IRAC: RETRIEVAL-AUGMENTED MULTI-AGENT COLLABORATION FOR MULTIMODAL SARCASM DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Sarcasm Detection (MSD) aims to identify ironic intent in image–text pairs, where sarcasm typically arises from inconsistencies or semantic incongruity between linguistic and visual cues. However, most MSD research still relies on coarse binary labels, and even recent efforts that enrich annotations seldom identify which modality drives the irony—that is, its underlying source. This limitation leads models to fit shallow statistical correlations rather than understanding deep cross-modal interactions, while traditional supervised methods suffer from restricted generalization in open-world scenarios. In this paper, we present **MIRS**, a large-scale benchmark for **M**ultimodal **I**rony **R**easoning and **S**ource-attribution, containing 11,000 samples with fine-grained four-layer hierarchical annotations. To tackle the demands of zero-shot explainable analysis—requiring both precise detection and human-aligned diagnostics—we propose **IRAC** (**I**rony-oriented **R**etrieval-**A**ugmented multi-agent **C**ollaboration). IRAC employs a specialized swarm of agents (Textual, Visual, and Cross-modal) and an orchestrator optimized via Group Relative Policy Optimization (GRPO) with irony-specific rewards. Experimental results demonstrate that IRAC achieves state-of-the-art performance (86.92% Acc) and significantly reduces hallucination rate, providing a new paradigm for explainable multimodal irony analysis. Our code and benchmark are available at `https://anonymous.4open.science/r/IRAC-1274778787`.

1 INTRODUCTION

Sarcasm is a sophisticated pragmatic phenomenon in which the intended meaning deviates from its literal surface form (Lu et al., 2025). Interpreting such expressions often requires identifying inconsistencies between textual and visual evidence (Wu et al., 2025; Zhao et al., 2026a). For example, the statement "What a wonderful day" paired with an image of a flooded street conveys frustration through **cross-modal incongruity**. As such expressions are pervasive in social media, political discourse, and internet memes, multimodal sarcasm detection (MSD) has become an important problem for sentiment analysis, opinion mining, and content moderation (Castro et al., 2019; Farabi et al., 2024).

Unlike conventional sentiment analysis, MSD is particularly challenging because sarcastic meaning often emerges not from either modality alone, but from their **synergistic interaction** (Cai et al., 2019; Liang et al., 2022; Qiao et al., 2023). Recent Multimodal Large Language Models (MLLMs) such as Qwen3-VL (Bai et al., 2025) and InternVL3 (Zhu et al., 2025) have demonstrated strong zero-shot multimodal reasoning capabilities, yet current approaches still suffer from three major limitations.

First, **Attribution Ambiguity in Benchmarks**. Many existing datasets formulate MSD as a binary classification task and primarily provide coarse sarcasm labels, without identifying the underlying source of irony. In practice, sarcasm may arise from linguistic irony, visual metaphors, or cross-modal incongruity (see Figure 1). Without fine-grained supervision, models tend to exploit superficial modality-specific shortcuts rather than learning interpretable reasoning patterns (Qin et al., 2023). Although recent efforts enrich sarcasm annotation with explanations (Desai et al., 2022; Kumar et al., 2022), mockery targets (Wang et al., 2022), conversational triggers (Singh et al., 2024), or

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

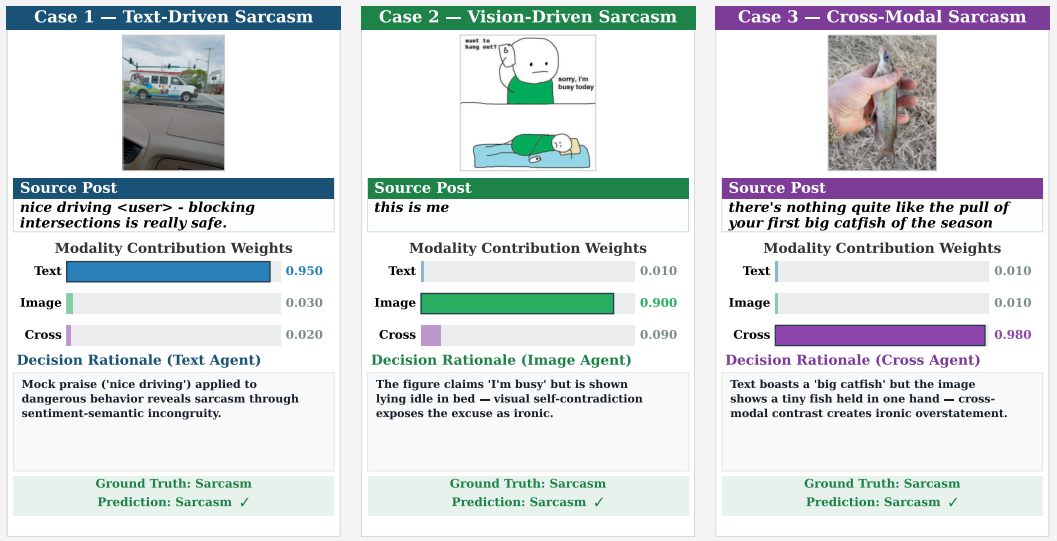

Figure 1: Case study of modality attribution and decision rationales produced by IRAC. Case 1 is dominated by textual sarcasm, Case 2 by visual incongruity, and Case 3 by cross-modal inconsistency.

multi-image and modality-level settings (Zhao et al., 2026a; Scott et al., 2026), explicit modality attribution—identifying which modality drives the sarcasm—remains largely unexplored.

Second, **Insufficient Pragmatic Knowledge**. Zero-shot MLLMs often lack the cultural and contextual knowledge needed to interpret memes, social norms, and event-specific references. While retrieval-augmented generation (RAG) (Lewis et al., 2020) has proven effective for knowledge-intensive NLP tasks, its application to multimodal sarcasm detection remains underexplored.

Third, **Monolithic Inference Paradigms**. Existing methods typically rely on single-pass prompting, requiring the model to jointly analyze textual, visual, and cross-modal evidence. This contrasts with human reasoning, which naturally decomposes the problem into modality-specific judgments before synthesizing a final decision. Although multi-agent reasoning (Wei et al., 2026) and reinforcement learning (Guo et al., 2025; Zhou et al., 2025) have improved complex reasoning, existing methods do not explicitly optimize how evidence from different modalities should be weighted.

To address these challenges, we introduce **MIRS**, a benchmark of 11,000 image–text pairs with four-layer annotations, including global sarcasm labels, modality attribution labels, and natural language rationales. MIRS supports the joint evaluation of classification accuracy and explanation faithfulness.

We further propose **IRAC**, a retrieval-augmented multi-agent framework for explainable MSD. As illustrated in Figure 2, IRAC consists of: (1) a retrieval module that supplements external contextual knowledge; (2) three specialized agents that independently analyze textual, visual, and cross-modal evidence; and (3) an orchestration agent that predicts modality attribution weights to aggregate the agent outputs. To align the orchestration process with human reasoning, we design **GRASP** (**G**eneralized **R**eward for **A**ttribution and **S**arcasm **P**lanning), a reinforcement learning objective that jointly optimizes prediction correctness, confidence calibration, attribution correctness, and reasoning consistency via Group Relative Policy Optimization (GRPO).

Experiments show that IRAC consistently outperforms state-of-the-art MLLMs in both classification and rationale quality. Retrieval augmentation improves contextual understanding, multi-agent decomposition produces more faithful explanations, and GRASP significantly reduces hallucinations while improving attribution consistency.

Our contributions are summarized as follows. **(1)** We introduce **MIRS**, a large-scale MSD benchmark that provides explicit modality attribution annotations together with hierarchical rationales, being among the first to combine large-scale coverage with source-level attribution and layered rationales in a single resource. **(2)** We propose **IRAC**, a retrieval-augmented multi-agent framework that refor-

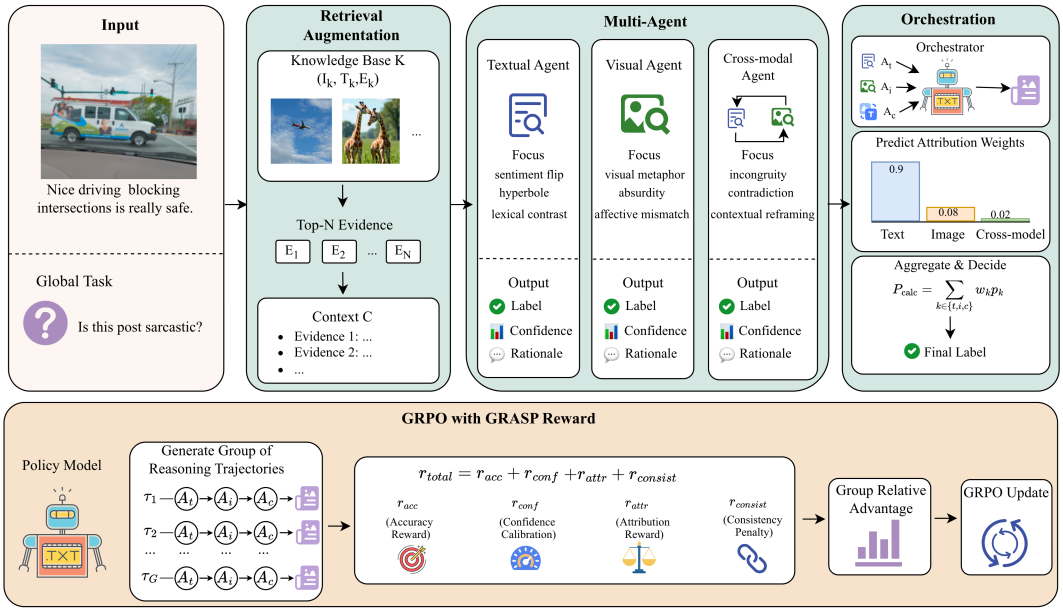

Figure 2: Overview of the IRAC framework. Given an image-text post, IRAC first retrieves external contextual knowledge to provide additional pragmatic evidence. Three specialized agents then independently analyze textual cues, visual evidence, and cross-modal incongruity, producing predictions, confidence scores, and natural language rationales. An orchestrator aggregates these outputs by predicting modality attribution weights over the agents to generate the final sarcasm label and explanation. During training, the orchestrator is optimized with GRPO using the proposed GRASP reward, which jointly considers prediction accuracy, confidence calibration, attribution correctness, and reasoning consistency.

mulates MSD as attribution-aware collaborative reasoning. **(3)** We design **GRASP**, a reinforcement learning objective that improves attribution reliability and reasoning interpretability for multimodal sarcasm detection. **(4)** IRAC achieves state-of-the-art performance on MIRS (86.92% Accuracy) and produces more faithful rationales than strong MLLM baselines.

2 Related Work

2.1 Multimodal Sarcasm Detection

Multimodal sarcasm detection aims to identify sarcastic intent by jointly reasoning over text and images, building on foundational datasets that established the importance of visual context (Castro et al., 2019; Cai et al., 2019). Subsequent research has centered on modeling cross-modal incongruity—the core signal underlying multimodal sarcasm—through graph-based interaction modeling (Liang et al., 2022), incongruity-aware fusion (Qiao et al., 2023), contrastive debiasing (Jia et al., 2024), and more recent advances in adaptive fusion, information bottleneck learning, and hierarchical chain-of-thought reasoning (Wu & Zang, 2025; Wang et al., 2026; Zhuang et al., 2026; Zhao et al., 2026b). Despite strong classification accuracy, most of these methods remain supervised black-box classifiers, offering limited insight into which modality or cross-modal interaction triggers the sarcastic meaning.

2.2 Multimodal Sarcasm Detection Benchmarks

The development of MSD has been supported by increasingly realistic datasets and benchmarks. Early works such as Castro et al. (2019) established widely used image–text sarcasm benchmarks, while later datasets introduced richer annotations and more complex settings. For example, MMSD3.0 (Zhao et al., 2026a) extends sarcasm detection to multi-image scenarios, and MuSaG (Scott et al., 2026) provides full-modal (audio–visual–text) annotations for explainable sarcasm analysis. These benchmarks

highlight the growing importance of explainability and source attribution in multimodal sarcasm detection.

Beyond binary classification, several efforts move toward finer-grained sarcasm understanding: MORE (Desai et al., 2022) and its target-augmented extension (Goel et al., 2025) generate natural-language explanations of why a post is sarcastic; MSTI (Wang et al., 2022) localizes the specific target being mocked; and in the conversational setting, WITS (Kumar et al., 2022) explains sarcasm in multi-party dialogues while SIRD (Singh et al., 2024) annotates the trigger and initiation of sarcasm. In contrast, IRAC predicts an explicit, quantitative *modality-attribution* distribution over textual, visual, and cross-modal sources, indicating *which modality drives* the sarcastic reading of each standalone image–text post—to our knowledge, the first framework to model such dynamic attribution weights and optimize them end-to-end through a reinforcement-learning reward.

2.3 MLLMs for Sarcasm Detection

MLLMs have recently been applied to zero-shot multimodal sarcasm detection. Basnet et al. (2025) systematically evaluated open-source MLLMs and found that they still struggle with subtle pragmatic reasoning. To improve performance, IRONIC (Ramakrishnan et al., 2025) introduces coherence-aware reasoning chains, S^3-MSD (Zhu et al., 2026) leverages MLLMs for explainable and generalizable sarcasm detection, and SEVADE (Liu et al., 2026) adopts a self-evolving multi-agent framework to reduce hallucinations. However, most of these methods do not explicitly model the relative contribution of text, image, and cross-modal evidence.

3 Methodology

In this section, we first formalize the sarcasm modality attribution task and then detail the architectural components of our IRAC framework.

3.1 Task Formulation

Given a multimodal dataset $\mathcal{D} = \{(I, T, y)\}$, the objective is to predict a global sarcasm label $y \in \{0, 1\}$ through a hierarchical reasoning process. For each irony source $k \in \{t, i, c\}$ (textual, visual, and cross-modal), the system first generates a modality-specific tuple $\langle \hat{y}_k, p_k, R_k \rangle$, consisting of a local judgment, confidence score, and rationale. Subsequently, these analytical outputs are aggregated by an orchestrator to produce the final system response: (1) a global prediction $\hat{y}$; (2) a weight vector $\mathbf{w}$ indicating the contribution of each source; and (3) a consolidated rationale R_{orch} that justifies the final decision. We note that the fine-grained modality attribution annotations in MIRS are *not* consumed as direct supervision targets for $\mathbf{w}$: the global label y is the only signal optimized against during training, and the attribution weights are learned indirectly through the GRASP reward (Section 3.5) rather than fit to gold attribution labels, which instead serve as a held-out resource for evaluating attribution quality (e.g., the Attribution Consistency metric).

3.2 Framework Overview

IRAC consists of three stages: knowledge retrieval, specialized reasoning, and collaborative orchestration. First, a retrieval module augments the input with external contextual knowledge. Then, three specialized agents $\mathcal{A} = \{A_t, A_i, A_c\}$ independently analyze textual cues, visual evidence, and cross-modal inconsistency, respectively. Finally, an orchestrator A_{orch} predicts modality attribution weights over the three agents and aggregates their outputs to produce the final prediction $\hat{y}$ and rationale R_{orch}.

3.3 Sarcasm Knowledge Retrieval

Given that irony often relies on implicit cultural tropes or memes, we utilize a knowledge base $\mathcal{K} = \{(I_m, T_m, E_m)\}_{m=1}^{M}$ to provide external evidence, where each entry pairs an image I_m, its text T_m, and a structured sarcasm explanation E_m. The explanations E_m are generated by a strong MLLM (DeepSeek-V4) and subsequently verified by human annotators to ensure faithfulness; the resulting knowledge base contains $M \approx 5{,}000$ high-confidence entries (see Appendix A.5). Crucially,

$\mathcal{K}$ is strictly disjoint from the evaluation set, so the retrieved evidence C serves purely as external pragmatic context rather than a shortcut to the label. For a query $X_q = (I_q, T_q)$, we retrieve contextual evidence C by computing a hybrid similarity score s_m against each entry m:

$$\begin{aligned} s_m = &\lambda \cdot \mathrm{sim}\big(\phi_i(I_q), \phi_i(I_m)\big) \\ &+ (1-\lambda) \cdot \mathrm{sim}\big(\phi_t(T_q), \phi_t(T_m)\big), \end{aligned} \tag{1}$$

where ϕ_i and ϕ_t are CLIP-based visual and textual encoders (Radford et al., 2021), $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, and $\lambda \in [0,1]$ is a hyperparameter that balances the importance of visual and textual relevance. The retrieved top-N structured explanations $\{E_m\}$ are concatenated to form the context C, allowing downstream agents to reason with few-shot pragmatic knowledge. In practice, each specialized agent is supplied with context retrieved under a modality-appropriate emphasis (i.e., a different setting of λ): A_t receives text-dominant retrieval, A_i vision-dominant retrieval, and A_c balanced joint retrieval (see the agent prompts in the Appendix).

3.4 Specialized Agent Swarm

The specialized agents operate in parallel, each focusing on a restricted input subspace to ensure modular reasoning.

3.4.1 Textual Reasoning Agent (A_t)

A_t serves as a pragmatic linguistic analyzer tasked with identifying internal linguistic dissonance within T. It interrogates stylistic markers such as: (1) Sentiment Polarity Flip; (2) Rhetorical Devices (e.g., hyperbole, rhetorical questions); and (3) Lexical Contrast. The agent outputs a tuple $\langle \hat{y}_t, p_t, R_t \rangle$.

3.4.2 Visual Reasoning Agent (A_i)

A_i functions as a semiotic interpreter of the image I. Moving beyond standard object detection, it decodes visual metaphors and compositional paradoxes by analyzing: (1) Absurd Juxtaposition; (2) Affective Mismatch; and (3) Visual Hyperbole. It produces a tuple $\langle \hat{y}_i, p_i, R_i \rangle$.

3.4.3 Cross-modal Interaction Agent (A_c)

A_c identifies sarcasm arising from inconsistencies between the image I and text T. To bridge the modality gap, it first selects a complementary cross-modal description d^* from a candidate set $\mathcal{D}_{cand}$ by maximizing:

$$\begin{aligned} d^* = \arg\max_{d \in \mathcal{D}_{cand}} \Big[&\alpha \cos(\mathbf{v}_I, \mathbf{t}_d) + \mu \cos(\mathbf{t}_T, \mathbf{t}_d) \\ &- \gamma \, \mathrm{Jaccard}(T, d) \Big]. \end{aligned} \tag{2}$$

where $\mathbf{v}_I = \mathrm{CLIP}_V(I)$ denotes the image embedding, $\mathbf{t}_T = \mathrm{CLIP}_T(T)$ and $\mathbf{t}_d = \mathrm{CLIP}_T(d)$ denote the text embeddings of the original post and candidate description, respectively, and α, μ, γ are non-negative balancing coefficients. The objective jointly encourages visual alignment, semantic relevance, and lexical complementarity. Specifically, the Jaccard penalty discourages redundant descriptions that simply restate the textual content. The selected description d^* is then used for conflict analysis between visual and textual semantics, producing $\langle \hat{y}_c, p_c, R_c \rangle$. The candidate set is not hand-crafted: for each input, a small pool of $|\mathcal{D}_{cand}| = 3$ candidate cross-modal tension descriptions is synthesized by GPT-5.4 (generation and analysis templates in Appendix, Figures 11 and 12).

3.5 Reinforcement-Trained Orchestration

The orchestrator agent A_{orch} acts as a dynamic router that aggregates the specialized judgments. It predicts the weight vector $\mathbf{w}$ to compute the aggregated confidence $P_{calc} = \sum_{k \in \{t,i,c\}} w_k p_k$, and *independently* emits a discrete final decision $\hat{y} \in \{0,1\}$ as part of its structured output. Rather than deriving $\hat{y}$ deterministically from P_{calc}, we let the orchestrator generate $\hat{y}$ from its holistic reasoning; the consistency penalty (Eq. 9) then encourages agreement between $\hat{y}$ and P_{calc} while still allowing the orchestrator to override the weighted vote when warranted. Since $\hat{y}$ is not confined to the convex

hull of the agent probabilities, the orchestrator can reject even a unanimous but hallucinated agent consensus. To optimize A_{orch} without direct supervision of $\mathbf{w}$, we utilize Group Relative Policy Optimization (GRPO) (Guo et al., 2025) using the GRASP reward r_{total}:

$$r_{total} = r_{acc} + r_{conf} + r_{attr} + r_{consist}. \tag{3}$$

1. Accuracy Reward (r_{acc}): To mitigate the common "Always-Yes" bias in sarcasm detection, we employ a weighted accuracy reward:

$$r_{acc} = (2 \cdot \mathbb{I}[\hat{y} = y] - 1) \cdot \mathcal{W}_{acc} \cdot c_{class}, \tag{4}$$

where $c_{class} = 1.5$ if the ground truth is non-ironic ($y = 0$) and 1.0 otherwise, and $\mathcal{W}_{acc} = 1.0$. This asymmetry encourages the model to be more discriminative rather than over-predicting irony.

2. Confidence Calibration (r_{conf}): Due to the inherent subjectivity of sarcasm, we incorporate confidence scores to provide smoother optimization signals beyond binary correctness rewards. Following ***ReflectRM*** (Qin et al., 2026), we derive the raw confidence c_{raw} from the lowest 10% of token log-probabilities in the orchestrator-generated rationale R_{orch}. Letting $\mathcal{B}$ denote this bottom-decile token set, we define

$$c_{raw} = \exp\left(\frac{1}{T\,|\mathcal{B}|}\sum_{j\in\mathcal{B}}\log p_j\right), \quad T = 5.0, \tag{5}$$

so that $c_{raw} \in (0, 1]$, where T is a temperature that smooths the estimate. Here R_{orch} denotes the concatenation of the orchestrator's `analysis` and `explanation` fields (see the output schema in the Appendix). This estimator focuses on the most uncertain reasoning steps while remaining robust to isolated outlier tokens.

To suppress "shallow heuristics," where the model assigns high confidence based on superficial cues, we scale the confidence by a reflection-depth factor $\rho = \text{len}(R_{orch})/L_{\text{target}}$, where $\text{len}(\cdot)$ counts the words in R_{orch} and $L_{\text{target}} = 200$:

$$c_{scaled} = 0.5 + (c_{raw} - 0.5) \cdot \min(1.0, \rho). \tag{6}$$

The reward is defined as:

$$r_{conf} = \begin{cases} (c_{scaled} - 0.5) \cdot \mathcal{W}_{conf}, & \text{if } \hat{y} = y, \\ (0.5 - c_{scaled}) \cdot \mathcal{W}_{conf}, & \text{otherwise.} \end{cases} \tag{7}$$

This design encourages confidence estimates that are both well-calibrated and supported by sufficiently detailed reasoning, with $\mathcal{W}_{conf} = 1.0$.

3. Attribution Reward (r_{attr}): This term rewards the orchestrator for correctly identifying which agent provided the most accurate evidence. Let $\mathcal{S} = \{t, i, c\}$ and $\mathbb{I}_k = \mathbb{I}[\hat{y}_k = y]$. The reward is formulated as:

$$r_{attr} = \left(\sum_{k\in\mathcal{S}} w_k \mathbb{I}_k - \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} \mathbb{I}_k\right) \cdot \mathcal{W}_{attr}. \tag{8}$$

By subtracting the average agent performance (uniform voting) from the orchestrator's weighted performance, we isolate the specific *advantage* gained from dynamic weight allocation. We set $\mathcal{W}_{attr} = 0.4$.

4. Consistency Penalty ($r_{consist}$): To ensure logical coherence, we penalize trajectories where the orchestrator's independently generated decision $\hat{y}$ contradicts the weight-aggregated confidence P_{calc}. Let $\Delta P = P_{calc} - 0.5$ and $\Delta\hat{y} = \hat{y} - 0.5$ denote the signed deviations from the decision boundary. The penalty is:

$$r_{consist} = -0.5 \cdot \mathbb{I}\big(|\Delta P| > \delta \wedge \Delta P \cdot \Delta\hat{y} \leq 0\big), \tag{9}$$

where $\delta = 0.05$. Since $\hat{y}$ is produced by the orchestrator rather than thresholded from P_{calc}, the two quantities can genuinely disagree; this term suppresses such logical dissonance, forcing the generated decision to be numerically aligned with the model's inner evidence accumulation.

Model	Modality			Performance Metrics			
	A	V	T	Acc	Macro-F1	Precision	Recall
Closed-source MLLMs							
GPT-4o (CoT)	✓	✓	✓	0.8162	0.8133	0.7431	0.9563
GPT-4o (CoT+KB)	✓	✓	✓	0.8608	0.8608	0.8385	0.8872
Claude Sonnet 4.5 (CoT)	×	✓	✓	0.7866	0.7845	0.7283	0.9016
Claude Sonnet 4.5 (CoT+KB)	×	✓	✓	0.8017	0.7999	0.8545	0.7184
Claude Haiku 4.5 (CoT)	×	✓	✓	0.6813	0.6590	0.6127	0.9542
Claude Haiku 4.5 (CoT+KB)	×	✓	✓	0.7443	0.7396	0.8133	0.6218
Open-source MLLMs							
Qwen3-VL-4B (CoT)	×	✓	✓	0.6120	0.5546	0.5604	**0.9857**
Qwen3-VL-4B (CoT+KB)	×	✓	✓	0.7733	0.7732	0.7530	0.8019
InternVL3-8B (CoT)	×	✓	✓	0.5957	0.5337	0.5492	0.9795
InternVL3-8B (CoT+KB)	×	✓	✓	0.6908	0.6803	0.6300	0.8912
Qwen3-VL-235B-A22B (CoT)	×	✓	✓	0.6643	0.6310	0.5960	0.9822
Qwen3-VL-235B-A22B (CoT+KB)	×	✓	✓	0.8191	0.8190	0.8119	0.8224
Qwen3-VL-32B (CoT)	×	✓	✓	0.6980	0.6774	0.6241	0.9679
Qwen3-VL-32B (CoT+KB)	×	✓	✓	0.8256	0.8253	0.8348	0.8040
Other Baselines							
DIP (Wen et al., 2023)	×	✓	✓	0.8103	0.8102	0.8104	0.8101
Multi-view CLIP (Qin et al., 2023)	×	✓	✓	0.8504	0.8483	**0.9373**	0.7452
IRONIC (Ramakrishnan et al., 2025)	×	✓	✓	0.8414	0.8413	0.8434	0.8313
LVLM-Sarcasm (Chen et al., 2025)	×	✓	✓	0.8312	0.8273	0.7678	0.9498
Commander-GPT (Zhang et al., 2025)	×	✓	✓	0.8464	0.8461	0.8063	0.9044
ProCrit (Xu et al., 2026)	×	✓	✓	0.7285	0.7284	0.7234	0.7244
IRAC (Base)	×	✓	✓	0.8579	0.8579	0.8402	0.8776
IRAC (GRPO)	×	✓	✓	**0.8692**	**0.8692**	0.8648	0.8695

Table 1: **Classification results on the MIRS dataset. Bold** and **green** indicate the best performance. The blue background highlights our IRAC framework. The **Modality** columns indicate each model's inherent modality support—**A**: Audio; **V**: Visual; **T**: Textual. CoT: Chain-of-Thought; KB: Knowledge Base.

To optimize the orchestrator policy π_θ using these rewards, we employ the GRPO algorithm (Guo et al., 2025). For each query, GRPO normalizes the total GRASP reward of each sampled trajectory against the mean of a group of G rollouts to form the relative advantage, and maximizes a clipped surrogate objective regularized by a KL penalty toward the reference model π_{ref}. This group-based normalization reduces the variance of the gradient estimate while limiting excessive policy drift. The full optimization objective is provided in Appendix A.3.

4 Experiments

4.1 MIRS Benchmark and Data Collection

We evaluate IRAC on our newly constructed **MIRS** benchmark. We aggregated 87,980 raw candidates from **Weibo** (80+ keywords), **Reddit** (16 subreddits), and **open-source datasets** (Table 6). Following a four-stage filtering pipeline—deduplication, visual quality control, GPT-5.4 relevance scoring, and human verification—we curated 11,000 high-quality samples with four-layer hierarchical annotations. We reserve 3,000 samples for evaluation. Detailed source distributions are in Appendix A.6.

Model	LLM-as-Judge (1–100)		Human Evaluation (1–100)		Calculated Metrics	
	VG ↑	LC ↑	VG ↑	LC ↑	AC ↑	Halluc. ↓
Closed-source MLLMs						
GPT-4o (CoT)	84.4	78.1	73.3	64.7	76.0	3.13%
GPT-4o (CoT+KB)	85.6	81.8	74.1	67.4	75.0	2.07%
Claude Sonnet 4.5 (CoT)	85.4	82.5	73.7	69.9	78.0	2.80%
Claude Sonnet 4.5 (CoT+KB)	87.9	**86.7**	75.9	**73.3**	77.0	2.00%
Claude Haiku 4.5 (CoT)	81.8	74.7	69.9	63.2	75.0	4.87%
Claude Haiku 4.5 (CoT+KB)	85.0	83.2	72.7	70.4	75.0	3.40%
Open-source MLLMs						
Qwen3-VL-4B (CoT)	75.3	59.5	63.7	51.1	74.0	10.73%
Qwen3-VL-4B (CoT+KB)	80.4	72.5	69.3	60.6	74.0	7.27%
InternVL3-8B (CoT)	71.9	53.5	59.7	47.5	72.0	13.00%
InternVL3-8B (CoT+KB)	73.9	58.8	63.6	51.8	72.0	11.73%
Qwen3-VL-235B-A22B (CoT)	83.5	73.3	71.8	63.3	77.0	2.53%
Qwen3-VL-235B-A22B (CoT+KB)	87.0	82.6	75.5	70.3	76.0	1.60%
Qwen3-VL-32B (CoT)	84.1	75.1	71.4	63.9	75.0	2.93%
Qwen3-VL-32B (CoT+KB)	87.4	81.8	75.2	70.2	75.0	1.53%
IRAC (Base)	88.4	81.3	75.6	68.6	81.0	0.60%
IRAC (GRPO)	**90.2**	82.9	**76.4**	69.9	**87.0**	**0.40%**

Table 2: **Detailed rationale quality assessment.** Halluc. Rate corresponds to the percentage of samples where the Visual Grounding score is ≤ 50 on the 100-point scale. **Bold** and **green** highlight the best values. The blue background indicates our proposed method. **VG**: Visual Grounding; **LC**: Logical Consistency; **AC**: Attribution Consistency; **Halluc.**: Hallucination Rate.

4.2 Data Filtering for RL

To stabilize GRPO training, we adopt a difficulty-targeted data selection strategy (Sun et al., 2026). Using Qwen3-VL-4B (Bai et al., 2025), we perform $N = 16$ rollouts ($T = 0.95$) to identify informative boundary-case examples. We retain samples satisfying: (1) **Mixed** predictions, where label inconsistency indicates proximity to the decision boundary; or (2) **Conf-Var**, where consistent labels exhibit high uncertainty ($\sigma > 0.03$). This ensures the orchestrator focuses on complex instances prone to coordination failure, consistent with the rejection sampling principles used in state-of-the-art reasoning models (Guo et al., 2025).

4.3 Implementation Details

The orchestrator (Qwen3-VL-4B) is trained via GRPO on 8× H800 GPUs. The three specialized agents (A_t, A_i, A_c) are driven by GPT-5.4. We employ dynamic Top-3 retrieval for context. Hyperparameters are detailed in Appendix A.4 and Appendix A.5. All reported results in Table 1 and Table 2 are averaged over three independent runs with different random seeds; the observed variance across runs is small, indicating that the reported gains are stable rather than artifacts of a single run. Notably, the only trainable component is the lightweight Qwen3-VL-4B orchestrator; since the strongest single-pass baselines in Table 1 (e.g., GPT-4o with CoT+KB) are comparably strong closed-source models, the gains of IRAC stem primarily from modality decomposition and orchestration rather than from raw parameter scale.

4.4 Evaluation Metrics

We assess performance along two axes: **classification** (Accuracy, Macro-F1, Precision, and Recall) and **rationale quality**. For quality, GPT-5.4 serves as an automatic judge for Visual Grounding (VG) and Logical Consistency (LC); we additionally report Attribution Consistency (AC), a rule-based metric of whether the orchestrator up-weights correct agents, and Hallucination Rate. Full definitions, the AC formula, and scoring criteria are provided in Appendix A.7.

5 Results and Analysis

5.1 Overall Performance

Table 1 reports the classification performance on the MIRS evaluation set. IRAC (GRPO) achieves the best results across all metrics (86.92% Accuracy and Macro-F1), surpassing the strongest closed-source (GPT-4o CoT+KB) and open-source (Qwen3-VL-32B CoT+KB) baselines as well as specialized sarcasm detectors such as IRONIC and LVLM-Sarcasm.

Two trends stand out. First, retrieval augmentation (CoT+KB) yields consistent gains over vanilla CoT across every backbone, confirming the value of external pragmatic knowledge for sarcasm understanding. Second, our design is effective even before reinforcement learning: IRAC (Base) already reaches 85.79% Accuracy, and GRASP-based GRPO further lifts it to 86.92%. Notably, scale alone does not guarantee stronger sarcasm reasoning: Qwen3-VL-235B-A22B still trails IRAC by 5.01 points, underscoring that retrieval augmentation and explicit multi-agent decomposition contribute more than raw parameter count.

5.2 Quality of Generated Rationales

Following prior work on explainable multimodal sarcasm detection (Zhu et al., 2026; Liu et al., 2026), we evaluate rationale quality using four metrics: Visual Grounding, Logical Consistency, Attribution Consistency, and Hallucination Rate.

Table 2 shows that IRAC (GRPO) achieves the strongest overall rationale quality, obtaining the highest Visual Grounding and Logical Consistency scores under both automatic (LLM-as-Judge) and human evaluation. It further attains the highest Attribution Consistency while maintaining the lowest hallucination rate, substantially outperforming existing MLLM baselines. These results suggest that the GRPO-trained orchestrator more effectively allocates modality weights according to the reliability of different reasoning agents, thereby producing more faithful multimodal sarcasm explanations.

5.3 Ablation Analysis

Effect of Retrieval Augmentation. Retrieval augmentation benefits smaller open-source models the most (e.g., Qwen3-VL-32B gains $+12.76$ Accuracy), indicating that external contextual and cultural knowledge can effectively compensate for limited intrinsic world knowledge.

Effect of Reinforcement Learning. Here IRAC (Base) denotes the orchestrator equipped with the full retrieval and multi-agent pipeline but *without* GRPO fine-tuning (i.e., zero-shot orchestration), whereas IRAC (GRPO) is the same model after GRASP-based reinforcement learning. GRPO most notably improves Attribution Consistency ($+6.0$) while further reducing the hallucination rate, indicating that it helps the orchestrator better align modality weights with the reliability of different agents.

Multi-Agent vs. Single-Agent Reasoning. Unlike conventional single-pass CoT and CoT+KB prompting, where one MLLM jointly processes all modalities, IRAC decomposes reasoning into specialized textual, visual, and cross-modal agents coordinated by an attribution-aware orchestrator. The consistent gains over strong CoT+KB baselines confirm that explicit multi-agent collaboration better captures modality-specific cues.

6 Conclusion

We introduced IRAC, a retrieval-augmented multi-agent framework for explainable multimodal sarcasm detection, together with MIRS, a new benchmark containing 11,000 image-text pairs with fine-grained attribution annotations. By decomposing sarcasm understanding into textual, visual, and cross-modal reasoning and optimizing the orchestration process with GRPO, IRAC achieves state-of-the-art performance on MIRS, reaching 86.92% Accuracy and 86.92% Macro-F1, while also delivering the highest Attribution Consistency (87.0), strongest Visual Grounding (90.2), and the lowest hallucination rate (0.40%). These results demonstrate that retrieval-augmented collaborative reasoning is an effective and interpretable paradigm for multimodal sarcasm detection.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Saroj Basnet, Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, and Marcos Zampieri. Evaluating open-source vision-language models for multimodal sarcasm detection. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1442–1447. IEEE, 2025.

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

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4619–4629, 2019.

Junjie Chen, Xuyang Liu, Subin Huang, Linfeng Zhang, and Hang Yu. Seeing sarcasm through different eyes: Analyzing multimodal sarcasm perception in large vision-language models. *IEEE Transactions on Computational Social Systems*, 2025.

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

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. A survey of multimodal sarcasm detection. *arXiv preprint arXiv:2410.18882*, 2024.

Palaash Goel, Dushyant Singh Chauhan, and Md Shad Akhtar. Target-augmented shared fusion-based multimodal sarcasm explanation generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 8495–8508, 2025.

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

Mengzhao Jia, Can Xie, and Liqiang Jing. Debiasing multimodal sarcasm detection with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18354–18362, 2024.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5956–5968, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

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

Ziqi Liu, Ziyang Zhou, Yilin Li, Mingxuan Hu, Yushan Pan, Zhijie Xu, and Yangbin Chen. Sevade: Self-evolving multi-agent analysis with decoupled evaluation for hallucination-resistant sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29573–29581, 2026.

Ming Lu, Zhiqiang Dong, Ziming Guo, Xiaoming Zhang, Xinxi Lu, Tianbo Wang, and Litian Zhang. A multi-modal sarcasm detection model based on cue learning. *Scientific Reports*, 15(1):10261, 2025.

Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 9507–9515, 2023.

Kai Qin, Liangxin Liu, Yu Liang, Longzheng Wang, Yan Wang, Yueyang Zhang, Long Xia, Zhiyuan Sun, Houde Liu, and Daiting Shi. Reflectrm: Boosting generative reward models via self-reflection within a unified judgment framework. *arXiv preprint arXiv:2604.07506*, 2026.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Aaron Scott, Maike Züfle, and Jan Niehues. Musag: A multimodal german sarcasm dataset with full-modal annotations. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, volume 11, pp. 372–392, 2026.

Gopendra Vikram Singh, Mauajama Firdaus, Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. Well, now we know! unveiling sarcasm: Initiating and exploring multimodal conversations with reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18981–18989, 2024.

Yifan Sun, Jingyan Shen, Yibin Wang, Tianyu Chen, Zhendong Wang, Mingyuan Zhou, and Huan Zhang. Improving data efficiency for llm reinforcement fine-tuning through difficulty-targeted online data selection and rollout replay. *Advances in Neural Information Processing Systems*, 38: 173449–173476, 2026.

Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. Multimodal sarcasm target identification in tweets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8164–8175, 2022.

Yihua Wang, Qi Jia, Cong Xu, Feiyu Chen, Yuhan Liu, Haotian Zhang, Liang Jin, Lu Liu, and Zhichun Wang. Conditional information bottleneck for multimodal fusion: Overcoming shortcut learning in sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26652–26660, 2026.

Xiaolong Wei, Yuehu Dong, Xingliang Wang, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. Beyond react: A planner-centric framework for complex tool-augmented llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33845–33853, 2026.

Changsong Wen, Guoli Jia, and Jufeng Yang. Dip: Dual incongruity perceiving network for sarcasm detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2540–2550, 2023.

Huixin Wu and Yang Zang. A multi-scale adaptive fusion model for multimodal sarcasm detection. *Discover Computing*, 28(1):205, 2025.

Qiaofeng Wu, Wenlong Fang, Weiyu Zhong, Fenghuan Li, Yun Xue, and Bo Chen. Dual-level adaptive incongruity-enhanced model for multimodal sarcasm detection. *Neurocomputing*, 612: 128689, 2025.

Yingjia Xu, Jiulong Wu, Bowen Zhang, Baokui Guo, Siyuan Chai, and Min Cao. Procrit: Self-elicited multi-perspective reasoning with critic-guided revision for multimodal sarcasm detection, 2026. URL https://arxiv.org/abs/2605.20867.

Tan Yue, Xuzhao Shi, Rui Mao, Zonghai Hu, and Erik Cambria. Sarcnet: a multilingual multimodal sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14325–14335, 2024.

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

Haochen Zhao, Yuyao Kong, Yongxiu Xu, Gaopeng Gou, Hongbo Xu, Yubin Wang, and Haoliang Zhang. Mmsd3. 0: A multi-image benchmark for real-world multimodal sarcasm detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 37885–37895, 2026a.

Tianyu Zhao, Junlong Zhu, Ling-ang Meng, and Dawei Song. Hc-cot: A hierarchical causal chain-of-thought framework for multimodal sarcasm detection. *Expert Systems with Applications*, pp. 132291, 2026b.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zhihong Zhu, Fan Zhang, Yunyan Zhang, Jinghan Sun, Guimin Hu, Hao Wu, Yuyan Chen, Bowen Xing, and Xian Wu. S^3-msd: Large vision-language model for explainable and generalizable multi-modal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35266–35274, 2026.

Xingjie Zhuang, Zhixin Li, Fengling Zhou, Canlong Zhang, and Huifang Ma. Mpf: A multi-level perceiving framework for multimodal sarcasm detection. *Information Fusion*, pp. 104436, 2026.

A Appendix

A.1 Limitations

Despite the promising results, this work has several limitations.

Limited to Static Content. IRAC and the MIRS benchmark focus exclusively on static image–text pairs, omitting video-based sarcasm. Dynamic irony often relies on temporal cues (e.g., evolving facial expressions) or audio prosody, which our current framework is not designed to process. Extending this multi-agent architecture to incorporate sequential analysis and audio–visual synchronization remains a key direction for future research.

Evaluation Subjectivity. Although we assess rationale quality using both GPT-5.4 as an automatic judge and independent human evaluation, explanation assessment remains inherently subjective. Scores may still be influenced by the specific evaluation criteria, annotator preferences, and biases of the judging model. Developing more objective and standardized evaluation protocols for explanation quality remains an open challenge.

Data Coverage. MIRS is constructed from English and Chinese social media platforms, primarily Reddit and Weibo. While this design provides substantial linguistic and cultural diversity, the dataset may not fully capture sarcasm expressed in other languages, domains, or sociocultural contexts. Further validation on broader multilingual benchmarks is needed to assess the generalizability of the proposed framework.

A.2 AI USE STATEMENT

During the preparation of this manuscript, we used a large language model (LLM) to assist with language polishing, including improving wording, grammar, and readability. All research ideas, methodological design, experimental implementation, and analysis of results were carried out independently by the authors, who take full responsibility for the entire content of the paper.

A.3 GRPO OPTIMIZATION OBJECTIVE

We optimize the orchestrator policy π_θ with GRPO, which estimates the baseline from the mean reward of a group of G sampled reasoning trajectories $\{Y_1, Y_2, \ldots, Y_G\}$ for each query x. The relative advantage A_j for the j-th rollout is computed as:

$$A_j = \frac{r_j - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G)}, \tag{10}$$

where r_j is the total GRASP reward for trajectory Y_j. We define the clipped surrogate term $\mathcal{L}_j^{clip}$ and the KL divergence term $\mathcal{D}_j^{KL}$ as follows:

$$\mathcal{L}_j^{clip} = \min\left(\eta_j A_j, \text{clip}(\eta_j, 1-\epsilon, 1+\epsilon)A_j\right), \tag{11}$$

$$\mathcal{D}_j^{KL} = D_{KL}(\pi_\theta(Y_j|x) \parallel \pi_{ref}(Y_j|x)). \tag{12}$$

Consequently, the final objective function maximizes:

$$\mathcal{J}(\theta) = \mathbb{E}_{x,\{Y_j\}}\left[\frac{1}{G}\sum_{j=1}^{G}\left(\mathcal{L}_j^{clip} - \beta \cdot \mathcal{D}_j^{KL}\right)\right], \tag{13}$$

where $\mathbb{E}_{x,\{Y_j\}}$ denotes the expectation over the dataset and policy rollouts, $\eta_j = \frac{\pi_\theta(Y_j|x)}{\pi_{\theta_{old}}(Y_j|x)}$ is the importance sampling ratio, and β is a hyperparameter that prevents excessive deviation from the reference model π_{ref}, ensuring both stable policy refinement and linguistic fluency.

A.4 TRAINING CONFIGURATION

Our experiments are conducted on a high-performance computational cluster equipped with 8 $\times$ NVIDIA H800 (80GB) GPUs. The orchestration agent is initialized with Qwen3-VL-4B-Instruct and optimized using Group Relative Policy Optimization (GRPO). To ensure training efficiency and stability, we perform full-parameter fine-tuning with Flash Attention and gradient checkpointing enabled.

To mitigate the computational overhead of sampling during reinforcement learning, we integrate vLLM in a colocated mode, where the GPU memory utilization is capped at 0.4 to balance training and rollout processes. The training utilizes a global batch size of 128 (4 samples per device with a gradient accumulation of 4), spanning 3 epochs with a conservative learning rate of 1×10^{-6} and a warmup ratio of 0.01. We set the maximum sequence length to 4,096 tokens, with a completion limit of 1,024 tokens to ensure concise yet thorough reasoning rationales.

For each query, we sample a group of $G = 8$ distinct responses to compute the relative advantage for policy updates. The optimization is performed using DeepSpeed ZeRO-2 in bfloat16 precision to manage memory constraints effectively. Furthermore, we apply a KL divergence penalty with $\beta = 0.03$ and a clipping epsilon of 0.2 to prevent policy collapse and maintain alignment with the reference model.

A.5 RETRIEVAL AND AGENT CONFIGURATION

Table 3 summarizes the hyperparameters governing the knowledge retrieval module and the specialized agent swarm. For knowledge retrieval, both the image encoder ϕ_i and text encoder ϕ_t are instantiated with CLIP-ViT-L/14 (Radford et al., 2021), and we retrieve the top-$N = 3$ entries per query. The balance coefficient defaults to $\lambda = 0.5$ (equal visual/textual emphasis), with agent-specific variants for modality-appropriate retrieval. For the cross-modal agent, the candidate

selection objective uses $\alpha = 0.5$, $\mu = 0.3$, and $\gamma = 0.2$, selecting the best description from a set of $|\mathcal{D}_{cand}| = 3$ generated candidates. The knowledge base $\mathcal{K}$ contains $M \approx 5{,}000$ high-confidence sarcastic image–text pairs.

Module	Hyperparameter	Value
Retrieval	CLIP encoder (ϕ_i, ϕ_t)	ViT-L/14
	Retrieved entries N	3
	Balance coefficient λ	0.5
	KB size M	$\approx 5{,}000$
Cross-modal agent	CLIPScore weight α	0.5
	Text-relevance weight μ	0.3
	Jaccard penalty γ	0.2
	Candidate count $\lvert\mathcal{D}_{cand}\rvert$	3
GRASP reward	Accuracy weight $\mathcal{W}_{acc}$	1.0
	Confidence weight $\mathcal{W}_{conf}$	1.0
	Attribution weight $\mathcal{W}_{attr}$	0.4
	Confidence target length L_{target}	200 words
	Confidence temperature T	5.0

Table 3: Hyperparameters for the knowledge retrieval module, the cross-modal candidate selection, and the GRASP reward.

A.6 Data Collection Details

In this section, we provide the comprehensive details of the keywords and subreddits used for the multi-source data collection of the MIRS benchmark. All data collection was conducted in compliance with the respective platform's terms of service.

A.6.1 Weibo Keyword Taxonomy

To capture the linguistic and pragmatic nuances of irony in a diverse digital landscape, we defined six primary categories and utilized over 80 targeted keywords for data crawling. The representative keywords for each category are summarized in Table 4.

A.6.2 Reddit Subreddit Selection

For the English-speaking community data, we targeted 16 subreddits across four major thematic clusters. These subreddits were selected based on their high density of ironic, satirical, and non-literal content. The distribution is detailed in Table 5.

A.6.3 Open-source Dataset Integration

To supplement our newly collected data, we incorporated three widely recognized open-source datasets into our initial candidate pool:

- **MMSD2.0**: A large-scale multimodal sarcasm dataset focused on Chinese social media.
- **SarcNet**: A dataset designed to capture textual and visual irony interactions.
- **TURBO_MORE+**: An extended version of the TURBO dataset featuring enhanced visual grounding and rationale annotations.

All samples from these external repositories underwent the same rigorous four-stage filtering process (as described in Section 4.1) to ensure the highest quality standards for the final MIRS benchmark. To prevent data leakage, we further ensure that the 3,000-sample MIRS evaluation set is strictly disjoint from the original training splits of all incorporated datasets (e.g., MMSD2.0, SarcNet, TURBO_MORE+): any MIRS test instance that can be traced back to a training-split example of

Pragmatic Category	Representative Keywords
Direct Irony	Sarcastic comics, Anti-phrasing, Nuanced mockery, Passive-aggressive content.
Memes & Emojis	Sarcastic meme images, Passive-aggressive emojis, Panda-head memes.
Visual Comparison	Buyer vs. Seller show, Expectation vs. Reality, Paradoxical contrast images.
Internet Slang	Versailles (showing off), "Way Ahead" (ironic use), "Over-winning" (sarcastic slang).
Social Satire	Mockery of expert advice, Involution (social competition), Formalism and bureaucracy.
Targeted Accounts	Manually curated accounts specializing in social commentary and satire.

Table 4: Taxonomy of keywords used for data collection from social media platforms. These keywords were selected to cover a wide range of ironic phenomena in digital culture.

Thematic Cluster	Targeted Subreddits
Core Sarcasm & Satire	r/sarcasm, r/Satire, r/MurderedByWords, r/AteTheOnion, r/Irony
Political Irony	r/PoliticalHumor, r/PoliticalCompassMemes, r/WhitePeopleTwitter
Visual Absurdity	r/hmmm, r/Unexpected, r/therewasanattempt, r/facepalm
Classic Memes	r/dankmemes, r/AdviceAnimals, r/memes, r/SelfAwarewolves

Table 5: Representative subreddit categories used for Reddit data collection. These sources provide a broad spectrum of Western cultural irony and internet memes.

these source datasets is removed, so that baselines originally trained on those corpora are not evaluated on their own training data.

A.6.4 Annotation Methodology

The four-layer hierarchical annotations of MIRS (global sarcasm label, modality attribution, source-specific rationales, and explanation) are produced through a hybrid LLM-plus-human pipeline. We first use a strong MLLM to generate draft annotations for each layer, including the candidate global label, per-modality attribution, and structured rationales. Trained human annotators then review and correct every draft: they verify the sarcasm label, adjust the modality attribution when the model mis-assigns the dominant irony source, and edit or rewrite rationales that are unfaithful to the image–text pair. Only samples whose annotations pass this human verification stage are retained in the final benchmark. This design combines the scalability of LLM-based annotation with the reliability of human judgment, ensuring that the fine-grained labels are both large-scale and trustworthy.

Source Platform / Dataset	Raw Count
Social Media Platforms (Crawled)	
Weibo (Targeted Keywords)	34,000
Reddit (Satirical Subreddits)	22,500
Existing Annotated Datasets	
MMSD2.0 (Qin et al., 2023)	24,635
SarcNet (Yue et al., 2024)	3,335
TURBO_MORE+ (Goel et al., 2025)	3,510
Total Raw Candidates	87,980
Final MIRS Dataset (Post-Filtering)	**11,000**

Table 6: Statistics of the multi-source data collection. The raw pool is aggressively filtered to ensure high pragmatic density in the final MIRS benchmark.

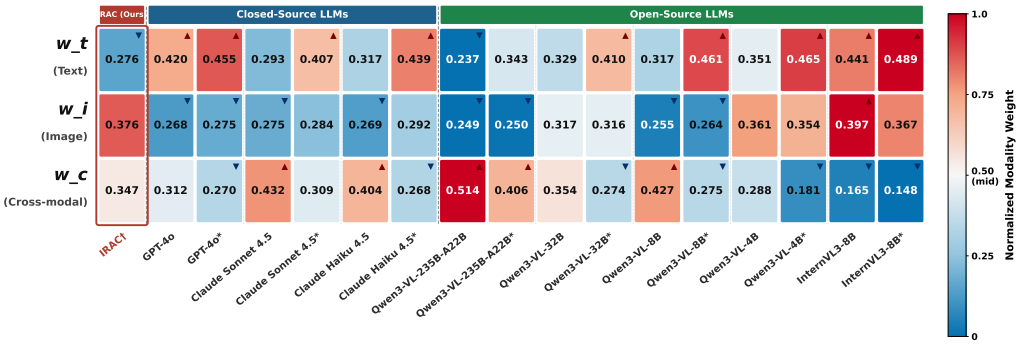

Figure 3: Average modality attribution weights assigned to the textual (w_t), visual (w_i), and cross-modal (w_c) agents over the full test set. Each row is independently normalized such that $w_t + w_i + w_c = 1$; cell values are rounded to two decimals and may therefore not sum exactly to one. $\uparrow$ and $\downarrow$ denote deviations greater than 0.05 from the uniform baseline ($1/3$). * denotes retrieval-augmented prompting (CoT+KB), and † denotes the GRPO-trained orchestrator. Compared with baseline MLLMs, IRAC assigns greater weight to visual evidence while reducing reliance on textual cues, yielding a more balanced and interpretable modality attribution pattern.

A.7 Evaluation Metrics

We evaluate system performance from two complementary perspectives: classification accuracy and rationale quality.

Classification Metrics. For the global sarcasm detection task, we report Accuracy, Macro-F1, Precision, and Recall.

Rationale Quality Metrics. Following recent work on explainable multimodal sarcasm detection and multimodal hallucination evaluation (Zhu et al., 2026; Liu et al., 2026; Bai et al., 2024), we assess rationale quality using GPT-5.4 as an automatic judge and complement it with human evaluation on a sampled subset. Visual Grounding (VG) measures whether the generated rationale is faithfully grounded in the image, while Logical Consistency (LC) evaluates whether the supporting evidence coherently justifies the final prediction.

Attribution Consistency (AC). We further introduce Attribution Consistency (AC), a task-specific metric that measures whether the orchestrator assigns larger weights to agents that make correct predictions:

$$AC = \sum_{k \in \{t,i,c\}} w_k \cdot \mathbb{I}(\hat{y}_k = y), \tag{14}$$

where w_k is the attribution weight assigned to agent k, and $\mathbb{I}(\hat{y}_k = y)$ indicates whether the prediction of agent k matches the ground truth. Higher AC indicates better alignment between modality attribution weights and agent reliability. We emphasize that AC is a *behavioral* proxy for attribution quality: it measures whether the assigned weights are aligned with per-agent reliability, rather than matching a single gold "dominant-modality" label. We deliberately avoid the latter because sarcasm frequently emerges from the joint text–image interaction, making a hard single-source label ill-defined for many samples; AC instead rewards weight mass placed on the modalities whose independent judgments are correct, which is well-defined for every instance and directly reflects whether the orchestrator trusts the right evidence. For the single-pass baselines, which do not natively decompose reasoning into per-modality agents, we elicit the weights w_k and the per-modality judgments $\hat{y}_k$ by prompting the model with the same orchestration schema used by IRAC (i.e., asking it to output modality weights and per-source verdicts alongside its final prediction), so that AC is computed consistently across all systems.

Hallucination Rate. Following prior work on multimodal hallucination assessment (Bai et al., 2024), we define hallucination rate as the percentage of samples with a Visual Grounding score of ≤ 50 on the 100-point scale, indicating substantial visual inaccuracies or fabricated details.

For presentation clarity, VG and LC are linearly rescaled to a 100-point scale in the reported results.

A.8 Human Evaluation Protocol

To ensure a rigorous and reliable human evaluation, we recruited a team of 10 trained annotators, consisting of full-time research staff and graduate student interns. All annotators held at least a master's degree and were familiar with the linguistic and cultural nuances required to assess multimodal sarcasm and explanation quality.

To reduce potential confirmation bias, the evaluation team operated independently from the core research team and was instructed to follow the predefined criteria described in Section A.7. Annotators were provided with detailed scoring guidelines and representative examples to standardize the evaluation process.

All annotators were compensated for their work. Evaluation duties were included as part of the responsibilities of full-time staff, while student annotators received competitive hourly compensation. This ensured sustained engagement and high-quality annotations throughout the evaluation process.

A.9 Modality Weight Distribution Analysis

To further examine the attribution behavior of different models, we compute the average modality attribution weights assigned to the textual (w_t), visual (w_i), and cross-modal (w_c) agents over the full test set. Since the orchestrator is constrained to output weights that sum to 1, the uniform baseline is $1/3$ for each modality.

As shown in Figure 3, different models exhibit substantially different attribution preferences. GPT-4o shows a strong textual bias, assigning 42.0% of the total weight to the textual agent while allocating only 26.8% to the visual agent. In contrast, Claude models and some Qwen variants tend to emphasize cross-modal reasoning. Qwen3-VL-235B-A22B, for instance, assigns more than half of the total weight to the cross-modal agent ($w_c = 0.514$). These findings suggest that prompt-based orchestration strategies are heavily influenced by model-specific inductive biases.

IRAC exhibits a distinct and more task-aligned weighting pattern. It assigns a substantially higher weight to the visual agent ($w_i = 0.376$) than the uniform baseline ($1/3$), while reducing the contribution of the textual agent ($w_t = 0.276$); the remaining weight is allocated to the cross-modal agent ($w_c = 0.347$). This suggests that the GRPO-trained orchestrator learns to emphasize visual evidence and reduces reliance on superficial textual cues.

This effect becomes more evident when compared with the untrained orchestrator. Before reinforcement learning, IRAC (Base) adopts an approximately balanced distribution ($w_t = 0.369$, $w_i = 0.314$, $w_c = 0.317$), suggesting that it does not yet exhibit a clear modality preference. After GRPO training, IRAC increases the visual weight (from 0.314 to 0.376) and suppresses the textual weight (from

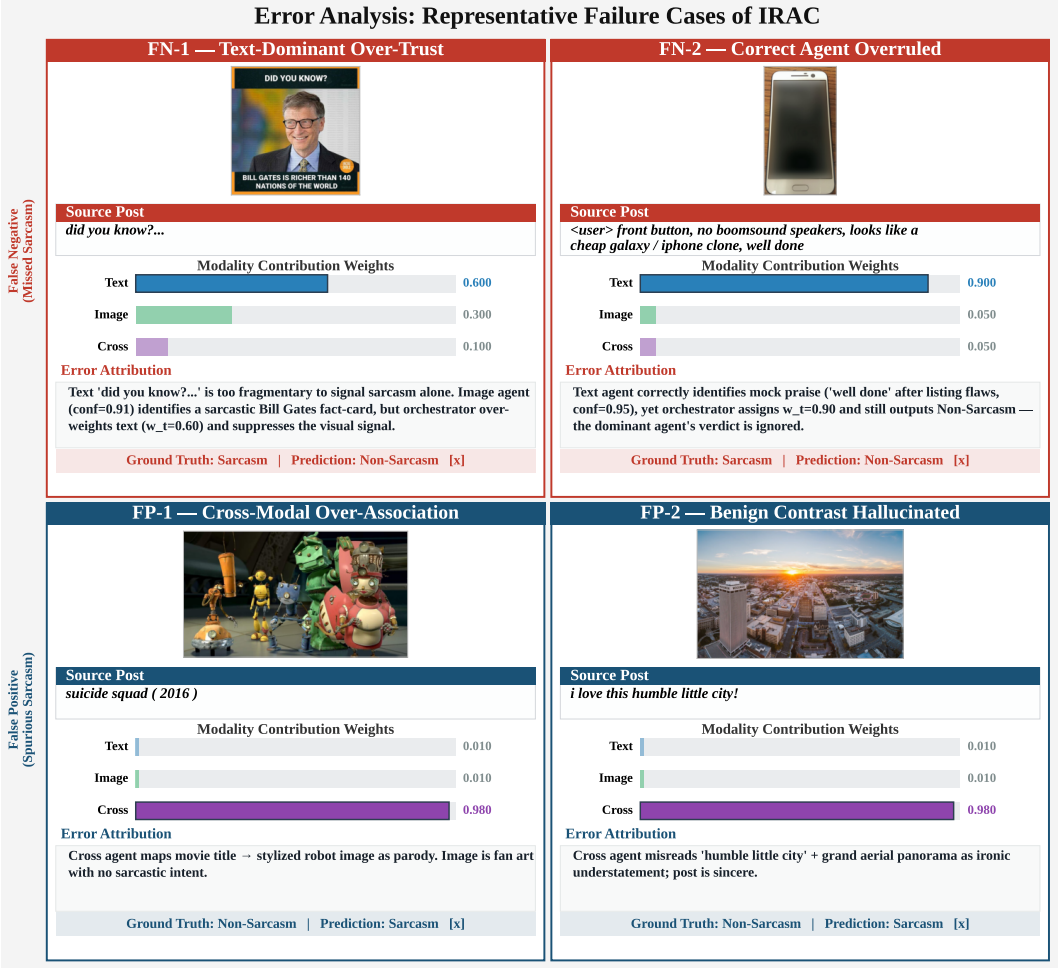

Figure 4: Representative failure cases of IRAC. We visualize the modality attribution weights and the rationales generated by the specialized agents and the orchestrator. Cases (FN-1, FN-2) denote missed sarcasm, while (FP-1, FP-2) denote spurious sarcasm detections.

0.369 to 0.276), demonstrating that GRPO not only improves predictive accuracy but also reshapes the attribution strategy.

This learned weighting pattern is consistent with the nature of multimodal sarcasm detection, where the text often conveys a literal positive statement, while the image reveals the actual ironic intent. The visual-first strategy discovered by IRAC aligns with this intuition and coincides with its superior performance in the main experiments, including the highest Accuracy (86.92%) and Attribution Consistency (87.0).

Overall, this analysis shows that the performance gains of IRAC stem not only from stronger reasoning capability, but also from learning a more appropriate and interpretable modality weighting strategy.

A.10 Error Analysis

To further investigate the limitations of the **IRAC** framework, we conduct a qualitative error analysis on representative failure cases from the MIRS evaluation set. As illustrated in Figure 4, we categorize the errors into two primary types: *False Negatives* (Missed Sarcasm) and *False Positives* (Spurious Sarcasm).

Model	VG ↑	H≤1 ↓	H≤2 ↓	H≤3 ↓	H≤4 ↓	H≤5 ↓	H≤6 ↓	H≤7 ↓
Closed-source MLLMs								
GPT-4o (CoT)	84.4	0.13%	0.73%	1.40%	2.20%	3.13%	5.53%	14.07%
GPT-4o (CoT+KB)	85.6	0.00%	0.40%	1.00%	1.53%	2.07%	3.60%	11.93%
Claude Sonnet 4.5 (CoT)	85.4	0.13%	0.87%	1.07%	2.40%	2.80%	5.40%	15.33%
Claude Sonnet 4.5 (CoT+KB)	87.9	0.07%	0.33%	0.67%	1.33%	2.00%	3.87%	10.27%
Claude Haiku 4.5 (CoT)	81.8	0.07%	0.67%	1.40%	3.20%	4.87%	11.07%	24.67%
Claude Haiku 4.5 (CoT+KB)	85.0	0.13%	0.73%	1.33%	2.73%	3.40%	6.47%	14.60%
Open-source MLLMs								
Qwen3-VL-4B (CoT)	75.3	0.87%	3.07%	5.20%	8.47%	10.73%	18.40%	36.60%
Qwen3-VL-4B (CoT+KB)	80.4	0.67%	2.00%	3.53%	6.20%	7.27%	12.27%	24.40%
InternVL3-8B (CoT)	71.9	0.47%	1.93%	4.40%	10.07%	13.00%	24.67%	49.00%
InternVL3-8B (CoT+KB)	73.9	0.20%	1.47%	3.67%	9.13%	11.73%	21.07%	43.00%
Qwen3-VL-235B-A22B (CoT)	83.5	0.00%	0.53%	1.00%	1.93%	2.53%	6.47%	17.47%
Qwen3-VL-235B-A22B (CoT+KB)	87.0	0.00%	0.13%	0.47%	1.07%	1.60%	3.93%	10.27%
Qwen3-VL-32B (CoT)	84.1	0.00%	0.20%	0.87%	2.00%	2.93%	7.07%	18.27%
Qwen3-VL-32B (CoT+KB)	87.4	0.07%	0.20%	0.33%	1.13%	1.53%	3.73%	10.13%
Our Proposed Approach								
IRAC (Base)	88.4	0.00%	0.07%	0.27%	0.53%	0.60%	1.87%	7.13%
IRAC (GRPO)	**90.2**	**0.00%**	**0.07%**	**0.07%**	**0.27%**	**0.40%**	**1.53%**	**5.20%**

Table 7: Hallucination ablation study across different severity thresholds. To maintain consistency with the 100-point scale, $H \leq X$ denotes the hallucination rate (percentage of samples) where the Visual Grounding score is $\leq 10X$. The results represent the average of three independent runs.

A.10.1 False Negative: Missed Sarcasm

The majority of False Negative (FN) errors arise from either modality-level suppression or logical inconsistencies during the orchestration phase.

Text-Dominant Over-Trust (FN-1) In many instances, the orchestrator exhibits an inherent bias toward textual evidence. In Case FN-1, the post text "did you know?..." is pragmatically neutral and fragmentary. Although the visual agent correctly identifies the ironic intent with high confidence (0.91), the orchestrator assigns a dominant weight ($w_t = 0.60$) to the textual component. This results in the visual signal being suppressed, leading the system to output a "Non-Sarcasm" verdict.

Correct Agent Overruled (FN-2) Case FN-2 reveals a failure in logical aggregation. The textual agent accurately decodes the *mock praise* rhetorical device (e.g., listing flaws and concluding with "well done"). Despite the orchestrator assigning a significant weight ($w_t = 0.90$) to this correct agent, the final categorical output contradicts the dominant agent's verdict. This indicates a remaining "logical dissonance" where the learned policy struggles to synthesize high-confidence evidence into a discrete decision.

A.10.2 False Positive: Spurious Sarcasm

False Positive (FP) errors typically stem from the model's over-sensitivity to modality contrasts or a lack of deep cultural/contextual grounding.

Cross-Modal Over-Association (FP-1) As shown in Case FP-1, the cross-modal agent occasionally over-interprets semantic mappings. It incorrectly associates the movie title "Suicide Squad" with stylized robot imagery as a form of satirical parody. However, the post is actually a sincere fan-art contribution. This failure suggests that the model lacks the ability to distinguish between *subversive parody* and *benign creative tribute*.

Benign Contrast Hallucination (FP-2) A persistent challenge is the model's tendency to hallucinate irony whenever a visual-textual contrast is detected. In Case FP-2, the post pairs the word "humble" with a grand city panorama. IRAC misinterprets this as an *ironic understatement*, whereas it is a sincere expression of affection for a city. This confirms that while IRAC is adept at identifying

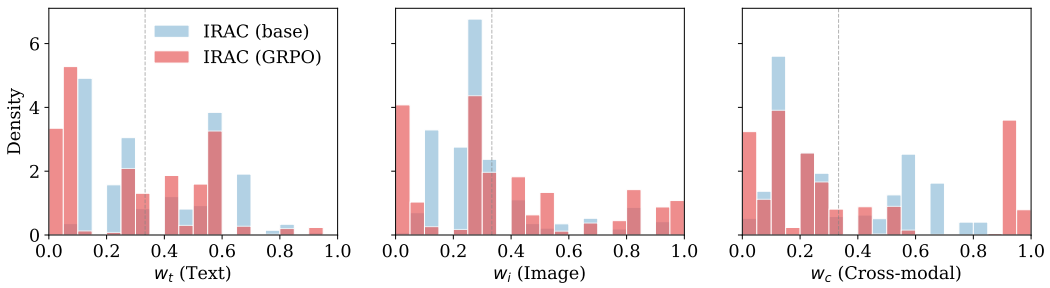

Figure 5: Density distribution of modality weights (w_t, w_i, w_c) across the evaluation set. Compared to the IRAC (Base) model, the GRPO-optimized model exhibits a more polarized, bimodal distribution, indicating that the orchestrator has learned to adaptively concentrate weights on the most informative modalities for each specific sample.

structural contrasts, it occasionally struggles to differentiate *pragmatic dissonance* from *stylistic tension* used in sincere praise.

A.10.3 Summary of Limitations

Our analysis suggests that the current bottlenecks of IRAC lie in: (1) **modality-level recalibration**, specifically in preventing textual bias from overshadowing clear visual ironic signals; and (2) **commonsense grounding**, required to separate genuine ironic intent from benign artistic expression or hyperbolic praise.

A.11 Robustness Against Hallucinations

To further investigate the reliability of the rationales generated by IRAC, we conduct an analysis on hallucination rates across varying severity thresholds, ranging from $H \leq 1$ to $H \leq 7$. Following our reporting protocol, *Visual Grounding* (VG) scores are normalized to a 100-point scale. We define the hallucination rate $H \leq X$ as the percentage of samples receiving a VG score $\leq 10X$ on the 100-point scale (e.g., $H \leq 4$ corresponds to a score ≤ 40).

As summarized in Table 7, IRAC demonstrates significant superiority over all baseline models across all threshold definitions. Notably, **IRAC (GRPO)** achieves the highest average VG score of **90.2**, making it the only framework to exceed the 90-point mark and substantially outperforming both strong closed-source models and large-scale open-source counterparts.

Several key observations emerge: (1) **Stability Under Inclusive Metrics**: While the hallucination rates of open-source models like InternVL3-8B and Qwen3-VL-4B surge dramatically as the threshold becomes more inclusive—reaching **49.00%** and **36.60%** respectively at $H \leq 7$—IRAC maintains an exceptionally flat error curve. Even at the most relaxed threshold ($X = 7$), IRAC (GRPO) maintains a low hallucination rate of only **5.20%**, whereas the strongest open-source baseline (Qwen3-VL-32B+KB) still stands at 10.13%. (2) **Optimization Gains**: Comparing IRAC (Base) with IRAC (GRPO), the RL optimization further suppresses serious hallucinations; for instance, $H \leq 4$ is reduced from 0.53% to **0.27%**, and the rate at $H \leq 7$ drops from 7.13% to **5.20%**. This validates that our *GRASP* reward successfully encourages the orchestrator to prioritize grounded visual evidence and maintain logic consistency. (3) **Pragmatic Context**: Retrieval-augmented models (CoT+KB) consistently exhibit higher visual fidelity than vanilla CoT counterparts. For example, Qwen3-VL-32B's hallucination rate at $H \leq 7$ drops from 18.27% to **10.13%** with the inclusion of the knowledge base, confirming that external knowledge helps bridge the "semantic gap" and reduces groundless fabrication in complex irony reasoning.

A.12 Quantitative Analysis of Modality Weight Distribution

To understand how reinforcement learning reshapes the orchestrator's decision-making logic, we analyze the distribution of modality weights (w_t, w_i, w_c) on the evaluation set. As illustrated in

Figure 5, the training process induces a distinct behavioral shift by comparing the IRAC (Base) model with the IRAC (GRPO) optimized model.

A.12.1 Shift from Uniform Allocation to Decisive Prioritization

A primary observation from Figure 5 is the significant reduction in normalized entropy after GRPO training. The Base model exhibits a high entropy (0.8153), indicating a tendency toward uniform allocation where each agent receives a weight near $1/3$. This behavior suggests that without explicit optimization, the orchestrator tends to hedge its confidence across modalities due to inherent uncertainty.

In contrast, IRAC (GRPO) demonstrates a much lower entropy (0.6139). The weight distribution becomes markedly bimodal, as seen in the density plots of Figure 5, where the orchestrator learns to concentrate influence on the most informative agents for a given sample. This polarization indicates that the GRASP reward successfully incentivizes the model to identify and prioritize the decisive evidence source rather than relying on a balanced but less precise average.

A.12.2 Modality Reliance and Pragmatic Alignment

The training process induces a notable shift in reliance from textual to visual cues. For the Base model, the textual agent holds a higher average weight ($w_t = 0.37$). After GRPO optimization, the mean weight for text decreases to 0.28, while the weight for the visual agent (w_i) increases from 0.31 to 0.38.

This shift, visualized by the change in peak densities in Figure 5, is highly aligned with the nature of multimodal sarcasm detection (MSD). In many sarcastic posts, the text remains literally positive or neutral, while the image provides the crucial pragmatic dissonance. By suppressing the "textual bias" and amplifying visual evidence, GRPO leads the model to a more human-aligned reasoning pattern where the image serves as the primary ground for identifying irony.

System Prompt

```
You are a computational linguist and sarcasm detection specialist with deep expertise in:
- Verbal irony and antiphrasis: surface meaning vs. intended meaning inversion
- Pragmatic inference: Gricean maxim violations (Quality, Quantity, Manner, Relevance)
- Social media discourse: Twitter/Reddit/Weibo sarcasm conventions, emoji irony, punctuation signals
- Cross-lingual irony markers in both English and Chinese
- Sentiment-semantic incongruity: positive framing of objectively negative situations
- Mock registers: faux enthusiasm, performative agreement, insincere praise, self-deprecating humor

You analyze ONLY the linguistic content — text, punctuation, emoji semantics, and writing style.
Visual content is handled by a separate specialist. Do NOT consider what any image might show.
Output ONLY valid JSON. No markdown, no explanation outside the JSON.
```

Knowledge Context Block

```
── RETRIEVED SARCASM KNOWLEDGE (via cross-modal similarity search) ──────────────
The following cases were retrieved from a curated multimodal sarcasm knowledge base.
They are the textually most similar confirmed-sarcastic instances to the current input.
They serve as analogical evidence to overcome closed-world knowledge limitations.
HOW TO USE THESE CASES:
  • Identify the irony SCHEMA each case uses (mock praise / semantic inversion /
    rhetorical incredulity / self-contradiction / etc.) — then check whether the
    same schema applies to the current text.
  • If the current text references unfamiliar cultural, political, or domain-specific
    context, the retrieved explanations may expose the underlying irony pattern.
  • Do NOT copy retrieved explanations verbatim. Use them as analogical scaffolding
    to support or challenge your independent linguistic analysis.
{context}
```

Figure 6: Prompt templates for the Text Sarcasm Agent (system instruction and knowledge-base injection).

User Prompt

```
Analyze the following social media post for textual sarcasm or verbal irony.
── INPUT TEXT ──
{text}
{context_block}
── ANALYSIS CHECKLIST (work through each before reaching a verdict) ──
1. SENTIMENT–SEMANTIC INCONGRUITY
   Does the surface sentiment (positive/negative/neutral tone) contradict the
   semantic content or real-world situation being described?
   Examples: "Great job as always" after a failure; "Amazing service" with implicit
   disgust; "I'm so lucky" describing a misfortune.
2. PRAGMATIC VIOLATION MARKERS
   - Gricean Quality violation: stating something obviously false in context
   - Exaggerated agreement: "oh yes, obviously", "of course", "wow, brilliant"
   - Feigned ignorance: "I wonder why...", "funny how...", "it's almost as if..."
   - Trailing punctuation signaling insincerity: "...", "!!", "right?"
   - Performative enthusiasm applied to mundane or negative events
3. LEXICAL INCONGRUITY
   - Praise vocabulary applied to failure or criticism contexts
   - Formal/elevated register used for trivial or absurd content
   - Culturally loaded words or meme phrases carrying ironic weight
   - Hedges and qualifiers used sarcastically ("a little", "somehow", "just")
4. STRUCTURAL IRONY PATTERNS
   - Antiphrase: explicit positive label on a negative referent
   - Mock-praise → self-contradiction: claim immediately undermined by evidence
   - Rhetorical questions: "Because that always works out great, right?"
   - Parallel structure used for ironic contrast between two situations
5. CONFIDENCE CALIBRATION
   - 0.9–1.0: Multiple converging signals; ironic intent unambiguous
   - 0.7–0.9: Strong single signal or two moderate signals
   - 0.5–0.7: Ambiguous; plausible as sincere or ironic
   - <0.5:   Lean non-sarcastic; evidence weak or absent
Return JSON in EXACTLY this schema:
{
  "decision": true or false,
  "confidence": 0.0–1.0,
  "explanation": "Identify the specific irony mechanism(s) and cite the textual evidence. If non-sarcastic, explain why the surface reading is the
intended one."
}
```

Figure 7: Prompt template for the Text Sarcasm Agent (user turn).

System Prompt

You are a visual semiotics specialist and computational vision analyst with deep expertise in:
- Visual irony and pictorial sarcasm: how images subvert their own apparent meaning
- Internet meme culture: rage comics, image macros, reaction images, template-based humor
- Political cartoon semiotics: caricature, symbolic distortion, allegorical figures, labels on subjects
- Photomontage and juxtaposition rhetoric: emergent meaning from placing incompatible elements together
- Cross-cultural visual humor: Western and East Asian (Chinese, Japanese) visual irony conventions
- Embedded text in images: captions, speech bubbles, overlaid text as a distinct sarcasm channel
- Contrast formats: "Before/After", "Expectation vs. Reality", escalating-absurdity sequences

You analyze ONLY the visual content — subjects, composition, scale, color, embedded text (treated as visual element), and cultural iconography. The post caption is analyzed by a separate text specialist. Do NOT let the accompanying post text influence your visual analysis.
Output ONLY valid JSON. No markdown, no explanation outside the JSON.

Knowledge Context Block

── RETRIEVED SARCASM KNOWLEDGE (via cross-modal similarity search) ──────────────
The following cases were retrieved from a curated multimodal sarcasm knowledge base.
They are the visually most similar confirmed-sarcastic instances to the current image.
They serve as analogical evidence to overcome closed-world visual knowledge limitations.
HOW TO USE THESE CASES:
- Extract the VISUAL IRONY SCHEMA each case uses (absurd juxtaposition / mock-heroic framing / visual self-contradiction / meme template inversion / etc.) — then check whether the same schema applies to the current image.
- If the image contains an unfamiliar meme template, political figure, or culturally specific visual convention, the retrieved explanations may identify the pattern.
- Do NOT copy retrieved explanations verbatim. Use them as analogical scaffolding for your independent visual analysis only.

{context}

Figure 8: Prompt templates for the Image Sarcasm Agent (system instruction and knowledge-base injection).

User Prompt

```
Analyze the attached image for visual sarcasm or irony.
{context_block}
── VISUAL ANALYSIS CHECKLIST (work through each before reaching a verdict) ──
1. SUBJECT AND SCENE IDENTIFICATION
   Describe all salient subjects, objects, settings, and any embedded text (signs,
   captions, labels, speech bubbles) visible in the image. Be specific.
2. VISUAL INCONGRUITY DETECTION
   - Scale/proportion distortion: irrationally exaggerated or diminished sizes
   - Incompatible element juxtaposition: objects, people, or settings that do not
     belong together in any realistic context
   - Role reversal: expected power dynamics, hierarchies, or behaviors inverted
   - Absurd causality: a visual chain of reasoning that is obviously circular or flawed
   - Ironic visual contrast: negative situation depicted in a positive/celebratory style,
     or trivial situation framed with dramatic/heroic visual language
3. MEME AND CULTURAL TEMPLATE RECOGNITION
   - Known meme formats: Drake, Distracted Boyfriend, Expanding Brain, clown progression,
     "This is fine" dog, Wojak variants, Chinese internet meme conventions, etc.
   - Political cartoon conventions: symbolic stand-ins, labels on figures, editorial exaggeration
   - Reaction images: is the subject's expression/body language deployed ironically?
   - Comparison formats: split-panel "expectation vs. reality", before/after, escalation
4. EMBEDDED TEXT AS VISUAL ELEMENT
   If the image contains overlaid text, captions, or labels:
   - Does the embedded text contradict or ironically reframe what is visually depicted?
   - Is the text register (formal/slogan/casual) incongruous with the image content?
   - Does embedded text make an implicit claim that the image visually disproves?
5. CONFIDENCE CALIBRATION
   - 0.9–1.0: Unambiguous visual irony (recognizable sarcastic meme, clear visual contradiction)
   - 0.7–0.9: Strong visual incongruity; ironic intent highly plausible
   - 0.5–0.7: Ambiguous; image could be ironic or sincere without additional context
   - <0.5:   Image appears straightforward; sarcasm unlikely from visual alone
── OUTPUT ──────────────────────────────────────────────────────────
Return JSON in EXACTLY this schema:
{
  "decision": true or false,
  "confidence": 0.0–1.0,
  "explanation": "Identify the specific visual element(s) producing irony and explain the mechanism. If non-sarcastic, explain why the visual
content appears sincere."
}
```

Figure 9: Prompt template for the Image Sarcasm Agent (user turn).

System Prompt

You are an expert in multimodal discourse analysis, specializing in the emergent sarcasm and
irony that arises EXCLUSIVELY from the INTERACTION between an image and its accompanying text.
Your expertise covers:
- Semantic reversal: text and image assert contradictory propositions about the same referent
- Expectation subversion: text establishes a frame that the image violently disrupts as punchline
- Contextual reframing: the image supplies context that transforms literal text into ironic opposite
- Complementary amplification: neither element is sarcastic alone, but their combination creates
irony through mutual reinforcement of an absurd or contradictory claim
- Pragmatic gap exploitation: image fills the underspecified gap in ambiguous text ironically
- Platform conventions: how Twitter/Weibo/Reddit users deploy image-text pairs for sarcasm
CRITICAL CONSTRAINT: You must identify sarcasm that EMERGES from the combination.
If text alone is sarcastic OR image alone is sarcastic, those are handled by separate specialists.
Your task is cross-modal EMERGENT irony only — irony that neither element produces independently.
Output ONLY valid JSON. No markdown, no explanation outside the JSON.

Knowledge Context Block

—— RETRIEVED SARCASM KNOWLEDGE (via cross-modal similarity search) ——————————
The following cases were retrieved from a curated multimodal sarcasm knowledge base
using joint image+text similarity search. They represent confirmed sarcastic instances
where irony emerges specifically from image-text INTERACTION — not from either alone.
They serve as schema-level analogical evidence to overcome closed-world limitations.
HOW TO USE THESE CASES:
• Extract the CROSS-MODAL IRONY SCHEMA each case uses:
A. SEMANTIC REVERSAL — text claims X; image depicts ¬X for the same referent
B. EXPECTATION SUBVERSION — text sets a premise; image delivers ironic punchline
C. CONTEXTUAL REFRAMING — image context makes text's literal reading impossible
D. SELF-UNDERMINING JUXTAPOSITION — combination exposes contradiction or hypocrisy
• Recognise whether the CURRENT image-text pair instantiates the same schema.
• Overcome cultural/domain knowledge gaps: retrieved explanations reveal HOW specific
combinations produce sarcasm, providing patterns even when direct knowledge is limited.
• Generate candidates that go BEYOND surface similarity — identify the schema, then
apply it precisely to the current post's specific image-text tension.
{context}

Figure 10: Prompt templates for the Cross-modal Interaction Agent.

Stage 1: Candidate Generation Prompt

```
Analyze the interaction between the attached image and the post text, then generate
{n_candidates} distinct hypotheses describing potential cross-modal sarcasm mechanisms.
── INPUT ──────────────────────────────────────────────────────────
Text: {text}
[Image attached above]
────────────────────────────────────────────────────────────────────
{context_block}
── TASK: GENERATE {n_candidates} CANDIDATE CROSS-MODAL TENSION DESCRIPTIONS ────────
Each candidate must describe a DISTINCT hypothesis about HOW image and text interact
to create (or fail to create) ironic meaning. Cover different interaction mechanisms:
 A. SEMANTIC REVERSAL
    The text makes a claim (X) about a subject, and the image visually depicts the
    opposite (¬X) for that same subject.
    Pattern: "Text praises X" + "Image shows X failing/being terrible/contradicted"
 B. EXPECTATION SUBVERSION
    The text establishes a premise or expectation, and the image functions as an
    ironic punchline that completely undermines it. Requires reading text before image.
    Pattern: "Text sets up positive/neutral framing" + "Image delivers absurd reversal"
 C. CONTEXTUAL REFRAMING
    The image supplies contextual information (identity, location, scale, evidence)
    that makes the text's surface meaning impossible to take seriously, exposing irony.
    Pattern: "Text makes a claim" + "Image provides undeniable counter-evidence"
 D. SELF-UNDERMINING JUXTAPOSITION
    Text and image are each credible alone, but placed together they expose an internal
    contradiction, hypocrisy, or absurd gap in the poster's stated position.
    Pattern: "Text expresses value/position" + "Image depicts behavior contradicting it"
QUALITY CRITERIA:
 - Each candidate MUST describe tension requiring BOTH text AND image to perceive
 - Be specific: name WHAT in the image contradicts/subverts WHAT in the text
 - Avoid vague descriptions like "the image and text contrast" — name the mechanism
 - Candidates should cover different mechanisms (do not generate 3 versions of the same idea)
Return JSON in EXACTLY this schema:
{
 "candidates": [
   "Candidate A: [mechanism type] — [specific description of what in image vs text creates tension]",
   "Candidate B: [mechanism type] — [specific description of a different tension hypothesis]",
   "Candidate C: [mechanism type] — [specific description of a third distinct hypothesis]"
 ]
}
```

Figure 11: Prompt template for the Cross-modal Interaction Agent, Stage 1: generation of the candidate cross-modal tension descriptions $\mathcal{D}_{cand}$.

Stage 2: Final Decision Prompt

Make a final cross-modal sarcasm judgment for the following post, guided by the
best-fitting tension description identified in the previous analysis step.

── INPUT ──

Text: {text}

[Image attached above]

Best cross-modal tension description: {best_candidate}

── JUDGMENT CRITERIA ──

VERDICT SARCASTIC (true) when ALL of:

✓ The tension description holds up upon re-examining image and text together

✓ The combination produces a meaning clearly opposite to or mocking the literal text

✓ A reasonable audience familiar with the cultural context would read this as ironic

✓ The irony would NOT be apparent from text or image viewed in isolation

VERDICT NOT SARCASTIC (false) when ANY of:

✗ The tension is superficial or coincidental — no ironic intent evident

✗ Both elements are sincere and the combination does not undermine either's surface meaning

✗ The tension description, upon review, overstates or misidentifies the relationship

✗ The sarcasm, if present, originates entirely from the text or image alone (not emergent)

CONFIDENCE CALIBRATION:

- 0.9–1.0: Unambiguous cross-modal irony; any reader would perceive the emergent sarcasm

- 0.7–0.9: Strong image-text tension with clearly plausible ironic intent

- 0.5–0.7: Genuine ambiguity — combination could be sincere or ironic

- <0.5: Cross-modal tension is weak or absent; combination appears sincere

── OUTPUT ──

Return JSON in EXACTLY this schema:

{

"decision": true or false,

"confidence": 0.0–1.0,

"explanation": "Explain precisely how the image-text interaction produces (or fails to produce) emergent sarcasm. Reference specific visual and textual elements. Name the irony mechanism type (A/B/C/D)."

}

Figure 12: Prompt template for the Cross-modal Interaction Agent, Stage 2: conflict analysis over the selected description d^*, producing $\langle \hat{y}_c, p_c, R_c \rangle$.

System Prompt

You are the Orchestrator Agent in the IRAC (Irony-oriented Retrieval-Augmented multi-agent Collaboration) framework for multimodal sarcasm detection. You have direct access to the original image and text post, as well as independent analyses from three specialist agents (Text Agent, Image Agent, Cross-Modal Agent). Your role is to assign trust weights to each agent, synthesize their evidence, and deliver a final verdict.

User Prompt

<image>

── Original Post ──

Text: {text_post}

[The image above is the original post image — examine it directly.]

── Specialist Agent Analyses ──

[Text Agent] (text-only · verbal irony, Gricean violations, sentiment incongruity)

Verdict: {text_verdict} | Sarcasm probability p_t: {text_prob:.2f}

Analysis: {text_exp}

[Image Agent] (image-only · visual metaphor, meme templates, embedded text)

Verdict: {image_verdict} | Sarcasm probability p_i: {image_prob:.2f}

Analysis: {image_exp}

[Cross-Modal Agent] (emergent sarcasm from image-text interaction only)

Verdict: {cross_verdict} | Sarcasm probability p_c: {cross_prob:.2f}

Best tension description: {best_candidate}

Analysis: {cross_exp}

── Your Task ──

You have both the original image/text AND the three specialist analyses above.

Step 1 — Reasoning: Look at the image and text directly. What do you actually observe — the visual content, the caption's tone, any tension or harmony between the two? Then consider how well each agent's evidence holds up: is the cited mechanism (verbal inversion, visual metaphor, emergent interaction) genuinely present in THIS post, or does the agent over-generalise? Identify which source of evidence is most decisive.

Step 2 — Weights: Assign trust weights that reflect the strength of evidence in THIS specific post. Weights must be non-negative and sum exactly to 1.0.

Do NOT use fixed default values.

- If the sarcasm is carried by language alone → w_t should dominate
- If a visual element alone creates the irony → w_i should dominate
- If the tension only appears when image and text interact → w_c should dominate
- If an agent's reasoning is vague or unsupported for this post → give it low weight

Final sarcasm probability: P_final = w_t·p_t + w_i·p_i + w_c·p_c

Step 3 — Explanation: Describe the decisive evidence that drove your verdict. Ground it in specific elements you actually observed — particular words, phrases, visual details, or the image-text relationship. Avoid generic descriptions; a reader should be able to verify your evidence by looking at the post directly.

Figure 13: Prompt templates for the Orchestrator Agent.

User Prompt

Step 4 — Verdict: Output "yes" if P_final > 0.5, otherwise "no". When agents strongly disagree, validate the dissenting view before dismissing it. When all agents are uncertain, go with the reading best supported by direct observation.

Output your response as a single JSON object with exactly these four fields. Do not output anything outside the JSON block.

Below are two examples — note that the explanation style can vary as long as the evidence cited is specific and verifiable:

Example A (sarcasm driven by text-image interaction):

{

"analysis": "The caption reads 'a bit of a chill in the air' — on its face a mild understatement. The image, however, is an NWS weather map plastered with wind-chill warnings of −25 to −40°F and frostbite advisories. That gap between the breezy phrase and the genuinely extreme cold is where the sarcasm lives. The Text Agent's broad irony reading is essentially correct but undersells the specific mechanism; the Image Agent misses the point because it reads the map in isolation without connecting it to the caption. The Cross-Modal agent correctly locates the sarcasm in the contrast.",

"weights": {"w_t": 0.10, "w_i": 0.20, "w_c": 0.70},

"explanation": "The sarcasm comes entirely from placing 'a bit of a chill' against a weather map showing lethal wind-chills — the understatement is only readable as ironic because the image makes the severity concrete. Without the image, the phrase is ambiguous; without the caption, the map is just a weather report.",

"result": "yes"

}

Example B (not sarcastic — agents over-predicted):

{

"analysis": "The post shows a broken car and the caption says 'car trouble again'. The Image Agent and Cross-Modal Agent both flagged potential irony, but looking at the image there is no visual exaggeration, meme framing, or unexpected juxtaposition — it is a straightforward photo of a damaged vehicle. The Text Agent's note about 'again' implying repetition is factually accurate but 'again' here expresses genuine frustration, not performative mock-complaint. Nothing in the image subverts that reading.",

"weights": {"w_t": 0.45, "w_i": 0.30, "w_c": 0.25},

"explanation": "Literal frustration is the most parsimonious reading: the image shows actual damage, the word 'again' marks recurring bad luck rather than ironic distance, and there is no visual or tonal cue that reframes the caption as performance. The agents' irony detectors fired on surface features without a genuine inversion to support them.",

"result": "no"

}

Figure 14: Prompt templates for the Orchestrator Agent.