83.1**	78.4	**80.7**
0.75	80.9	82.4	**78.5**	80.1
1	78.4	82.5	77.8	79.9

Table 15: Ablation on proposal agent rewards.

	MMSD2.0 (F1)		MMSD (F1)	
Setting	Draft	Revise	Draft	Revise
Full	**81.1**	**83.1**	78.4	**80.7**
w/o r_{eval}	79.9	82.4	77.8	80.3
w/o $r^{\text{d}}_{\text{acc}}$	80.2	82.4	77.6	80.1
w/o $r^{\text{d}}_{\text{fmt}}$	80.5	82.5	77.7	80.0
w/o r_{imp}	80.7	82.6	**79.0**	79.7
w/o $r^{\text{r}}_{\text{acc}}$	80.9	82.2	78.1	80.2
w/o $r^{\text{r}}_{\text{fmt}}$	79.9	82.7	77.9	80.2

Table 16: Ablation on critic agent rewards.

	MMSD2.0 (F1)	MMSD (F1)
Setting	Revise	Revise
Full	**83.1**	**80.7**
w/o r_{align}	82.2	79.6
w/o r_{act}	82.6	79.4
w/o $r^{\text{c}}_{\text{fmt}}$	82.4	80.1

optimization is thus preferable in-domain, whereas RedEval favors a larger λ (Appendix D.5); we use $\lambda = 0.5$ throughout.

Proposal agent rewards. Table 15 ablates the reward components of both modes. In Draft mode (Eq. 6), removing the critic evaluation reward r_{eval} leads to the largest drop, suggesting that the critic's score provides finer-grained supervision than binary accuracy alone; removing $r^{\text{d}}_{\text{acc}}$ or $r^{\text{d}}_{\text{fmt}}$ also degrades performance. In Revise mode (Eq. 7), removing r_{imp} reduces Revise F1 by 1.0 on MMSD and 0.5 on MMSD2.0 but not on RedEval (Appendix D.5); we keep r_{imp} for its explicit credit assignment to genuine corrections.

Critic agent rewards. Table 16 shows that both r_{align} and r_{act} contribute: on MMSD2.0, removing r_{align} decreases Revise F1 from 83.1 to 82.2 and removing r_{act} to 82.6. The former aligns the critic's scores with correctness; the latter encourages feedback that the proposal agent can act on. Ablating $r^{\text{c}}_{\text{fmt}}$ also hurts, consistent with complementary roles of the three terms.

D.5 Complete RedEval Ablations

We repeat the complete proposal and critic ablation suite on RedEval. Table 17 covers training stages, critic rewards, and revision modes; Table 18 covers the balance weight and proposal rewards.

Table 17: Core RedEval ablations. All F1 values are positive-class F1.

(a) Training stages

SFT	RL	Draft	Revise
✗	✓	83.5	**84.7**
✓	✗	79.3	82.9
✓	✓	**84.8**	83.7

(b) Critic rewards

Setting	Revise
Full	**83.7**
w/o r_{align}	83.5
w/o r_{act}	83.1
w/o $r^{\text{c}}_{\text{fmt}}$	82.3

(c) Revision modes

Setting	F1	Acc
None (Draft)	**84.8**	**87.2**
Self-revision	80.1	83.2
Initialized critic	79.9	83.2
Proposal-refined critic	83.7	85.7

On RedEval the ablations agree with the in-domain results on the main points but not uniformly: the complete SFT+RL pipeline gives the strongest Draft F1, the full critic objective gives the best Revise F1 among critic-reward variants, and the proposal-refined critic remains the best revision-based mode, whereas several removals change Revise F1 by about one point or less in either direction (e.g., omitting SFT or $r^{\text{d}}_{\text{acc}}$ yields 84.7 and 83.9 Revise F1 against 83.7 for the full model), and the balance

Table 18: RedEval ablations of the proposal objective.

(a) Draft–Revise weight

λ	Draft	Revise
0	83.6	83.4
0.25	84.6	84.1
0.50	84.8	83.7
0.75	**84.9**	84.1
1	80.7	**84.2**

(b) Proposal rewards

Setting	Draft	Revise
Full	**84.8**	83.7
w/o r_{eval}	83.4	82.7
w/o $r^{\text{d}}_{\text{acc}}$	**84.8**	**83.9**
w/o $r^{\text{d}}_{\text{fmt}}$	83.8	83.1
w/o r_{imp}	84.5	**83.9**
w/o $r^{\text{r}}_{\text{acc}}$	82.9	83.5
w/o $r^{\text{r}}_{\text{fmt}}$	83.1	83.5

weight prefers $\lambda = 0.75$ for Draft and $\lambda = 1$ for Revise. RedEval also highlights the operating-point trade-off: Draft yields higher F1 and accuracy, while full revision raises recall from 91.4 to 93.7. We therefore read the RedEval ablations as confirming the training pipeline and the critic design, while indicating that the reward-level choices are within noise out-of-domain.

D.6 Optimization Order and Training Stability

The default schedule optimizes the critic first and the proposal second, freezing the counterpart in each stage. Reversing this order on MMSD2.0 changes Draft/Revise F1 from 81.1/83.1 to 80.9/82.6. Across the monitored reward, KL, and gradient-norm curves, both stages completed without sustained divergence or collapse.

Table 19: Optimization-order sensitivity and three-seed variation on MMSD2.0. Standard deviations are sample standard deviations across training runs.

Order	**Draft**	**Revise**
Critic first	**81.1**	**83.1**
Proposal first	80.9	82.6

Mode	**Seed 42**	**Seed 43**	**Seed 44**	**Mean $\pm$ std.**
Draft	81.1	80.4	81.5	81.0 ± 0.56
Revise	83.1	82.6	83.0	82.9 ± 0.26

Beyond optimization-order sensitivity, we assess run-to-run variation by training the default critic-first configuration with three random seeds (42, 43, and 44). Across these runs, Revise improves over its paired Draft by 2.0, 2.2, and 1.5 F1 points, respectively, with a mean paired gain of 1.9 ± 0.36. The improvement is positive for all three seeds.

D.7 Cross-Task Transfer to Hateful Memes

We extend ProCrit to multimodal hate-speech detection on the Hateful Memes benchmark (Kiela et al., 2020). We retain the Qwen2.5-VL-7B backbone and the proposal-critic design, optimizing the proposal directly with GRPO and initializing the critic from 10,000 synthesized feedback examples for the new task. All methods are evaluated on the same 2,408 test examples with available images. AUROC is computed from normalized next-token probabilities of "yes" and "no" conditioned on the generated rationale.

Table 20: Cross-task evaluation on Hateful Memes. Best and second-best values are bold and underlined, respectively.

Method / Mode	**AUROC**	**Acc**	**Macro-F1**	**Hate-F1**
Qwen2.5-VL zero-shot	69.5	65.9	59.9	44.3
Qwen2.5-VL 2-shot	72.3	67.7	66.9	61.4
Plain CoT Reasoner	80.3	73.1	72.1	66.7
Fixed-role Reasoner	<u>81.3</u>	74.1	73.0	67.4
ProCrit Draft	<u>81.3</u>	<u>76.0</u>	<u>75.4</u>	<u>71.4</u>
ProCrit Revise	**83.5**	**78.2**	**77.8**	**74.0**

ProCrit achieves the best result across all four metrics, improving over the Fixed-role Reasoner by 2.2 AUROC, 4.1 Accuracy, 4.8 Macro-F1, and 6.6 Hate-F1 points. It also consistently improves

over ProCrit Draft, showing that critic-guided revision remains effective on the new task. Because the proposal is trained here with GRPO under its drafting prompt without synthesized rollouts, this experiment tests the transfer of adaptive-perspective prompting and of proposal–critic collaboration, not of the full annotation-synthesis pipeline; within that scope, both components carry over to another multimodal understanding task that hinges on subtle image–text interactions.

D.8 Additional Ablations

Effect of Multiple Revision Rounds. We investigate whether iterating the critique–revise cycle beyond a single round yields further improvements. Starting from the same draft, the proposal agent receives feedback from the critic agent and revises its reasoning; this process is repeated for up to five rounds on MMSD2.0. Table 21 reports the results.

Table 21: Effect of multiple revision rounds on MMSD2.0. Each round applies one critique–revise cycle using the same critic agent.

Round	F1	Acc	P	R
Draft	81.1	82.5	75.7	87.4
Revise 1	83.1	84.0	76.1	91.5
Revise 2	83.1	84.1	76.6	90.7
Revise 3	82.8	83.8	76.2	90.7
Revise 4	83.1	84.1	76.4	91.0
Revise 5	83.3	84.2	76.6	91.1

The first revision round contributes the dominant improvement, raising F1 from 81.1 to 83.1 (+2.0 points) with a substantial recall gain from 87.4 to 91.5. Subsequent rounds yield marginal and inconsistent changes: F1 fluctuates within a narrow range of 82.8–83.3 across rounds 2–5, without a steady improvement. This indicates that a single critique–revise cycle is sufficient to capture the actionable feedback from the critic agent, and additional iterations provide limited additional gains. In practice, ProCrit therefore adopts a single revision round, which achieves the majority of the performance gain while keeping inference cost manageable.

D.9 Efficiency Analysis

We analyze the computational cost of ProCrit from both training and inference perspectives.

Training cost. ProCrit's training pipeline involves four stages, each operating on a 7B-parameter backbone (Qwen2.5-VL-7B) with 8 NVIDIA A800 GPUs under DeepSpeed ZeRO-3. Table 22 summarizes the data scale and approximate GPU hours for each stage. The reported stage-wise training budget totals 42.8 hours on the 8-GPU setup. Notably, all stages use modest data scales: the largest being critic SFT with 20K examples and the RL stages operate on only 2K and 5K instances, respectively. This is substantially smaller than the hundreds of thousands of samples typically used in general-purpose reasoning RL (Guo et al., 2025a), reflecting the efficiency of targeted, task-specific optimization.

Table 22: Training cost breakdown across the full pipeline on 8×A800 GPUs.

Stage	Data	Epochs	GPU Hours
Proposal SFT	9K	3	0.70
Critic SFT	20K	3	1.80
Critic GRPO	5K	4	19.50
Proposal GRPO	2K	4	20.80
Total			42.8

Inference cost. At test time, ProCrit offers two operating modes with different cost–performance trade-offs. In *draft-only* mode (ProCrit Draft), the proposal agent performs a single forward pass,

incurring the same cost as any single-pass reasoning method while already achieving state-of-the-art performance (e.g., 81.1 F1 on MMSD2.0). In *draft–critique–revise* mode (ProCrit Revise), three sequential passes are required: the proposal agent generates a draft, the critic agent evaluates it and provides feedback, and the proposal agent produces a revised analysis. This introduces additional latency relative to single-pass methods; however, we note that existing multi-agent sarcasm detection approaches such as S3Agent (3 independent reasoning agents with a decision agent) and Commander-GPT (6 sub-task agents plus a decision module) also require multiple inference calls per sample. Importantly, a single revision round captures the dominant improvement (Table 21), and no iterative looping is needed—the overhead is bounded and predictable.

Generation length. The proposal agent's maximum generation length is 512 tokens and the critic agent's is 768 tokens (Table 7). In practice, the average draft length is considerably shorter: the majority of reasoning trajectories comprise 3 analytical steps (Table 5), with each step producing a concise, evidence-focused analysis. The revision pass generates output of comparable length to the draft, as it produces a new analysis from scratch rather than appending to the original. The critic's output consists of a quality score and a short natural-language feedback paragraph, typically much shorter than the 768-token budget.

E Qualitative Analysis

The examples from the MMSD2.0 test set illustrate how the proposal and critic agents interpret multimodal cues and refine their reasoning. In Figure 7, the critic identifies misinterpreted pragmatic cues or overlooked contextual irony in an incorrect draft and provides targeted feedback that guides the proposal toward a corrected analysis. Figure 8 shows that feedback can also improve a draft whose prediction is already correct: the critic identifies gaps in its reasoning, and the revision grounds the judgment more explicitly in the image and text. The campus example in Figure 9 further illustrates how the choice of cues affects the interpretation. The SFT-only proposal treats "it's too sunny" as a literal description of the lighting and predicts non-sarcasm, whereas the RL-trained proposal attends to the winter clothing, interprets the caption as ironic, and predicts sarcasm. The alarm-meme example in Figure 10 shows how feedback can prompt a different reading when self-revision retains the initial interpretation. Both the draft and self-revision read the laughter as sincere amusement and predict non-sarcasm. The critic instead points to self-deprecating irony in the enthusiastic identification with repeated alarms; the feedback-guided revision adopts this interpretation and predicts sarcasm, illustrating how a specific pragmatic cue can guide a change in judgment.

F Limitations

The draft–critique–revise paradigm requires three sequential forward passes at inference time (proposal draft, critic evaluation, and proposal revision), increasing computational cost compared to single-pass methods. We provide a detailed efficiency analysis in Appendix D.9 and note that the draft-only output already achieves competitive performance, offering a practical alternative when inference cost is a concern. Moreover, the current evaluation is conducted on English-language social media datasets (Twitter and Reddit). Sarcastic expression varies considerably across languages and cultural contexts, and validating the framework in multilingual and cross-platform settings remains an important direction for future work. Additionally, sarcasm perception is inherently subjective—the same expression may be interpreted as sarcastic by some audiences but not others, depending on shared context, cultural background, and interpersonal familiarity. The current benchmarks adopt majority-vote labels that inevitably smooth over such disagreements. Incorporating richer contextual signals (e.g., conversational history or author profile) and modeling annotator disagreement as soft labels could better capture this subjectivity and further improve robustness.

Furthermore, the critic's RL rewards are outcome-based: r_{align} and r_{act} both derive from label correctness, so the critic's score is label-correlated by construction and process quality is shaped only through SFT on teacher critiques. Relatedly, revision improves recall and in-domain F1 but lowers F1 on the out-of-domain RedEval, and human raters find revised rationales slightly less grounded than drafts; a natural remedy is to revise selectively when the critic flags a draft. Finally, the two agents are optimized for a single alternation round; whether further rounds help remains open. The teacher model may also have encountered the public MMSD corpora during pre-training; the student



Text

on mondays , i wear silk scarfs and posh shoes and sit on nice rugs . what ' s your monday like ? - - - - - - - - - - - - -

Reasoning

<think>Step1: Analyze the typical association of Mondays. Mondays are commonly linked to fatigue, casual attire, and a lack of enthusiasm for starting the workweek, contrasting with the "glamorous" elements mentioned in the text (silk scarfs, posh shoes, nice rugs).Step2: Evaluate the contrast between the described behavior and common Monday experiences. The text presents a polished, luxurious Monday routine, which deviates from the average person's likely experience of Mondays. This deviation suggests irony, as the speaker may be exaggerating or mockingly portraying an idealized version of a Monday to highlight the disparity with reality.Step3: Assess the tone of the question. By asking "what's your monday like?" after describing their own "fancy" Monday, the speaker invites comparison, implying that their experience is unusual or humorous compared to others' more mundane Mondays. This reinforces the ironic tone, as the statement is likely not meant to be taken literally but rather to emphasize the absurdity of such a polished start to the week.</think><answer>yes</answer>



Text

welcome back alan rickman . founder of mughal empire . sorrynotsorry

Reasoning

<think>Step1: Identify key references and factual mismatches. Alan Rickman was a British actor (known for roles like Severus Snape), while the Mughal Empire was founded by Babur, a historical figure from the 16th century. The text incorrectly attributes the Mughal Empire's founding to Rickman, creating a deliberate factual error.Step2: Analyze tonal cues. "Welcome back" references Rickman's death (he passed in 2016), so the phrase is sarcastic—mocking the idea of "welcoming back" someone who cannot return. "Sorrynotsorry" is a common sarcastic tag, indicating the speaker does not genuinely apologize for the joke, reinforcing the ironic tone.Step3: Evaluate contrast between expectation and reality. The text pairs a modern actor with a historical achievement he had no connection to, using absurdity to create humor. This intentional misalignment of facts and context is a hallmark of sarcasm/irony, as it subverts expectations for comedic effect.</think><answer>yes</answer>



Text

i love my dissertation !

Reasoning

<answer>Step1: Analyze the context of the text and image. The image shows a person who appears to be a student (indicated by the lanyard and academic posters in the background), and the text states "i love my dissertation !".Step2: Consider common attitudes toward dissertations. Dissertations are widely known to be challenging, stressful, and time-intensive tasks for students, so expressing genuine love for one is unusual.Step3: Evaluate the tone of the text. The exclamation mark and casual phrasing suggest emphasis, but given the typical negative association with dissertations, the statement is more likely sarcastic (meaning the opposite of what is said) rather than literal.Step4: Determine if there's a contrast between the stated sentiment and expected reality. Since loving a dissertation contradicts the common experience of students, the text uses sarcasm to convey frustration or dislike disguised as affection.</think><answer>yes</answer>



Text

you found that someone & we couldn 't be any happier that you are officially jadine !

Reasoning

<think>Step1: Analyze the original post's intent. The original text expresses a genuine desire for companionship ("need someone that will help me think of some besides myself") and a longing for connection ("#iwantsomeonetowalkthruethefrontdoorwith").
Step2: Examine the comment's tone. The comment "you found that someone & we couldn't be any happier that you are officially jadine !" uses positive language ("couldn't be any happier," "officially jadine") to celebrate the original poster's new relationship. There is no indication of saying the opposite of what is meant (sarcasm) or a contrast between expectation and reality (irony). The comment aligns with the original post's theme of seeking connection, making it sincere rather than sarcastic.</think><answer>no</answer>



Text

this one deserves to be put on a picture frame and place it on the wall a family picture tho ahhhh this is so goood

Reasoning

<think>Step1: Analyze the textual expression for signs of insincerity or contrast between stated sentiment and implied meaning. The text uses phrases like "deserves to be put on a picture frame," "family picture," and "so goood" (with elongated vowels), which initially appear to express genuine enthusiasm and approval.Step2: Evaluate contextual cues. The image depicts a posed group photo, likely of a musical group, with members smiling and arranged in a typical "family photo" style. The term "family picture" is commonly used affectionately for close-knit groups (e.g., bands), suggesting the speaker views the group as cohesive and the photo as worthy of display. There is no obvious contradiction between the text's positive tone and the image's content, nor are there markers of sarcasm (e.g., mocking language, exaggerated disbelief).Step3: Conclude based on tone and context. The text conveys sincere admiration for the photo, framing it as a cherished "family picture" suitable for display. No elements of sarcasm or irony are present, as the stated sentiment aligns with the visual content and typical affectionate language for such photos.</think><answer>no</answer>

Figure 6: Examples of reasoning trajectory synthesized through dynamic-role agentic rollout. Each analytical perspective is spawned sequentially within a shared context, progressively building a coherent multi-perspective analysis.

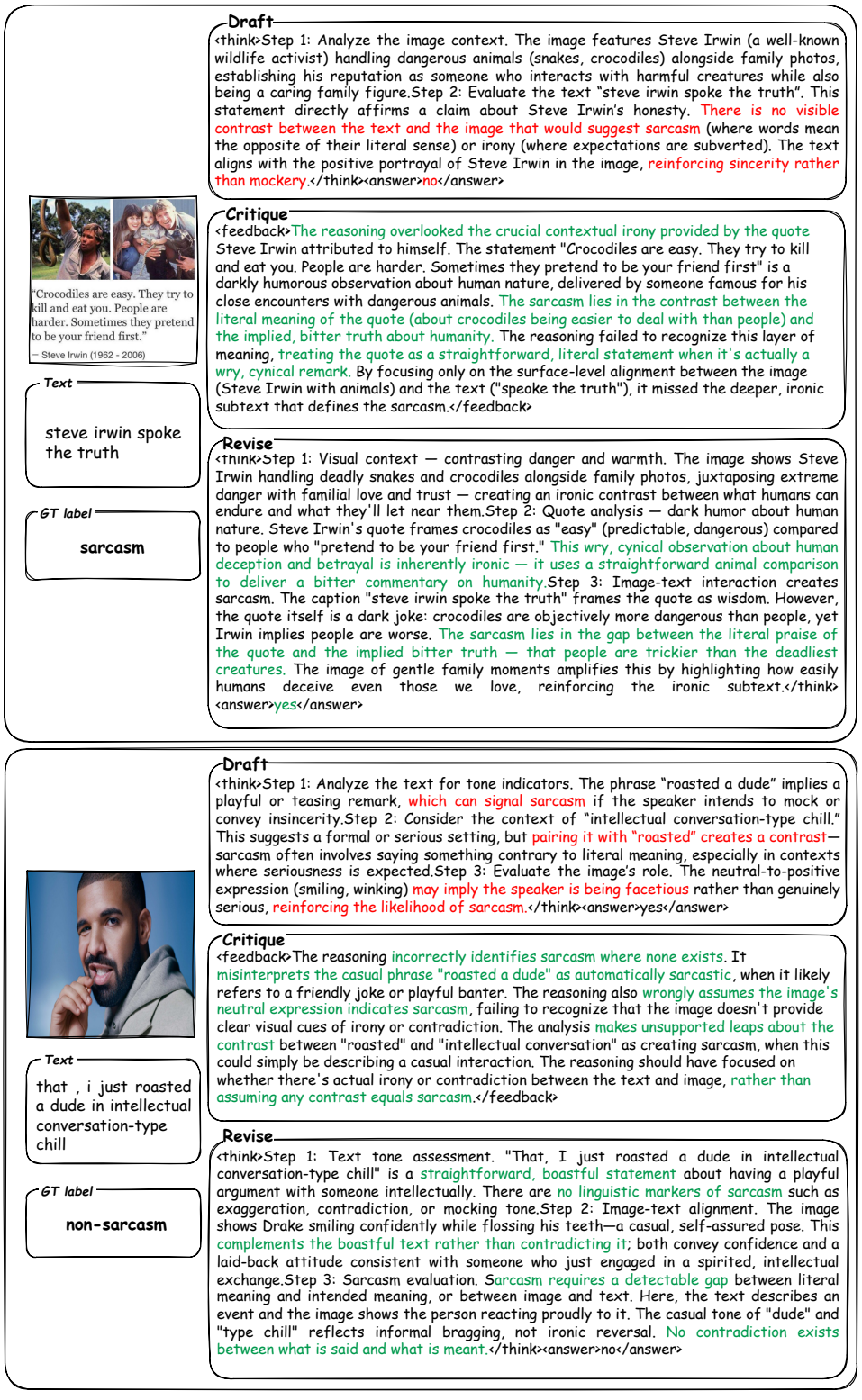


Figure 7: Successful answer correction via critic-guided revision. The draft produces an incorrect sarcasm judgment due to flawed cross-modal reasoning. The critic pinpoints the reasoning deficiency and provides targeted feedback, enabling the proposal agent to generate a corrected reasoning from scratch.



Text

thanks for the wonderful ride <user>

GT label

sarcasm

Draft

<think>Step 1: Analyze the image context. The image shows a person's feet on a train floor with seats visible, suggesting a public transportation setting. The scene appears ordinary without obvious signs of discomfort or negative experience.Step 2: Evaluate the text "thanks for the wonderful ride <user>". The phrase "wonderful ride" typically expresses positive sentiment. However, if the actual experience (implied by the image or prior context) were negative (e.g., uncomfortable seating, delays, noise), the text would likely be sarcastic—using positive language to convey the opposite emotion.Step 3: Determine if there's a contrast between the text and implied reality. Since the image doesn't provide explicit negative details but the structure of the comment ("thanks for the [positive] ride") often signals sarcasm when paired with a subpar experience, the text is likely sarcastic.</think><answer>yes</answer>

Critique

<feedback>The reasoning correctly identifies that the text "thanks for the wonderful ride" could be sarcastic when paired with the image of a train interior, and it connects the image and text by noting the contrast between positive language and what might be an implied negative experience. However, the reasoning could be stronger by being more specific about what visual elements in the image suggest a negative experience (e.g., crowded space, worn seats, uncomfortable posture). The analysis is somewhat general and could benefit from pointing to particular visual cues that would make the sarcastic intent more apparent.</feedback>

Revise

<think>Step 1: Visual evidence of an uncomfortable train ride. The image shows a man's legs in athletic shoes with orange laces, wearing black shorts, seated on what appears to be a commuter train with blue carpeted floors and metal seats. His legs are crossed awkwardly at the ankles, and his posture suggests cramped, uncomfortable seating—he has little legroom, and the overall scene looks cramped and unpleasant.Step 2: Text-image incongruity analysis. The caption "thanks for the wonderful ride" uses overtly positive language, yet the visual context directly contradicts this. The cramped train interior, limited legroom, and casual posture while wearing athletic shoes on public transport all signal an uncomfortable journey—not a "wonderful" one.Step 3: Sarcastic intent confirmation. The deliberate contrast between the enthusiastic "wonderful ride" and the visibly cramped, uncomfortable train environment creates classic verbal irony. The speaker is mocking the poor conditions by feigning gratitude, making the composite message clearly sarcastic.</think><answer>yes</answer>



Text

calum with a fan today ! (oh goodness he 's so handsome) | via blink . 1998 on ig

GT label

non-sarcasm

Draft

<think>Step 1: Analyze the text content. The text states "calum with a fan today ! (oh goodness he 's so handsome)" which expresses admiration for Calum, describing him as "handsome."Step 2: Evaluate the tone and intent. The phrase "oh goodness he's so handsome" is a direct, positive compliment without any indication of insincerity, mockery, or contrast between the stated sentiment and underlying meaning.Step 3: Consider the context. The mention of "a fan" suggests a casual, friendly interaction, and the overall tone of the text aligns with genuine appreciation rather than sarcasm or irony.</think><answer>no</answer>

Critique

<feedback>The reasoning correctly identifies that the text expresses genuine admiration for Calum and concludes there's no sarcasm, which matches the ground truth. However, it focuses primarily on the text content without sufficiently connecting it to the visual elements in the image. The analysis would be stronger if it explained how the image (showing two people together outdoors) reinforces the text's positive tone and relationship context. While the conclusion is correct, the reasoning could better leverage the multimodal nature of the task by explicitly showing how the visual and textual elements work together to convey sincerity rather than sarcasm.</feedback>

Revise

<think>Step 1: Visual content analysis. The image shows a young man wearing sunglasses and a green jacket posing closely with a smiling woman outdoors under a blue sky, suggesting a friendly, fan-celebrity encounter at an event like a concert or meet-and-greet.Step 2: Text-image alignment check. The text calls this person "Calum" and refers to "a fan," placing them in a genuine fan-meeting context. The parenthetical "(oh goodness he's so handsome)" is a straightforward compliment, and the image confirms this—Calum appears attractive, and the woman is visibly happy posing with him, consistent with a fan's admiring reaction.Step 3: Sarcasm/irony evaluation. Sarcasm requires a mismatch between literal meaning and intended meaning, or between text and image. Here, both modalities reinforce the same sincere sentiment: a fan genuinely admiring someone at an event. The visual warmth (smiling faces, close proximity) and textual enthusiasm align perfectly, leaving no ironic gap.</think><answer>no</answer>

Figure 8: Reasoning quality improvement via critic feedback. The draft predicts the correct label but provides shallow, surface-level reasoning. After receiving the critic's feedback identifying analytical gaps, the revision produces a substantially deeper analysis that explicitly captures the cross-modal sarcastic forms.



Text

mikkeli university of applied sciences . it 's too sunny .

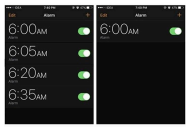
SFT-only Agent Reasoning Result

<think>Step 1: Analyze the image context. The image shows two individuals wearing sunglasses in front of a sign for MAMK University (Kasarmin Kampus Main Campus), with bright lighting suggesting sunny weather. Step 2: Evaluate the text's tone and intent. The text states "mikkeli university of applied sciences . it 's too sunny ." Since the image depicts a sunny environment, the statement is a direct observation rather than a contradiction or mockery. There is no indication of sarcasm or irony.</think><answer>no</answer>

RL-trained Agent Reasoning Result

<think>Step 1: Visual evidence analysis. Two young women wear bright blue sunglasses and knit hats/caps while taking a selfie in front of a MAMK (University of Applied Sciences) campus sign—cold weather attire despite what appears to be bright sunlight. Step 2: Text-image contradiction detection. The caption says "it's too sunny," yet they're bundled up in winter gear. Mikkeli is indeed known for severe winters, so "too sunny" is absurdly literal—it's actually too cold to be out in this kind of weather, making the statement the opposite of what's true. Step 3: Sarcastic intent confirmation. The humor arises from the deliberate mislabeling: calling cold weather "too sunny" when everyone knows it's freezing outside. This is classic verbal irony—the speaker means the exact opposite of the words to express frustration or humor about the miserable conditions.</think><answer>yes</answer>

Figure 9: The SFT-only and RL-trained proposals interpret the same campus image–text pair differently. The SFT-only proposal focuses on the bright lighting and predicts non-sarcasm, whereas the RL-trained proposal considers the cold-weather clothing and predicts sarcasm. Colored spans highlight the contrasting cues and interpretations.



Text

that 's me ! ! ! hahahahah

Initial Draft

<think>Step 1: Visual content analysis. The image shows several alarms set shortly after 6:00 AM. Step 2: Text interpretation. The laughter suggests enthusiasm or amusement, and the user finds humor in it. Step 3: Sarcasm assessment. The text directly affirms the image's content; the user genuinely relates to the behavior, with no insincerity or hidden meaning.</think><answer>no</answer>

Self-revision output

<think>Step 1: Visual content identification. The meme contrasts early risers and late sleepers through multiple alarms versus one alarm. Step 2: Text-image interaction. The text expresses enthusiastic self-identification with the late-sleeper category, while the laughter conveys genuine amusement rather than mockery. Step 3: Sarcasm assessment. The text straightforwardly affirms the image's message; the laughter is authentic self-deprecation, not ironic detachment, so no contradiction exists.</think><answer>no</answer>

Critic feedback

<feedback>Step 1: Missed tone. The reasoning overlooks self-deprecating sarcasm. Step 2: Missed textual cue. The exaggerated enthusiasm and laughter function as ironic mockery. Step 3: Missed contrast. The implicit contradiction lies between the cheerful affirmation and being late or unprepared. The alarms imply repeated failure to wake up on time, and the laughter is an ironic acknowledgment of this failure.</feedback><score>0</score>

Critic-guided revision

<think>Step 1: Visual content analysis. The multiple overlapping alarms suggest sleeping through the earliest one. Step 2: Text-image interaction. The text uses exaggerated enthusiasm, but the reality is being late or oversleeping. The laughter functions as ironic mockery and creates self-deprecating humor. Step 3: Sarcasm identification. The contrast between over-the-top excitement and relying on multiple alarms signals sarcasm; the speaker is sarcastically acknowledging the habit.</think><answer>yes</answer>

Figure 10: Self-revision and critic-guided revision produce different judgments for the alarm meme. The draft and self-revision predict non-sarcasm; the critic identifies self-deprecating irony and supplies targeted feedback, after which the proposal predicts sarcasm. The panels show both revision outputs and the critic's feedback.

is trained only on the MMSD2.0 training split, and its gains on RedEval, drawn from a different platform, are unlikely to reflect such exposure.

G BROADER IMPACT

Failure to recognize sarcasm can affect the accurate interpretation of sentiment and communicative intent in applications such as opinion mining, public sentiment analysis, and online content moderation. By improving the accuracy and robustness of multimodal sarcasm detection, ProCrit may help such systems better capture the intended meaning behind social media posts. In addition, the reasoning chains generated alongside predictions can provide greater transparency into the decision-making process, enabling more effective human oversight in sensitive deployment scenarios. We also acknowledge potential risks. Sarcasm detection systems could be misused for surveillance or censorship of online speech. Moreover, since sarcastic expression is culturally situated, deploying a model trained predominantly on English-language social media in other cultural contexts may produce systematic errors. We therefore recommend that ProCrit be used as a decision-support tool complementing human judgment rather than as a fully autonomous content filter.