ProCrit: Self-Elicited Multi-Perspective Reasoning with Critic-Guided Revision for Multimodal Sarcasm Detection

Yingjia Xu[1,3] **Jiulong Wu**[2] **Bowen Zhang**[1] **Baokui Guo**[3] **Siyuan Chai**[3] **Min Cao**[1*]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Baidu Inc., Beijing, China
[3]Zhipu AI, Beijing, China
{yjxuxuyj, mcao}@suda.edu.cn

Abstract

Multimodal sarcasm detection requires reasoning over cross-modal incongruities between literal expression and intended meaning, yet the specific analytical perspectives needed vary across samples due to the diversity of sarcastic mechanisms. 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, ProCrit synthesizes process-level reasoning annotations through a *dynamic-role agentic rollout*: a strong vision-language model sequentially spawns analytical roles within a shared context, and the resulting multi-role trajectories are flattened into sequences that preserve cross-perspective dependencies while enabling efficient autoregressive generation. Second, to improve reasoning reliability, ProCrit adopts a *draft–critique–revise* paradigm in which an independent critic identifies reasoning deficiencies and provides targeted natural-language feedback for directed revision. Finally, we develop a *mutual-refinement* training framework that jointly optimizes proposal drafting and feedback-guided revision via dual-stage reinforcement learning, while refining the critic agent according to the actual effectiveness of its feedback. Experiments on three widely used benchmarks demonstrate the effectiveness of ProCrit.

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 [1–3]. 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 mechanisms—verbal irony, hyperbole, rhetorical questions, cultural allusion, among others. Exposing the underlying

*Corresponding author.

Preprint.

incongruity requires examining the sample from multiple *analytical perspectives*—distinct angles of inspection such as probing a cultural reference, decoding exaggerated expressions, or evaluating scene plausibility against commonsense expectations, as illustrated in Fig. 6. Which perspectives are needed and how they should be combined varies across samples, as different instances of sarcasm rely on distinct mechanisms and distribute critical cues across text and image in different ways.

To address this challenge, early works [4–6] have explored increasingly sophisticated cross-modal fusion strategies through attention, graph-based, and contrastive learning mechanisms, in an effort to capture the diverse signals that underlie sarcastic intent. Recent methods 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. This calls for *self-elicited multi-perspective reasoning* in which the model autonomously explores and integrates different analytical perspectives on its own and progressively builds a sample-adaptive, coherent analysis.

In this work, we make the first attempt to equip multimodal sarcasm detection with *self-elicited multi-perspective reasoning*, a capability that enables the model to dynamically generate the analytical perspectives required for each input sample and progressively integrate them into a coherent analysis, where each perspective builds upon evidence uncovered by preceding analyses. 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 [7, 8] 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.** Because sarcastic intent often depends on subtle, implicit, and cross-modal cues, a model equipped with self-elicited multi-perspective reasoning capability may still leave important aspects unaddressed, misinterpret weak signals, or fail to resolve the underlying incongruity, making the resulting reasoning incomplete or flawed.

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 a 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 multi-role collaborative dynamics is then flattened into a unified reasoning sequence and distilled into a student model, effectively transforming explicit cross-role context propagation into implicit information flow, eliminating the need to re-encode prior context and yielding rich process-level supervision. ❷ **Draft–critique–revise paradigm**. To improve reasoning reliability, we introduce a *draft–critique–revise* paradigm that decouples generation from evaluation rather than relying on self-correction, which prior work shows can be unreliable without external signals [9–14]: the proposal agent first drafts an initial multi-perspective reasoning; the independent critic agent then evaluates it and provides targeted natural-language feedback that points out specific deficiencies; the proposal agent revises its reasoning from scratch under this feedback, enabling directed improvement rather than uninformed retry. ❸ **Mutual-refinement training strategy**. We further develop a *mutual-refinement* training strategy where the two agents are alternately optimized through reciprocal training signals: the proposal agent's drafting and revision abilities are jointly optimized via dual-stage reinforcement learning using the critic agent's feedback as training signal, while the proposal's revision outcomes reciprocally refine the critic—grounding its feedback quality in actual revision effectiveness.

Our contributions are summarized as follows:

- We introduce **ProCrit**, a two-agent framework that reformulates multimodal sarcasm detection from manually prescribing perspectives to self-elicited multi-perspective reasoning, where analytical perspectives are generated adaptively and integrated progressively.

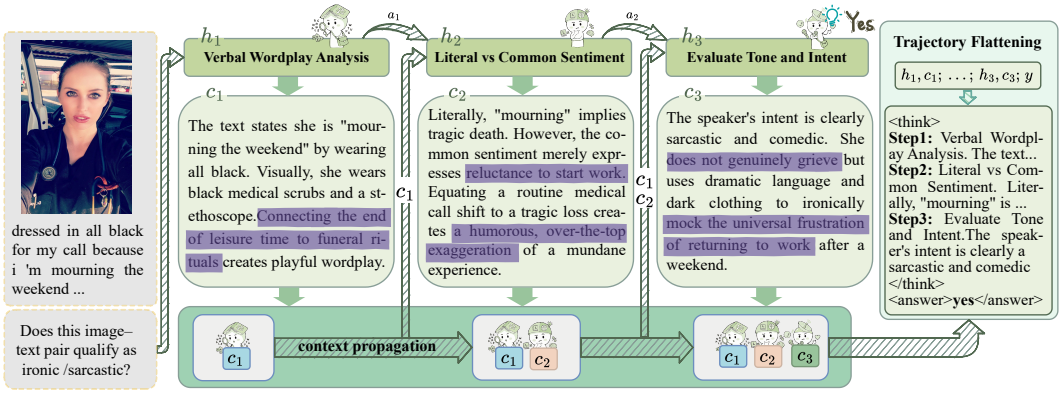


Figure 1: Illustration of the reasoning annotation synthesis process.

- We introduce a process-level reasoning annotation synthesis strategy based on *dynamic-role agentic rollout*, which generates progressive, sample-specific analytical trajectories.
- We propose a critic-guided *draft–critique–revise* paradigm with *mutual-refinement training strategy*, which improves reasoning reliability through external critique while optimizing revision and critique quality in a reciprocal optimization process.

2 Method

We develop **ProCrit**, a Proposal-Critic two-agent framework for multi-perspective reasoning in multimodal sarcasm detection, comprising a proposal agent that drafts and revises multi-perspective reasoning and a critic agent that provides external evaluation and targeted revision guidance. The framework contains three components: (1) *process-level reasoning annotation synthesis* uses a dynamic-role agentic rollout to dynamically generate analytical roles based on the findings established by preceding analyses and the remaining unaddressed aspects; (2) a *draft–critique–revise paradigm* improves reasoning reliability by providing external evaluation and targeted guidance for revision; (3) a *mutual-refinement training strategy* that alternately optimizes the two agents through reciprocal training signals: the critic agent's feedback improves the proposal agent's drafting and revision capabilities, while the proposal agent's revision outcomes reciprocally refine the critic agent.

2.1 Process-Level Reasoning Annotation Synthesis

While existing sarcasm corpora [7, 8] 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 7. To obtain process-level supervision for coherent, sample-adaptive multi-perspective reasoning, we design a *dynamic-role agentic rollout* in which strong vision-language model, denoted as $\mathcal{T}$, sequentially spawns analytical roles within a shared context.

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 agentic rollout above produces a multi-turn trajectory τ in which each role conditions on the full history of all preceding roles. However, directly using this multi-turn format

for training would require the shared multimodal context to be re-encoded at every turn, leading to redundant computation and quadratically growing input length. 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 resulting data provide rich process-level supervision that preserves multi-role analytical dependencies in a single reasoning sequence: a model trained on these sequences learns to produce all analytical perspectives in one autoregressive pass, where information from earlier roles flows to later ones through hidden states within the same forward pass, eliminating the need to re-encode prior outputs at each turn.

2.2 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 §2.3). 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 [9]. Therefore, to further 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>`$\cdots$`</think>` tags, followed by the predicted sarcasm label in `<answer>`$\cdots$`</answer>` tags.

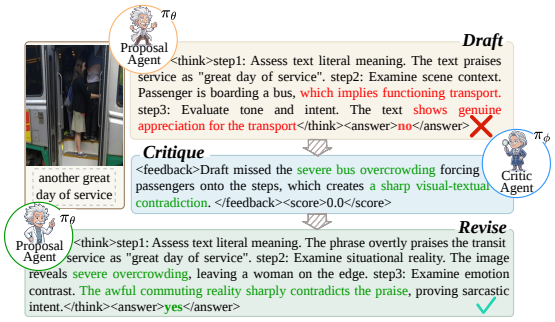


Figure 2: Illustration of the draft–critique–revise paradigm.

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>`$\cdots$`</feedback>` tags that identifies specific reasoning deficiencies and provide targeted revision guidance, and $s \in \{0, 1, 2\}$ is a quality score within `<score>`$\cdots$`<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 generates a revised reasoning from scratch under a revision prompt p_{revise} (see Appendix C.2):

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

The revised reasoning $\mathbf{d}$ follows the same structured format as the draft reasoning $\bar{\mathbf{d}}$, 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 $(\bar{d}, \hat{y}_{draft})$ remains available as the first-pass result for training and analysis.

2.3 Mutual-Refinement Training

We further develop a *mutual-refinement* training strategy that alternately optimizes the two agents, each serving as a training signal for the other. Specifically, when optimizing the proposal agent, the critic agent is frozen and provides quality scores and targeted feedback as reward signals for dual-stage reinforcement learning for drafting and revision (§2.3.2); when optimizing the critic agent, the proposal agent is frozen and its revision outcomes measure whether the critic's feedback genuinely improves reasoning (§2.3.3). Through this alternation, both agents are progressively refined: the proposal agent produces stronger reasoning, and the critic agent provides more actionable feedback.

2.3.1 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 (§2.1) 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 §3.1.

2.3.2 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. We compute group-relative advantages within each set:

$$\begin{aligned} \hat{A}_{\text{draft}}(d_i) &= \frac{r_{\text{draft}}(d_i) - \mu}{\sigma + \epsilon}, \qquad \mu = \text{mean}_{d \in \mathcal{D}}\, r_{\text{draft}}(d), \qquad \sigma = \text{std}_{d \in \mathcal{D}}\, r_{\text{draft}}(d), \\ \hat{A}_{\text{revise}}(o_{k,j}) &= \frac{r_{\text{revise}}(o_{k,j}) - \mu_k}{\sigma_k + \epsilon}, \quad \mu_k = \text{mean}_{o \in \mathcal{O}_k}\, r_{\text{revise}}(o), \quad \sigma_k = \text{std}_{o \in \mathcal{O}_k}\, r_{\text{revise}}(o). \end{aligned} \tag{9}$$

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:

$$\mathcal{L}_{\text{proposal}} = (1 - \lambda)\Phi(\mathcal{D}, \hat{A}_{\text{draft}}) + \lambda \frac{1}{K} \sum_{k=1}^{K} \Phi(\mathcal{O}_k, \hat{A}_{\text{revise}}), \tag{10}$$

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.

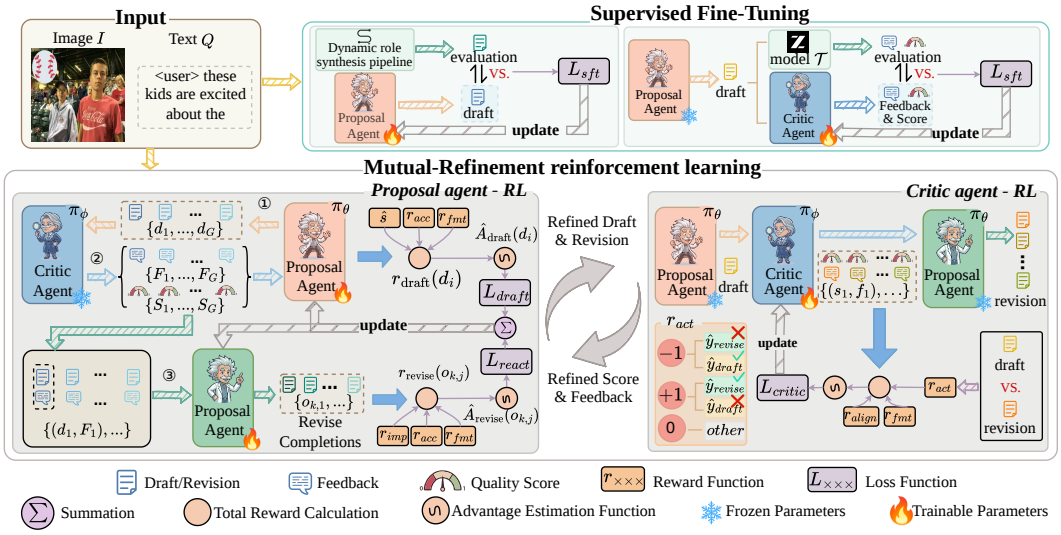


Figure 3: Overview of mutual-refinement training. **Top:** Both agents are initialized via SFT. **Bottom left:** The critic is optimized while the proposal is frozen, using revision outcomes as rewards. **Bottom right:** The proposal is optimized in dual stages (draft and revision) while the critic is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

2.3.3 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:

$$r_{\text{align}} = \begin{cases} s - 1 & \text{if } \hat{y} = y^* \\ 1 - s & \text{if } \hat{y} \neq y^*. \end{cases} \tag{11}$$

This rewards the critic for assigning high scores to correct proposals and low scores to incorrect ones.

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.

$$r_{\text{act}} = \begin{cases} +1 & \text{if } \hat{y}_{\text{draft}} \neq y^* \text{ and } \hat{y}_{\text{revise}} = y^* \quad \text{(fix)} \\ -1 & \text{if } \hat{y}_{\text{revise}} \neq \hat{y}_{\text{draft}} \text{ and } \hat{y}_{\text{revise}} \neq y^* \quad \text{(damage)} \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

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.

2.3.4 Training Pipeline

Both agents are first initialized via SFT (§2.3.1), 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.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate on three widely used multimodal sarcasm detection benchmarks: (1) **MMSD** [4], 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** [7], 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** [15], 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 [7, 16], 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 [17], with the vision encoder and multimodal projector frozen and only the language model parameters updated. For *reasoning annotations generation*, we use GLM-4.6V [18] 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 alternately optimized, with the frozen counterpart deployed via vLLM [19] 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 [20] 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*), which correspond to the labels used in Table 1. (1) *None* denotes plain chain-of-thought reasoning without explicitly specified analytical perspectives; this category includes Zero-shot CoT [21], Automatic Prompt Engineer [22], Plan-and-Solve [23], Generated Knowledge Prompting [24]. We also implement Plain CoT Reasoner, a fine-tuned variant of Zero-shot CoT, trained to produce generic chain-of-thought reasoning. (2) *Predefined* denotes reasoning with a fixed set of manually specified perspectives; this category includes IRONIC [25], S3Agent [16], Commander-GPT [26]. We additionally implement a fine-tuned variant of S3Agent trained with the three predefined perspectives, termed Fixed-role Reasoner. (3) *Adaptive* denotes sample-adaptive perspective generation where the analytical perspectives are self-elicited per instance. The main table reports both zero-shot and fine-tuned systems under these 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.

3.2 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 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

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 [21]	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 [22]	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 [23]	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 [24]	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
IRONIC [25]	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 [16]	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 [26]	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 [21]	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 [16]	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**

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†).

3.3 Ablation Study

We conduct systematic ablations on MMSD2.0 and MMSD to isolate the contribution of each component. 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 training pipeline—combining SFT and RL—consistently achieves the highest F1 scores in both Draft and Revise modes on MMSD2.0 and MMSD. Removing SFT results in the most significant performance drop, underscoring its critical role in establishing the proposal agent's sample-adaptive multi-perspective reasoning capability. Removing RL also leads to a consistent decline, indicating its complementary function in refining reasoning quality through feedback-driven optimization. Overall, SFT and RL serve complementary roles, and ablating either stage degrades performance across both generation modes.

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 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

Ablation on Draft–Revise Balance Weight λ. Table 4 presents the sensitivity of performance to the trade-off parameter λ in Eq. 10. The proposal agent achieves optimal results when both Draft and Revise modes are jointly optimized, with the best F1 score attained at $\lambda = 0.5$. In contrast, extreme settings degrade performance: $\lambda = 0$ (optimizing only the Draft mode) and $\lambda = 1$ (optimizing only the Revise mode) both yield weaker overall results. This demonstrates that balancing the two

Table 4: 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 5: 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

objectives is essential—overemphasizing either stage compromises the agent's ability to generate high-quality initial proposals and effectively refine them.

Ablation on Reward Components. (1) *Proposal agent rewards.* We ablate the reward components in both Draft and Revise modes. As shown in Table 5, in the Draft mode (Eq. 6), removing the critic evaluation reward r_{eval} leads to the largest performance drop, confirming that the critic's quality signal offers finer-grained supervision beyond binary accuracy. Ablating either the draft accuracy reward $r^{\text{d}}_{\text{acc}}$ or $r^{\text{d}}_{\text{fmt}}$ also results in consistent degradation. In the Revise mode (Eq. 7), removing r_{imp} reduces revise-stage F1 on MMSD, demonstrating that explicitly rewarding successful corrections is crucial for effective revision. (2) *Critic agent rewards.* As shown in Table 3, both r_{align} and r_{act} are essential. Removing r_{align} decreases Revise F1 from 83.1 to 82.2, while removing r_{act} drops it to 82.6. The former ensures the critic's scores correlate with actual reasoning quality; the latter encourages feedback that provides actionable guidance for the proposal agent. Additionally, ablating r^{c}_{fmt} also harms performance. The best results are achieved when all three critic rewards are used jointly, highlighting their complementary roles.

Ablation on Revise Mode Table 6 compares four revise modes with the proposal agent fixed.

Table 6: 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**

Self-revision, where the proposal agent regenerates its output without external feedback, performs substantially worse than the initial draft reasoning with no revision (None). This indicates that unguided regeneration is unreliable: in the absence of diagnostic signals, a second pass is more likely to corrupt correct reasoning than to correct errors. In contrast, incorporating external critique from an independent critic agent consistently improves performance over both baselines. Specifically, the initialized critic, obtained via supervised distillation from teacher annotations, yields measurable gains, highlighting the effectiveness of targeted natural-language feedback. Further improvement is achieved through mutual-refinement reinforcement learning, where the proposal-refined critic attains the best results. This validates our co-evolutionary design: revision outcomes from the proposal agent provide outcome-grounded signals that enable the critic to transcend mere imitation and generate more actionable and context-aware guidance. Finally, we observe that a single revision round captures the dominant improvement (see Appendix D.1).

4 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 adopts a critic-guided draft–critique–revise paradigm with a mutual-refinement training strategy that jointly optimizes reasoning and revision quality through external critique, while reciprocally refining the critic through actual revision outcomes. Extensive experiments on MMSD2.0, MMSD, and redEval demonstrate that

ProCrit achieves the strongest overall performance across all three benchmarks, and ablation studies confirm the contribution of each component. We provide further discussion on limitations and broader impact in the Appendix F and G.

References

[1] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 5, pp. 1–22, 2017.

[2] D. G. Maynard and M. A. Greenwood, "Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis," in *Lrec 2014 proceedings*. ELRA, 2014.

[3] M. Abulaish, A. Kamal, and M. J. Zaki, "A survey of figurative language and its computational detection in online social networks," *ACM Transactions on the Web (TWEB)*, vol. 14, no. 1, pp. 1–52, 2020.

[4] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 2506–2515.

[5] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, and R. 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)*, 2022, pp. 1767–1777.

[6] B. Liang, L. Gui, Y. He, E. Cambria, and R. Xu, "Fusion and discrimination: A multimodal graph contrastive learning framework for multimodal sarcasm detection," *IEEE Transactions on Affective Computing*, vol. 15, no. 4, pp. 1874–1888, 2024.

[7] L. Qin, S. Huang, Q. Chen, C. Cai, Y. Zhang, B. Liang, W. Che, and R. Xu, "Mmsd2. 0: Towards a reliable multi-modal sarcasm detection system," in *Findings of the association for computational linguistics: ACL 2023*, 2023, pp. 10 834–10 845.

[8] P. Desai, T. Chakraborty, and M. S. Akhtar, "Nice perfume. how long did you marinate in it? multimodal sarcasm explanation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 563–10 571.

[9] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou, "Large language models cannot self-correct reasoning yet," *arXiv preprint arXiv:2310.01798*, 2023.

[10] R. Kamoi, Y. Zhang, N. Zhang, J. Han, and R. Zhang, "When can llms actually correct their own mistakes? a critical survey of self-correction of llms," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1417–1440, 2024.

[11] K. Tsui, "Self-correction bench: Uncovering and addressing the self-correction blind spot in large language models," *arXiv preprint arXiv:2507.02778*, 2025.

[12] Z. Zhan, M. Cui, and R. Zhang, "Can large language models self-correct in medical question answering? an exploratory study," *arXiv preprint arXiv:2604.00261*, 2026.

[13] Y. Li, "Decomposing llm self-correction: The accuracy-correction paradox and error depth hypothesis," *arXiv preprint arXiv:2601.00828*, 2025.

[14] J. J. Arimbur, "How many tries does it take? iterative self-repair in llm code generation across model scales and benchmarks," *arXiv preprint arXiv:2604.10508*, 2026.

[15] B. Tang, B. Lin, H. Yan, and S. 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)*, 2024, pp. 1732–1742.

[16] P. Wang, Y. Zhang, H. Fei, Q. Chen, Y. Wang, J. Si, W. Lu, M. Li, and L. Qin, "S3 agent: Unlocking the power of vllm for zero-shot multi-modal sarcasm detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 11, pp. 1–16, 2025.

[17] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[18] T. Glm, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao *et al.*, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," *arXiv preprint arXiv:2406.12793*, 2024.

[19] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the 29th symposium on operating systems principles*, 2023, pp. 611–626.

[20] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "Zero: Memory optimizations toward training trillion parameter models," in *SC20: international conference for high performance computing, networking, storage and analysis*. IEEE, 2020, pp. 1–16.

[21] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[22] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," in *The eleventh international conference on learning representations*, 2022.

[23] L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, and E.-P. 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)*, 2023, pp. 2609–2634.

[24] J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, and H. Hajishirzi, "Generated knowledge prompting for commonsense reasoning," in *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 2022, pp. 3154–3169.

[25] A. A. Ramakrishnan, A. A. Ramakrishnan, and D. Lee, "Ironic: Coherence-aware reasoning chains for multi-modal sarcasm detection," *arXiv preprint arXiv:2505.16258*, 2025.

[26] Y. Zhang, C. Zou, B. Wang, J. Qin, and P. Tiwari, "Commander-gpt: Dividing and routing for multimodal sarcasm detection," *arXiv preprint arXiv:2506.19420*, 2025.

[27] T. Yue, R. Mao, H. Wang, Z. Hu, and E. Cambria, "Knowlenet: Knowledge fusion network for multimodal sarcasm detection," *Information Fusion*, vol. 100, p. 101921, 2023.

[28] H. Zhou, J. Yan, Y. Chen, R. Hong, W. Zuo, and K. Jin, "Ldgnet: Llms debate-guided network for multimodal sarcasm detection," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[29] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[30] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in neural information processing systems*, vol. 36, pp. 11 809–11 822, 2023.

[31] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[32] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Forty-first international conference on machine learning*, 2024.

[33] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. 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*, 2024, pp. 17 889–17 904.

[34] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang *et al.*, "Self-refine: Iterative refinement with self-feedback," *Advances in neural information processing systems*, vol. 36, pp. 46 534–46 594, 2023.

[35] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," *Advances in neural information processing systems*, vol. 36, pp. 8634–8652, 2023.

[36] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," in *The twelfth international conference on learning representations*, 2023.

[37] P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. 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)*, 2024, pp. 9426–9439.

[38] D. Zhang, J. Lei, J. Li, X. Wang, Y. Liu, Z. Yang, J. Li, W. Wang, S. Yang, J. 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*, 2025, pp. 9050–9061.

[39] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[40] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[41] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[42] J. Hu, Y. Zhang, Q. Han, D. Jiang, X. Zhang, and H.-Y. Shum, "Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model," *arXiv preprint arXiv:2503.24290*, 2025.

[43] T. Chu, Y. Zhai, J. Yang, S. Tong, S. Xie, D. Schuurmans, Q. V. Le, S. Levine, and Y. Ma, "Sft memorizes, rl generalizes: A comparative study of foundation model post-training," *arXiv preprint arXiv:2501.17161*, 2025.

[44] H. Shen, P. Liu, J. Li, C. Fang, Y. Ma, J. Liao, Q. Shen, Z. Zhang, K. Zhao, Q. Zhang *et al.*, "Vlm-r1: A stable and generalizable r1-style large vision-language model," *arXiv preprint arXiv:2504.07615*, 2025.

[45] W. Huang, B. Jia, Z. Zhai, S. Cao, Z. Ye, F. Zhao, Z. Xu, X. Tang, Y. Hu, and S. Lin, "Vision-r1: Incentivizing reasoning capability in multimodal large language models," *arXiv preprint arXiv:2503.06749*, 2025.

Appendix

Contents

A Related Work

A.1 Multimodal Sarcasm Detection

Early multimodal sarcasm detection methods treat the task as end-to-end classification, fusing cross-modal features through attention mechanisms [4], graph convolutional networks [5], contrastive learning [6], or external knowledge injection [27], and mapping the fused representation directly to a sarcasm label. More recent variants incorporate LLMs as feature enhancers via demonstration retrieval [15] or debate-guided text generation [28]. While these methods have steadily improved detection accuracy, their analytical process remains implicit in the learned representations, with no explicit reasoning over the input. A growing line of work makes the analytical process explicit by prompting LLMs and VLMs to produce intermediate reasoning alongside predictions. S3 Agent [16] assigns three fixed analytical perspectives and Commander-GPT [26] routes inputs through six sub-task agents, each aggregated by a decision module, while IRONIC [25] predicts predefined discourse relations to guide interpretive reasoning. However, these methods rely on predefined perspectives that operate independently, limiting adaptability to diverse sarcasm mechanisms and preventing later analyses from building on earlier findings. Our work instead realizes *self-elicited multi-perspective reasoning*, where the model dynamically generates sample-adaptive perspectives and progressively integrates them into a coherent analysis, with an independent critic agent providing external evaluation and targeted revision guidance.

A.2 Reasoning and Self-Refinement

Chain-of-thought prompting [29] and its extensions—Tree-of-Thoughts [30], self-consistency decoding [31], and multi-agent debate [32, 33]—have established that explicit step-by-step reasoning and diverse-path exploration substantially improve LLM performance. A related line of work explores *self-refinement*: Self-Refine [34] and Reflexion [35] demonstrate that iterative generate-critique-revise loops can improve output quality at inference time. However, a growing body of evidence [9–13] indicates that *intrinsic* self-correction without external signals remains fundamentally unreliable—motivating the use of a *independent* critic rather than self-evaluation. Existing critic models provide scalar quality signals but not revision guidance: process reward models [36] and Math-Shepherd [37] score reasoning steps, and Critic-V [38] extends this to VLM reasoning, yet none produces natural-language feedback. Our critic agent generates both calibrated scores and actionable feedback, with feedback quality explicitly optimized via RL to improve the proposal agent's revision outcomes.

A.3 Reinforcement Learning for LLM Reasoning

RLHF [39] is the standard post-training alignment paradigm. For reasoning, Group Relative Policy Optimization (GRPO) [40] eliminates the value model by computing group-relative advantages, and has been scaled in DeepSeek-R1 [41] and Open-Reasoner-Zero [42]. Zhai et al. [43] show that SFT stabilizes format while RL drives generalization, supporting SFT-then-RL pipelines. Recent work extends GRPO-style training to vision-language models [44, 45] for multimodal reasoning. However, existing RL approaches optimize single-pass generation with no opportunity to revise, and rely on outcome-based or scalar process rewards without natural-language feedback. Our dual-stage GRPO (§2.3.2) jointly optimizes draft generation and feedback-driven revision within a single model, with a frozen critic supplying both scalar scores and natural-language feedback during training.

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 mechanisms and require different analytical perspectives to expose the underlying incongruity (see Fig. 6), 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).

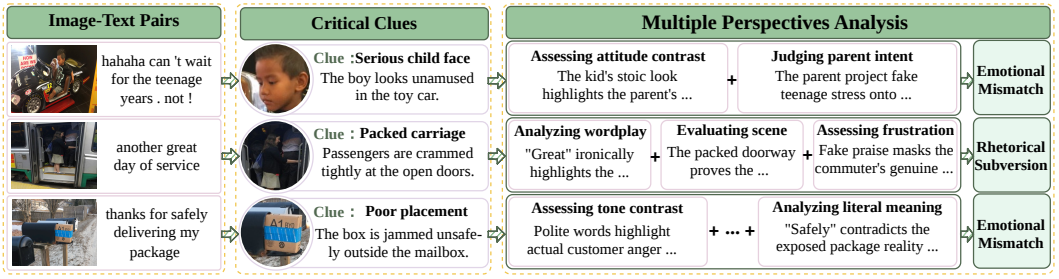


Figure 6: Examples illustrating that different sarcastic samples involve distinct mechanisms and demand different combinations of analytical perspectives to expose the text–image incongruity.

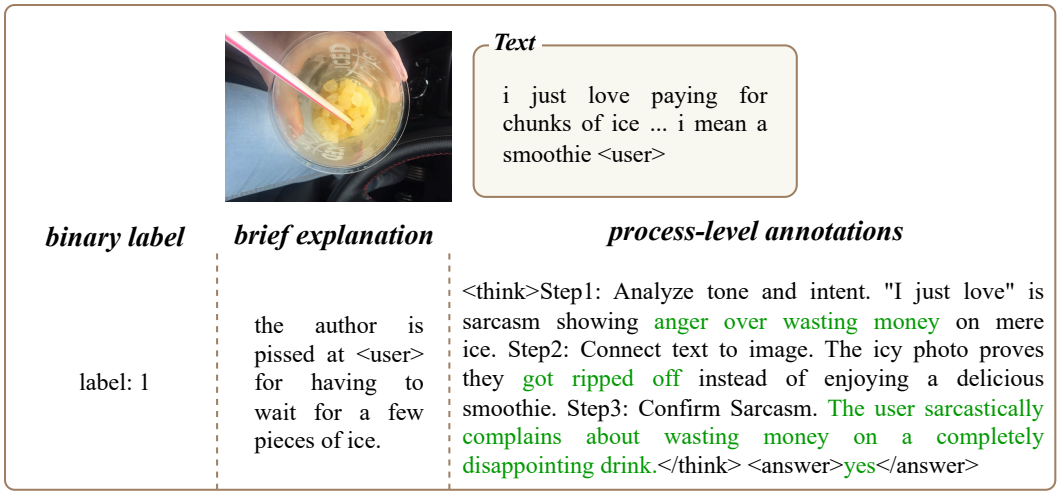


Figure 7: Comparison between existing dataset annotations and the process-level reasoning annotations synthesized by ProCrit.

B.1 Dynamic-Role Rollout Details

While existing sarcasm corpora [7, 8] 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 7, the dynamic-role agentic rollout (§2.1) addresses this by synthesizing multi-perspective, context-aware reasoning annotations from raw image-text samples. Specifically, we use GLM-4.6V [18] 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 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 6 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 6: 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.

B.3 Synthesis Examples

Figures 8 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 §3.1. In this section, we provide more information about the hyper-parameters used in our experiment. Tables 7 and 8 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 7: 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 8: 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 Prompts

This subsection provides the complete prompts used in reasoning annotation synthesis (§2.1), proposal agent inference (drafting and revision), and critic evaluation, as referenced throughout §2.2.

Reasoning annotation synthesis. During the dynamic-role agentic rollout (§2.1), 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.

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
  - 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}
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}

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>

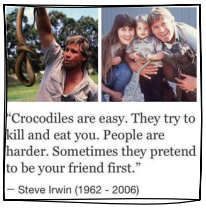
Text

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 8: 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.



Text

steve irwin spoke the truth

GT label

sarcasm

Draft

<think>Step 1: Analyze the image context. The image features Steve Irwin (a well-known wildlife activist) handling dangerous animals (snakes, crocodiles) alongside family photos, establishing his reputation as someone who interacts with harmful creatures while also being a caring family figure.Step 2: Evaluate the text "steve irwin spoke the truth". This statement directly affirms a claim about Steve Irwin's honesty. There is no visible contrast between the text and the image that would suggest sarcasm (where words mean the opposite of their literal sense) or irony (where expectations are subverted). The text aligns with the positive portrayal of Steve Irwin in the image, reinforcing sincerity rather than mockery.</think><answer>no</answer>

Critique

<feedback>The reasoning overlooked the crucial contextual irony provided by the quote Steve Irwin attributed to himself. The statement "Crocodiles are easy. They try to kill and eat you. People are harder. Sometimes they pretend to be your friend first" is a darkly humorous observation about human nature, delivered by someone famous for his close encounters with dangerous animals. The sarcasm lies in the contrast between the literal meaning of the quote (about crocodiles being easier to deal with than people) and the implied, bitter truth about humanity. The reasoning failed to recognize this layer of meaning, treating the quote as a straightforward, literal statement when it's actually a wry, cynical remark. By focusing only on the surface-level alignment between the image (Steve Irwin with animals) and the text ("speoke the truth"), it missed the deeper, ironic subtext that defines the sarcasm.</feedback>

Revise

<think>Step 1: Visual context — contrasting danger and warmth. The image shows Steve Irwin handling deadly snakes and crocodiles alongside family photos, juxtaposing extreme danger with familial love and trust — creating an ironic contrast between what humans can endure and what they'll let near them.Step 2: Quote analysis — dark humor about human nature. Steve Irwin's quote frames crocodiles as "easy" (predictable, dangerous) compared to people who "pretend to be your friend first." This wry, cynical observation about human deception and betrayal is inherently ironic — it uses a straightforward animal comparison to deliver a bitter commentary on humanity.Step 3: Image-text interaction creates sarcasm. The caption "steve irwin spoke the truth" frames the quote as wisdom. However, the quote itself is a dark joke: crocodiles are objectively more dangerous than people, yet Irwin implies people are worse. The sarcasm lies in the gap between the literal praise of the quote and the implied bitter truth — that people are trickier than the deadliest creatures. The image of gentle family moments amplifies this by highlighting how easily humans deceive even those we love, reinforcing the ironic subtext.</think><answer>yes</answer>

Text

that , i just roasted a dude in intellectual conversation-type chill

GT label

non-sarcasm

Draft

<think>Step 1: Analyze the text for tone indicators. The phrase "roasted a dude" implies a playful or teasing remark, which can signal sarcasm if the speaker intends to mock or convey insincerity.Step 2: Consider the context of "intellectual conversation-type chill." This suggests a formal or serious setting, but pairing it with "roasted" creates a contrast—sarcasm often involves saying something contrary to literal meaning, especially in contexts where seriousness is expected.Step 3: Evaluate the image's role. The neutral-to-positive expression (smiling, winking) may imply the speaker is being facetious rather than genuinely serious, reinforcing the likelihood of sarcasm.</think><answer>yes</answer>

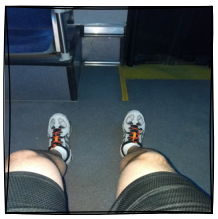
Critique

<feedback>The reasoning incorrectly identifies sarcasm where none exists. It misinterprets the casual phrase "roasted a dude" as automatically sarcastic, when it likely refers to a friendly joke or playful banter. The reasoning also wrongly assumes the image's neutral expression indicates sarcasm, failing to recognize that the image doesn't provide clear visual cues of irony or contradiction. The analysis makes unsupported leaps about the contrast between "roasted" and "intellectual conversation" as creating sarcasm, when this could simply be describing a casual interaction. The reasoning should have focused on whether there's actual irony or contradiction between the text and image, rather than assuming any contrast equals sarcasm.</feedback>

Revise

<think>Step 1: Text tone assessment. "That, I just roasted a dude in intellectual conversation-type chill" is a straightforward, boastful statement about having a playful argument with someone intellectually. There are no linguistic markers of sarcasm such as exaggeration, contradiction, or mocking tone.Step 2: Image-text alignment. The image shows Drake smiling confidently while flossing his teeth—a casual, self-assured pose. This complements the boastful text rather than contradicting it; both convey confidence and a laid-back attitude consistent with someone who just engaged in a spirited, intellectual exchange.Step 3: Sarcasm evaluation. Sarcasm requires a detectable gap between literal meaning and intended meaning, or between image and text. Here, the text describes an event and the image shows the person reacting proudly to it. The casual tone of "dude" and "type chill" reflects informal bragging, not ironic reversal. No contradiction exists between what is said and what is meant.</think><answer>no</answer>

Figure 9: 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 10: 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 mechanism.

D Additional Experiments

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

D.1 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 9 reports the results.

Table 9: 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.2 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 10 summarizes the data scale and approximate GPU hours for each stage. The total training budget is approximately 20 GPU hours. 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 [41], reflecting the efficiency of targeted, task-specific optimization.

Table 10: 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 9), 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 8). In practice, the average draft length is considerably shorter: the majority of reasoning trajectories comprise 3 analytical steps (Table 6), 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

This section presents additional examples to illustrate the behavior of the draft–critique–revise pipeline on concrete samples from the MMSD2.0 test set. As shown in Figure 9, the critic agent is able to identify specific reasoning deficiencies in incorrect drafts—such as misinterpreted pragmatic cues or overlooked contextual irony—and provide targeted feedback that guides the proposal agent toward a corrected analysis generated from scratch. As demonstrated in Figure 10, the critic also contributes when the draft already arrives at the correct prediction but with insufficiently grounded reasoning: it pinpoints analytical gaps, prompting the proposal agent to produce a substantially deeper revision that explicitly anchors its judgment in concrete multimodal evidence. These cases highlight that the critic serves a dual role—not only as an error-correction mechanism but also as a quality-enhancement signal that strengthens the analytical rigor of the reasoning process.

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.2 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.

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.