

Perceive, Localize, Reflect: Evidence-Anchored Multi-Agent Debate for MLLM Hallucination Mitigation

Yebo Wu^{1,2†}, Han Jin^{2†}, Wei Yang Bryan Lim^{1*}, Li Li^{2*}

¹Nanyang Technological University

²State Key Laboratory of IOTSC, University of Macau

bryan.limwy@ntu.edu.sg, llili@um.edu.mo

Abstract

Multimodal Large Language Models (MLLMs) demonstrate remarkable reasoning capabilities, yet they remain susceptible to hallucinations. While Multi-Agent Debate (MAD) presents a promising collaborative solution, unguided debates inevitably devolve into a stochastic martingale process¹ and expanding conversational contexts trigger severe visual attention decay. In this paper, we propose EVIAMAD, an Evidence-Anchored Multi-Agent Debate framework. To proactively expose internal hallucinations, EVIAMAD bifurcates perception into two distinct roles: a Global Observer capturing the holistic scene and a Local Observer scrutinizing salient details. When conflicts arise, a dedicated Analyzer intervenes, distilling the linguistic disagreement into a deterministic bimodal evidence anchor: a physically cropped visual patch of the disputed region paired with key semantic directives. During the subsequent reflection phase, a surgical re-allocation mechanism explicitly amplifies the attention weights of this anchor. By compelling agents to ground their reasoning in strictly verified evidence, EVIAMAD effectively elicits collective intelligence. Extensive evaluations demonstrate that EVIAMAD significantly mitigates hallucinations while elevating the reasoning capabilities of the multi-agent system.

1 Introduction

Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Bai et al., 2025) have fundamentally transformed the landscape of artificial intelligence by bridging visual perception and linguistic reasoning (Wu et al., 2026d,c,b; Xu et al., 2026b,c, 2025). However, they remain

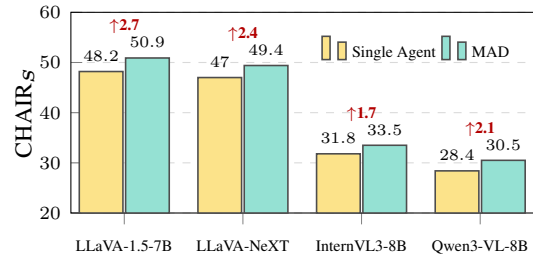


Figure 1: Performance comparison between the single-agent baseline and standard MAD on CHAIR. CHAIR_S evaluates the sentence-level object hallucination rate.

plagued by severe hallucinations (Leng et al., 2024; Wu et al., 2026a), wherein they frequently generate highly confident yet visually ungrounded content. This pervasive flaw undermines the reliability of MLLMs, posing a formidable bottleneck to their deployment in safety-critical domains (Zhao et al., 2024; Wu et al., 2026e; Xu et al., 2024).

To address this challenge, existing literature has explored a spectrum of strategies, spanning training-based alignment (Sun et al., 2024b; Zhao et al., 2025), post-hoc rectification (Zhou et al., 2024), and inference-time interventions (Suo et al., 2025; Wang et al., 2026). Despite achieving substantial improvements, these approaches remain confined to a single-agent paradigm, overlooking the collective intelligence enabled by multi-agent collaboration (Tran et al., 2025; Xu et al., 2026a).

Multi-Agent Debate (MAD) (Tillmann, 2025) presents a compelling framework to unlock this synergy through structured, iterative peer exchange. However, directly employing MAD for multimodal hallucination mitigation is surprisingly counterproductive. As shown in Figure 1, MAD increases CHAIR_S by up to 2.7%. This is because unguided debates induce a stochastic martingale over agent beliefs (Choi et al., 2025), where peer exchange without external factual grounding inherently fails to improve expected correctness. Furthermore, the

*Corresponding authors.

†Equal contribution.

¹A martingale describes a process with no expected drift. Applied to unguided MAD, it means peer exchange acts as a random walk; without factual grounding, debate merely redistributes uncertainty rather than improving correctness.

expanding textual context of multi-round debates severely dilutes visual attention, driving agents to abandon physical evidence and ultimately collapsing the debate into a fabricated consensus.

In this paper, we propose EVIAMAD, an Evidence-Anchored Multi-Agent Debate framework tailored for MLLM hallucination mitigation. Specifically, to proactively expose latent hallucinations, we bifurcate the reasoning process into two perceptual roles: a Global Observer capturing the holistic scene and a Local Observer scrutinizing salient details. Moreover, we introduce a Conflict Analyzer to shatter the stochastic martingale stagnation inherent in unguided debates. When the observers’ perspectives conflict, this Analyzer intervenes, translating the linguistic disagreement into a deterministic bimodal evidence anchor: a physically cropped visual patch of the disputed region coupled with its key semantic directives. During the subsequent reflection phase, a surgical attention reallocation mechanism explicitly upweights this bimodal anchor. By compelling agents to reason over strictly verified evidence, EVIAMAD effectively elicits collective intelligence and drives the system toward a factually grounded consensus. In summary, our main contributions are as follows:

- 1) We systematically diagnose the failure of standard MAD in multimodal hallucination mitigation, attributing it to two core pathologies: the stochastic martingale trap and context-induced visual attention decay.
- 2) We introduce EVIAMAD, an Evidence-Anchored Multi-Agent Debate framework, which features a specialized tri-agent system to proactively expose internal hallucinations and dictate the debate trajectory.
- 3) We conduct extensive evaluations on nine MLLMs and eleven benchmarks. The results demonstrate that EVIAMAD effectively mitigates hallucinations while elevating the reasoning capabilities of the multi-agent system.

2 A Closer Look at Multi-Agent Debate

In this section, we aim to answer the following two research questions (RQs):

- **RQ1:** Does standard MAD reliably reduce multimodal hallucinations in practice?
- **RQ2:** What root mechanisms cause MAD to break down in multimodal settings?

Model	Acc ^S (%)	Acc ^M (%)	σ (%)
LLaVA-1.5-7B	85.38	83.45 ($\downarrow 1.93$)	2.27
LLaVA-NeXT	86.17	84.36 ($\downarrow 1.81$)	2.11
InternVL3-8B	87.90	86.41 ($\downarrow 1.49$)	1.72
Qwen3-VL-8B	88.62	86.54 ($\downarrow 2.08$)	2.37

Table 1: Single-agent vs. MAD performance on POPE.

2.1 Hallucination Exacerbation under MAD

To answer **RQ1**, we evaluate a suite of MLLMs under two protocols: a Single-Agent baseline (“S”) and a standard MAD setup (“M”). Let $\mathcal{Q}$ denote the full evaluation set, with $\mathcal{Q}_V^P$ and $\mathcal{Q}_X^P$ representing the subsets of correctly and incorrectly answered queries under protocol $P \in \{S, M\}$. To comprehensively quantify the impact of MAD, we adopt two metrics. First, **Accuracy** ($\text{Acc}^P = |\mathcal{Q}_V^P|/|\mathcal{Q}|$) measures overall task performance. Second, we introduce the **Factual Shift Rate** (σ) to quantify the fraction of initially correct responses that are corrupted by the unguided debate process.

$$\sigma = |\mathcal{Q}_X^M \cap \mathcal{Q}_V^S| / |\mathcal{Q}_V^S|. \quad (1)$$

Finding 1: MAD consistently underperforms a single agent. As shown in Table 1, Acc^M falls below Acc^S across all models. Specifically, the performance degradation ranges from 1.49% (InternVL3-8B) to a striking 2.08% (Qwen3-VL-8B). The long-held intuition that “more agents $\Rightarrow$ better reasoning” completely breaks down in the multimodal domain.

Finding 2: Peer exchange erodes factual correctness. Table 1 reveals that σ reaches up to 2.37%, indicating that a proportion of queries originally answered correctly in $\mathcal{Q}_V^S$ is erroneously shifted into $\mathcal{Q}_X^M$ during peer exchange. This degradation suggests that agents are compelled to abandon their initial visual grounding, conforming instead to a hallucinated textual consensus.

2.2 Probing the Root Causes of Failure

To address **RQ2**, we dissect the failure factors through both theoretical and empirical lenses.

2.2.1 The Stochastic Martingale Trap

Following the Bayesian framework for MAD (Choi et al., 2025), we model each agent’s internal state as a belief distribution over candidate answers governed by a Dirichlet-Compound-Multinomial process, which captures both an agent’s internal uncertainty and the stochasticity of its generation process. Formally, at round t , each agent i maintains a belief

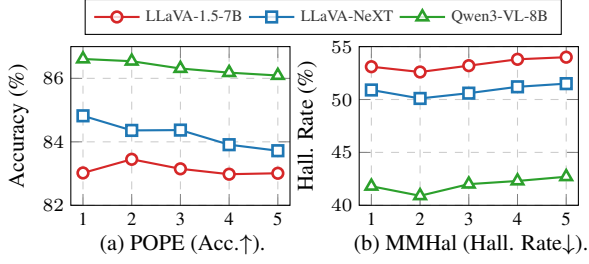


Figure 2: Average performance of the multi-agent system across debate rounds.

vector $\alpha_{i,t} \in \mathbb{R}_+^K$ over K candidate answers. To produce a response, the agent first samples a probability vector $\theta_{i,t} \sim \text{Dirichlet}(\alpha_{i,t})$, then draws $y_{i,t} \sim \text{Categorical}(\theta_{i,t})$. By Bayesian conjugacy, the marginal probability of generating answer k is:

$$P(y_{i,t} = k \mid \alpha_{i,t}) = \frac{\alpha_{i,t}^{(k)}}{\sum_{j=1}^K \alpha_{i,t}^{(j)}}. \quad (2)$$

Upon observing peer responses, agent i performs a Bayesian posterior update:

$$\alpha_{i,t} = \alpha_{i,t-1} + \mathbf{c}_{i,t}, \quad (3)$$

where $\mathbf{c}_{i,t} \in \mathbb{N}^K$ is the count vector tallying the responses of agent i 's observable neighbors, denoted as $\mathcal{N}(i)$, at round $t-1$. Let $p_{i,t} := \alpha_{i,t}^{(1)} / \sum_k \alpha_{i,t}^{(k)}$ denote agent i 's belief in the correct answer (indexed as $k = 1$) at round t .

Martingale Behavior of MAD. For any agent i at round $t > 0$, if the average neighbor belief equals the agent's own belief,

$$\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} p_{j,t-1} = p_{i,t-1}, \quad (4)$$

then the sequence $\{p_{i,t}\}_{t \geq 0}$ forms a *martingale*:

$$\mathbb{E}[p_{i,t} \mid \alpha_{t-1}] = p_{i,t-1}. \quad (5)$$

This property dictates that unguided debate cannot improve an agent's belief in the correct answer. Lacking a deterministic visual anchor, peer exchanges are zero-sum in expectation, merely redistributing linguistic uncertainty rather than converging on factual truth. To empirically verify this, we track p_t (estimated via mean per-agent performance) over debate rounds. As shown in Figure 2, the resulting trajectories exhibit no positive drift. Instead, they manifest a persistent negative drift: POPE accuracy declines (e.g., from 84.82% to 83.72% for LLaVA-NeXT), while MMHal hallucination rates worsen (climbing to 54.0% for

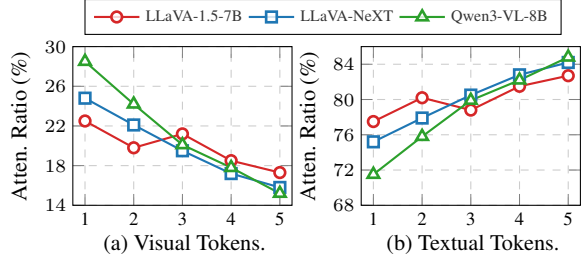


Figure 3: Attention allocation dynamics across rounds.

LLaVA-1.5-7B). This degradation confirms that unguided textual exchange cannot break perceptual deadlocks; rather, prolonged debate introduces cascading errors, degenerating into a stochastic random walk entirely hijacked by language priors.

2.2.2 The Pathology of Visual Attention Decay

To further investigate *why* unguided debate precipitates the continuous performance degradation observed in Figure 2, we track the internal attention mass allocated to visual and textual tokens across debate rounds. As depicted in Figure 3, we observe a striking phenomenon: as the debate progresses, MLLMs experience profound **visual attention decay**. The models' finite attention budget is progressively hijacked by text tokens, compelling agents to gradually abandon the original visual evidence.

This decay arises because, as agents append peer responses to their context, the expanding textual history overwhelms the model's finite self-attention capacity. Consequently, the attention allocated to visual tokens plummets (e.g., dropping from 28.5% to 15.2% for Qwen3-VL-8B by round 5), while textual attention surges (71.5% $\rightarrow$ 84.8%). Deprived of adequate grounding in the visual evidence, agents are forced to fall back on language priors, inevitably inducing and reinforcing hallucinations.

Summary: The Vicious Cycle. The martingale trap renders unguided debate a zero-sum game incapable of improving correctness, while visual attention decay progressively blinds agents to the factual signals that could shatter this deadlock. These insights directly motivate our EVIAMAD, which effectively breaks this cycle by anchoring debates in deterministic visual evidence.

3 Methodology: EVIAMAD

As illustrated in Figure 4, EVIAMAD operates through a three-stage pipeline: **Perception**, **Localization**, and **Reflection** to proactively surface and resolve perceptual conflicts.

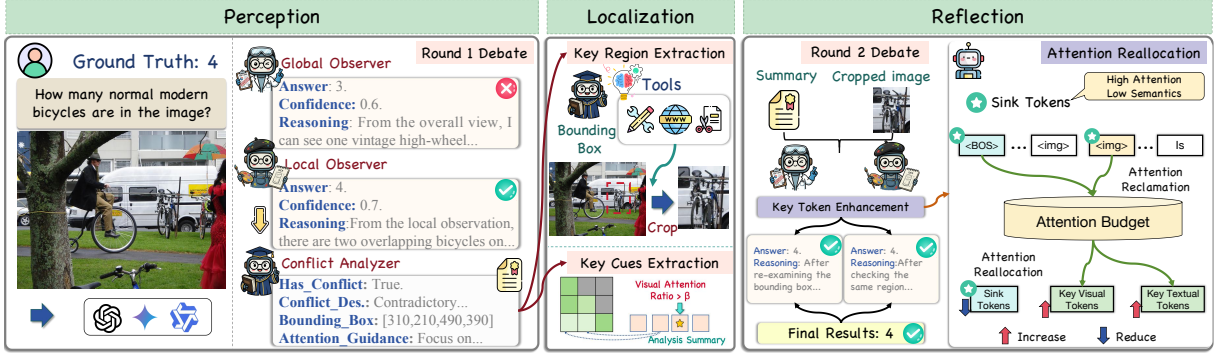


Figure 4: Overview of EVIAMAD. **(Left) Perception:** A Global Observer and a Local Observer analyze the input from complementary stances to proactively expose latent hallucinations. Upon detecting a contradiction, a Conflict Analyzer intervenes to summarize the disagreement and predict a precise spatial Bounding Box alongside explicit Attention Guidance. **(Middle) Localization:** The linguistic disagreement is translated into a bimodal evidence anchor: the disputed visual region is physically cropped, and key semantic directives are extracted from the text. **(Right) Reflection:** Attention is reclaimed from uninformative sink tokens and redistributed to the bimodal anchor.

3.1 Perception: Exposing Latent Conflicts

The martingale trap is fundamentally rooted in *epistemic homogeneity*: when all agents share an identical perceptual context, their prior distributions over candidate answers are highly correlated. Consequently, peer exchange merely redistributes this shared uncertainty without injecting any new factual signals. To address this, EVIAMAD engineers *asymmetric* perceptual stances to proactively surface the internal inconsistencies inherent to hallucinations. If an agent hallucinates a detail, the error often manifests as a fine-grained discrepancy that may be invisible at the holistic scene level but detectable upon targeted local scrutiny. This engineered asymmetry serves as the informational wedge that breaks the zero-sum deadlock.

Bifurcated Perceptual Roles. To operationalize this asymmetry, EVIAMAD deliberately bifurcates visual perception into two complementary roles:

- **Global Observer ($\mathcal{A}_G$):** Operating on the full-resolution input, $\mathcal{A}_G$ performs holistic, top-down reasoning. While adept at capturing macroscopic scene semantics, this stance structurally overlooks fine-grained details.
- **Local Observer ($\mathcal{A}_L$):** Instructed to adopt a bottom-up, detail-first reasoning strategy, $\mathcal{A}_L$ acts as a forensic scrutinizer. It provides a complementary factual perspective that frequently disagrees with the global view when fine-grained visual evidence is ambiguous.

Both observers independently output a structured (Answer, Confidence, Reasoning) tuple. This deliberate bifurcation transforms the unresolvable,

latent uncertainty of a homogeneous debate into a concrete, diagnosable factual disagreement.

The Conflict Analyzer ($\mathcal{A}_C$). When $\mathcal{A}_G$ and $\mathcal{A}_L$ yield contradictory answers, the Conflict Analyzer intervenes. Rather than passively moderating the debate, $\mathcal{A}_C$ acts as an active factual arbiter, performing a rigorous analysis of the disagreement to produce a four-part structured output:

- **Has_Conflict** $\in \{\text{True}, \text{False}\}$: A flag verifying the existence of a factual contradiction.
- **Conflict_Description**: A concise summary distilling the linguistic disagreement.
- **Bounding_Box** $\in \mathbb{R}^4$: Predicted coordinates $[x_1, y_1, x_2, y_2]$ that precisely localize the disputed visual region.
- **Attention_Guidance**: A targeted directive instructing the observers on *what* specific visual features to re-examine.

This structured output distills a purely textual dispute into a deterministic bimodal evidence anchor, driving the subsequent Localization stage.

3.2 Localization: Anchoring Factual Evidence

Once a perceptual conflict is exposed, the system transitions to the Localization phase. EVIAMAD executes two extraction operations to instantiate the bimodal evidence anchor.

Key Region Extraction. EVIAMAD invokes external tools to *physically crop* the image $\mathcal{V}$ according to the coordinates. The resulting localized patch, $\mathcal{V}_{bbox}$, unambiguously isolates the disputed

visual locus and strips away irrelevant background noise. This process acts as a visual magnifying glass, compelling the agents to directly process the raw, high-fidelity pixels of the contested area.

Key Cues Extraction. While cropping resolves spatial ambiguity, the Attention_Guidance directive remains susceptible to semantic dilution. To isolate these pivotal cues, EVIAMAD computes the cross-modal attention ratio for each token j within the guidance. Specifically, we calculate the fraction of j 's total outgoing attention that is directed toward the visual tokens. A token is retained in the critical cue set $\mathcal{I}_{text}$ if this ratio exceeds β :

$$\mathcal{I}_{text} = \left\{ j \mid \frac{a_j^{\text{vis}}}{\sum_k a_{j,k}} > \beta \right\}, \quad (6)$$

where $a_j^{\text{vis}} = \sum_{k \in \mathcal{V}} a_{j,k}$ denotes the attention mass that token j allocates to all image tokens $\mathcal{V}$, and the denominator represents its total attention over the full context. This filtering rigorously isolates the semantic core of the diagnostic directive.

Together, the cropped visual patch $\mathcal{V}_{bbox}$ and the distilled textual cue set $\mathcal{I}_{text}$ constitute the **bimodal evidence anchor**. This anchor acts as a compact, precise, and verifiable grounding signal to explicitly steer the subsequent Reflection stage.

3.3 Reflection: Surgical Attention Calibration

Even with the bimodal evidence, context-induced visual attention decay remains a critical threat. As the multi-round debate unfolds, the expanding context enables text tokens to hijack the agents' finite attention budget, marginalizing the carefully extracted evidence. To address this challenge, we employ a lightweight, training-free attention reallocation mechanism to actively defend the bimodal evidence anchor against context dilution.

Bimodal Target Set. In the reflection stage, each Observer receives both the original image and the cropped image patch to preserve global context while emphasizing the disputed region. Let $\mathcal{I}_{bbox}$ denote the visual tokens corresponding to the cropped image patch $\mathcal{V}_{bbox}$, and let $\mathcal{I}_{text}$ be the key cues. They jointly constitute the bimodal target set:

$$\mathcal{I}_{target} = \mathcal{I}_{bbox} \cup \mathcal{I}_{text}. \quad (7)$$

By enhancing both simultaneously, EVIAMAD enforces a synergized visual-textual focus that compels agents to reason jointly over the disputed region and its semantic interpretation.

Attention Reclamation from Sink Tokens. Due to softmax normalization, directly adding attention to target tokens necessitates removing it from elsewhere. Instead, EVIAMAD leverages a well-documented property of the attention mechanism: a small set of **sink tokens** (e.g., <BOS>) accumulate disproportionately large attention scores despite encoding minimal semantic content (Kang et al., 2025). This presents a unique opportunity to safely reclaim this wasted attention and reallocate it to the verified evidence. Formally, following the sink token identification strategy proposed in VAR (Kang et al., 2025) (Appendix A), let $\tilde{\mathcal{I}}$ denote the set of identified sink tokens. For each attention head, the reclaimable attention budget is defined as:

$$\Omega = \sum_{j \in \tilde{\mathcal{I}}} \gamma \cdot \alpha_{i,j}, \quad (8)$$

where $\alpha_{i,j}$ is the original attention score from query i to key j , and $\gamma \in (0, 1)$ is the reclamation ratio controlling how aggressively the sink tokens are dampened. The attention allocated to sink tokens is correspondingly reduced:

$$\tilde{\alpha}_{i,j} = (1 - \gamma) \cdot \alpha_{i,j}, \quad \forall j \in \tilde{\mathcal{I}}. \quad (9)$$

Attention Redistribution. The harvested budget Ω is redistributed to the target tokens in proportion to their *original* attention scores:

$$\alpha_{i,k}^* = \alpha_{i,k} + \Omega \cdot \frac{\alpha_{i,k}}{\sum_{m \in \mathcal{I}_{target}} \alpha_{i,m}}, \quad \forall k \in \mathcal{I}_{target}. \quad (10)$$

This process incurs no additional forward passes, making it a lightweight and efficient augmentation.

Achieving Factual Convergence. Upon entering the second round of debate, the two Observers strictly ground their reasoning in the enhanced bimodal evidence anchor, thereby effectively shattering the stochastic martingale trap and overcoming the visual attention decay. Ultimately, this synergized intervention drives the system to converge on a factually grounded consensus.

4 Experiments

4.1 Experimental Settings

Evaluation Models and Baselines. We evaluate EVIAMAD across nine representative MLLMs: LLaVA-1.5-7B (Liu et al., 2023), LLaVA-NeXT, Qwen2.5-VL-7B (Hui et al., 2024), Gemma3-12B (Team et al., 2024), InternVL3 (Zhu et al.,

Base Agent	POPE		CHAIR		MME		MMHal	
	Acc.↑	F1↑	CHAIR _S ↓	CHAIR _I ↓	Perception↑	Cognition↑	Total↑	Hall. Rate↓
LLaVA-1.5-7B	85.4	85.2	48.2	14.5	1243.3	255.3	1498.6	51.9
+DoLa	85.6	85.5	47.3	14.0	1249.4	255.2	1504.6	51.3
+VCD	86.2	86.3	46.5	13.8	1254.6	256.8	1511.4	50.6
+EVIAMAD	87.2	87.7	43.7	12.8	1272.5	261.2	1533.7	46.3
LLaVA-NeXT	86.2	86.0	47.0	14.1	1254.0	257.5	1511.5	49.8
+DoLa	86.2	86.4	46.7	13.9	1262.5	257.9	1520.4	48.6
+VCD	87.1	87.3	45.4	13.5	1273.1	261.4	1534.5	46.3
+EVIAMAD	88.1	88.0	41.4	12.0	1282.3	263.0	1545.3	44.2
Qwen2.5-VL-7B	86.9	86.8	42.9	13.2	1268.4	260.2	1528.6	44.3
+DoLa	87.2	87.1	42.6	12.9	1272.8	261.5	1534.3	44.0
+VCD	87.5	87.4	41.1	12.3	1277.6	262.2	1539.8	43.6
+EVIAMAD	88.1	87.9	40.5	11.9	1285.8	264.4	1550.2	42.7
Gemma3-12B	88.5	88.7	28.9	9.3	1334.5	279.0	1613.5	39.8
+DoLa	88.7	88.9	28.6	9.0	1337.6	280.9	1618.5	39.4
+VCD	89.1	89.5	27.2	8.5	1340.9	282.5	1623.4	38.6
+EVIAMAD	89.9	90.2	26.7	7.8	1346.5	283.8	1630.3	37.0
InternVL3-8B	87.9	88.3	31.8	9.6	1383.7	284.0	1667.7	40.7
+DoLa	88.0	88.5	30.2	9.3	1388.1	284.7	1672.8	40.4
+VCD	88.5	88.8	28.8	9.1	1392.8	286.0	1678.8	39.9
+EVIAMAD	89.2	89.3	27.3	8.1	1400.9	288.2	1689.1	38.8
Qwen3-VL-8B	88.6	88.5	28.4	9.0	1402.7	288.0	1690.7	38.6
+DoLa	88.7	88.8	28.2	8.8	1406.1	288.9	1695.0	38.4
+VCD	89.2	89.3	27.4	8.6	1409.8	289.6	1699.4	38.0
+EVIAMAD	89.5	89.9	26.9	7.9	1416.0	290.4	1706.4	37.1

Table 2: Performance comparison of EVIAMAD against single-agent baselines on POPE, CHAIR, MME, and MMHal benchmarks. The best results are highlighted in **bold**.

2025) (8B and 14B), and the Qwen3-VL (Bai et al., 2025) series (2B, 4B, and 8B). Beyond the base models, we compare EVIAMAD with two representative single-agent hallucination mitigation baselines, DoLa (Chuang et al., 2024) and VCD (Leng et al., 2024).

Evaluation Benchmarks. To comprehensively assess our framework, we employ two complementary benchmark suites: **1) Multimodal Hallucination Benchmarks:** POPE (Li et al., 2023b), CHAIR (Rohrbach et al., 2018), MME (Fu et al., 2025), and MMHal (Sun et al., 2024b); and **2) General Reasoning Benchmarks:** MM-Vet (Yu et al., 2024), GQA (Hudson and Manning, 2019), VQA^{v2} (Goyal et al., 2017), MMBench (Liu et al., 2025), SEED-Bench (Li et al., 2023a), VizWiz (Gurari et al., 2018), and MMVP (Tong et al., 2024).

4.2 Main Results

Hallucination Mitigation. Table 2 shows that EVIAMAD consistently outperforms the single-agent baselines across all evaluated settings. On LLaVA-1.5-7B, EVIAMAD improves POPE ac-

curacy to 87.2% and reduces CHAIR_S to 43.7%, surpassing the strongest VCD baseline by 1.0% and 2.8%, respectively. Similar gains appear on LLaVA-NeXT, where EVIAMAD lowers CHAIR_S from VCD’s 45.4% to 41.4% and further reduces the MMHal hallucination rate to 44.2%. The trend also holds for stronger models such as Gemma3-12B, InternVL3-8B, and Qwen3-VL-8B, where EVIAMAD achieves the best results across all benchmarks. These results indicate that our bimodal evidence anchors successfully compel agents to ground their reasoning in verifiable visual truth, regardless of the model architecture.

Reasoning Enhancement. Beyond hallucination mitigation, EVIAMAD also enhances general multimodal reasoning. As shown in Table 3, EVIAMAD achieves the highest average score across all evaluated models, consistently outperforming the single-agent baselines. For instance, on LLaVA-1.5-7B, EVIAMAD improves the average score from 50.1% to 52.6%, exceeding VCD by 1.1%. Notably, on Qwen3-VL-8B, EVIAMAD still attains the best average score of 69.8%, indicating

Base Agent	MM-Vet $\uparrow$	GQA $\uparrow$	VQA $^{v2}\uparrow$	MMB $\uparrow$	SEED $\uparrow$	VizWiz $\uparrow$	MMVP $\uparrow$	Avg. $\uparrow$
LLaVA-1.5-7B	32.9	61.5	77.9	64.8	59.6	48.1	5.6	50.1
+DoLa	33.3	62.7	78.1	65.1	59.7	48.5	7.9	50.8
+VCD	33.9	63.6	78.5	66.0	60.5	48.9	9.4	51.5
+EVIAMAD	34.8	64.8	79.5	66.9	61.6	49.4	11.5	52.6
LLaVA-NeXT	33.4	62.9	78.3	66.1	61.5	49.0	21.9	53.3
+DoLa	33.8	63.5	78.5	66.4	61.6	49.4	24.2	54.0
+VCD	34.4	63.9	78.9	67.3	62.1	49.8	25.9	54.7
+EVIAMAD	35.0	64.2	80.8	68.1	62.4	50.7	27.5	55.5
InternVL3-8B	55.2	68.7	82.1	78.3	67.8	55.2	66.7	67.7
+DoLa	55.6	69.0	82.3	78.4	67.9	55.3	69.0	68.4
+VCD	56.2	69.6	82.7	78.7	68.1	55.7	70.5	69.1
+EVIAMAD	56.7	70.1	83.4	79.1	68.5	56.1	72.8	69.5
Qwen3-VL-8B	58.4	71.2	83.0	77.8	67.4	54.2	65.9	68.3
+DoLa	58.3	71.5	83.2	78.1	67.5	54.6	68.2	69.0
+VCD	58.4	71.7	83.6	79.0	68.0	55.0	69.7	69.7
+EVIAMAD	58.1	72.4	83.9	79.4	68.2	55.2	71.7	69.8

Table 3: Performance comparison of EVIAMAD against single-agent baselines on general reasoning tasks.

Method	POPE		CHAIR	
	Acc. $\uparrow$	F1 $\uparrow$	CHAIR $_S\downarrow$	CHAIR $_I\downarrow$
LLaVA-1.5-7B				
+ EVIAMAD	87.2	87.7	43.7	12.8
w/o Roles	83.5	83.5	50.9	15.1
w/o Analyzer	85.9	86.1	46.3	13.6
w/o Enhance.	85.4	85.6	47.4	14.1
LLaVA-NeXT				
+ EVIAMAD	88.1	88.0	41.4	12.0
w/o Roles	84.4	84.2	49.4	14.9
w/o Analyzer	87.4	87.2	43.2	13.7
w/o Enhance.	86.8	86.5	45.5	13.2

Table 4: Ablation study of EVIAMAD.

that evidence-anchored debate strengthens the reasoning capabilities of the multi-agent system.

4.3 Ablation Study

To verify the effectiveness of EVIAMAD, we conduct an ablation study on three core mechanisms: 1) *w/o Roles*, which removes distinct perceptual roles; 2) *w/o Analyzer*, which omits the Conflict Analyzer; and 3) *w/o Enhancement*, which disables targeted attention amplification. As shown in Table 4, all variants underperform compared to EVIAMAD, confirming the necessity of each component. In particular, *w/o Roles* causes the largest degradation, with POPE accuracy dropping by up to 3.7%, showing that perceptual asymmetry is crucial for exposing latent hallucinations. Removing the Analyzer weakens factual grounding (e.g., CHAIR $_S$ increases from 41.4% to 43.2% on LLaVA-NeXT), while disabling attention enhancement leads to visible attention dilution (e.g., a 1.8% POPE accuracy

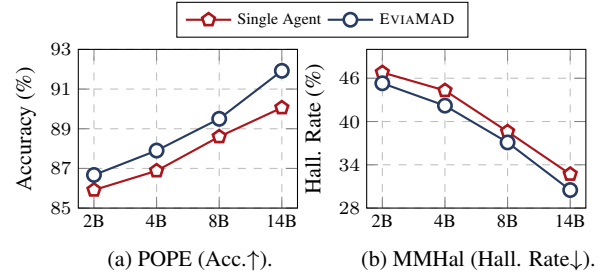


Figure 5: Scaling analysis on POPE and MMHal.

drop on LLaVA-1.5-7B). These results demonstrate that role specialization, evidence extraction, and attention calibration are complementary components that jointly enable factually grounded convergence.

4.4 More Analysis

Scaling Analysis. To investigate the scalability of EVIAMAD, we evaluate its performance across four distinct parameter scales, instantiated by the Qwen3-VL series (2B, 4B, and 8B) alongside InternVL3 (14B). As depicted in Figure 5, EVIAMAD yields consistent and robust gains over the single-agent baseline across every evaluated scale. For example, our framework drives absolute improvements in POPE accuracy by 0.76%, 1.02%, 0.90%, and 1.86% for the 2B, 4B, 8B, and 14B models, respectively. These scale-invariant enhancements demonstrate that our method fundamentally cures the structural pathologies of unguided debate, proving that explicit visual anchoring delivers an orthogonal cognitive improvement entirely independent of mere parameter scaling.

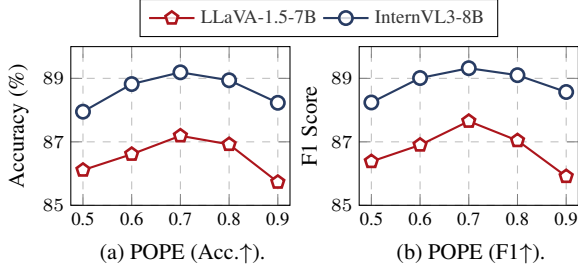


Figure 6: Impact analysis of the key-cues threshold β .

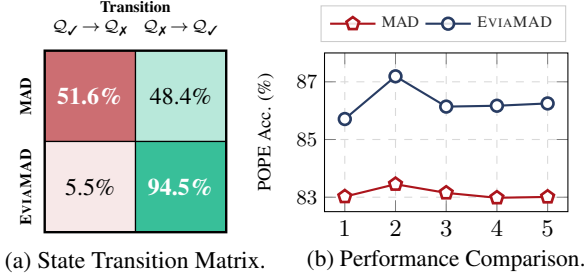


Figure 7: Empirical analysis of the martingale trap (LLaVA-1.5-7B, POPE). (a) State transition matrices. (b) Accuracy dynamics across five debate rounds.

Sensitivity Analysis. We evaluate the sensitivity of EVIAMAD to the key-cues threshold β , ranging from 0.5 to 0.9. Figure 6 demonstrates that performance remains robust across the moderate range, peaking consistently at $\beta = 0.7$ for both models. This behavior is expected: relaxing the threshold ($\beta \rightarrow 0.5$) amplifies excess textual tokens and introduces semantic noise, whereas aggressive filtering ($\beta \rightarrow 0.9$) obliterates the essential semantic anchor. This dynamic confirms that β strictly regulates the precision-coverage trade-off of textual evidence.

Breaking the Martingale Trap. To empirically verify that EVIAMAD escapes the stochastic martingale trap, we analyze agent state transitions from Round 1 to Round 2. As illustrated in Figure 7(a), the factual transitions under standard MAD are nearly symmetric: 51.6% of shifts are regressions ($Q_{\checkmark} \rightarrow Q_{\times}$) and 48.4% are corrections ($Q_{\times} \rightarrow Q_{\checkmark}$). This near-identical balance visualizes the zero-sum nature of the martingale property. In contrast, EVIAMAD triggers a profound directional rectification. Upon introducing bimodal evidence anchors, an overwhelming 94.5% of transitions become positive corrections, thereby validating that our framework shatters the stochastic martingale trap. Furthermore, Figure 7(b) illustrates that EVIAMAD reaches peak performance at Round 2 (87.19%). Performance subsequently saturates as more rounds offer no further visual grounding, only risking the accumulation of noise.

These results confirm that our framework mitigates hallucinations with superior convergence speed.

5 Related Work

MLLM Hallucination Mitigation. Training-based methods (Sun et al., 2024b) utilize alignment techniques like DPO to penalize hallucinatory outputs. For instance, HA-DPO (Zhao et al., 2025) reframes hallucination reduction as a preference selection task. However, they incur prohibitive computational costs and rely heavily on meticulously curated datasets. Post-hoc rectification (Zhou et al., 2024) leverages external tools to verify and rewrite inaccurate descriptions, such as Woodpecker’s (Yin et al., 2024) multi-stage correction pipeline. Nevertheless, this strategy introduces significant inference latency. Inference-time interventions (Huang et al., 2024) offer a training-free alternative by calibrating outputs or manipulating attention scores, exemplified by VCD’s (Leng et al., 2024) contrastive decoding over distorted visual inputs. However, this method is prone to injecting unintended noise into the reasoning process. Departing from these single-agent methods, EVIAMAD leverages structured debate to mitigate multimodal hallucinations.

Multi-Agent Debate. Multi-Agent Debate (MAD) (Tillmann, 2025) harnesses collective intelligence to enhance LLM reasoning, achieving significant success across diverse domains such as mathematical reasoning (Zhang and Xiong, 2025) and software issue resolution (Li et al., 2025). Subsequent works have optimized debate efficiency through structured communication topologies. For instance, CortexDebate (Sun et al., 2025) constructs a sparse communication graph, restricting exchanges to strictly beneficial peers. However, unguided debate induces a stochastic martingale process over agents’ belief trajectories. Moreover, the expanding textual context in multimodal settings dilutes visual attention. EVIAMAD resolves these challenges through evidence-anchored debate.

6 Conclusion

In this paper, we propose EVIAMAD, an Evidence-Anchored Multi-Agent Debate framework. By bifurcating perceptual roles and deploying a dedicated Conflict Analyzer, EVIAMAD distills linguistic disagreements into deterministic bimodal evidence anchors. Combined with an attention reallocation mechanism, our framework effectively

compels agents to ground their reasoning in verifiable visual truth. Extensive evaluations demonstrate the effectiveness of EVIAMAD.

Limitations

While EVIAMAD significantly mitigates multi-modal hallucinations and enhances reasoning, it introduces an inherent trade-off in computational overhead. Because the framework necessitates multi-round, multi-agent interactions, the inference latency and token consumption are naturally higher than those of a standard single-agent forward pass. Future work will explore more efficient debate stopping criteria and early-exit mechanisms to optimize the inference speed while preserving the cognitive synergy of the multi-agent system.

Ethical Considerations

In this work, all empirical evaluations are conducted on publicly available, open-source benchmarks (e.g., POPE, CHAIR), ensuring no collection or exposure of sensitive personal information. All datasets and pre-trained MLLMs are utilized strictly within their intended research scopes and in full compliance with their respective open-source licenses. Because EVIAMAD operates as a training-free, inference-time intervention, it explicitly avoids modifying or redistributing any proprietary model weights, thereby strictly adhering to usage permissions and minimizing computational carbon footprint compared to retraining.

From a societal perspective, our methodology actively mitigates multimodal hallucinations stemming from cognitive inertia and perceptual deadlocks. By anchoring multi-agent debates in deterministic visual evidence, EVIAMAD enhances the safety and reliability of MLLM deployments in critical real-world applications. However, we acknowledge that because EVIAMAD does not alter the underlying parameters of the base models, the system may still inherit pre-existing biases or the potential for generating toxic content embedded in the pre-training corpora. Aside from this inherent limitation of foundation models, to the best of our knowledge, there are no additional ethical concerns associated with our proposed framework.

Acknowledgements

This work is supported in part by the Science and Technology Development Fund of Macau (0107/2024/RIA2, 0061/2025/RIB2), Joint Science

and Technology Research Project with Hong Kong and Macau in Key Areas of Nansha District’s Science and Technology Plan (EF2024-00180-IOTSC) and the Multi-Year Research Grant of University of Macau (MYRG-GRG2023-00211-IOTSC-UMDF, MYRG-GRG2024-00180-IOTSC). This work is also supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award MOE-T2EP20125-0005).

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. 2025. [Debate or vote: Which yields better decisions in multi-agent large language models?](#) In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 101732–101764. Curran Associates, Inc.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. [Dola: Decoding by contrasting layers improves factuality in large language models](#). In *International Conference on Learning Representations*, volume 2024, pages 54158–54183.
- Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267. Curran Associates, Inc.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xianwu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. [Mme: A comprehensive evaluation benchmark for multimodal large language models](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. [Vizwiz grand challenge: Answering visual questions from blind people](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. [Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.
- Drew A. Hudson and Christopher D. Manning. 2019. [Gqa: A new dataset for real-world visual reasoning and compositional question answering](#). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. [See what you are told: Visual attention sink in large multimodal models](#). In *International Conference on Learning Representations*, volume 2025, pages 87676–87703.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. [Mitigating object hallucinations in large vision-language models through visual contrastive decoding](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Han Li, Yuling Shi, Shaoxin Lin, Xiaodong Gu, Heng Lian, Xin Wang, Yantao Jia, Tao Huang, and Qianxiang Wang. 2025. Swe-debate: Competitive multi-agent debate for software issue resolution. *arXiv preprint arXiv:2507.23348*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2025. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, pages 216–233, Cham. Springer Nature Switzerland.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. [Object hallucination in image captioning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.
- Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024a. [Massive activations in large language models](#). In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.
- Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. 2025. [CortexDebate: Debating sparsely and equally for multi-agent debate](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9503–9523, Vienna, Austria. Association for Computational Linguistics.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024b. [Aligning large multi-modal models with factually augmented RLHF](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.
- Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. [Octopus: Alleviating hallucination via dynamic contrastive decoding](#). In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29904–29914.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Arne Tillmann. 2025. Literature review of multi-agent debate for problem-solving. *arXiv preprint arXiv:2506.00066*.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. [Eyes wide shut? exploring the visual shortcomings of multimodal llms](#). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.
- Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.
- Yujun Wang, Jinhe Bi, Soren Pirk, Yunpu Ma, et al. 2026. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10306–10314.
- Yebo Wu, Han Jin, Zhijiang Guo, and Li Li. 2026a. [Spotlight and shadow: Attention-guided dual-anchor introspective decoding for MLLM hallucination mitigation](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 13219–13233,

- San Diego, California, United States. Association for Computational Linguistics.
- Yebo Wu, Jingguang Li, Zhijiang Guo, and Li Li. 2026b. [Developmental federated tuning: A cognitive-inspired paradigm for efficient LLM adaptation](#). In *The Fourteenth International Conference on Learning Representations*.
- Yebo Wu, Jingguang Li, Zhijiang Guo, and Li Li. 2026c. [Don't reinvent the wheel, just realign the spokes: Resource-efficient federated fine-tuning via rank-wise expert assembly](#). In *Forty-third International Conference on Machine Learning*.
- Yebo Wu, Feng Liu, Ziwei Xie, Zhiyuan Liu, Changwang Zhang, Jun Wang, and Li Li. 2026d. Tsembed: Unlocking task scaling in universal multimodal embeddings. *arXiv preprint arXiv:2603.04772*.
- Yebo Wu, Chunlin Tian, Jingguang Li, He Sun, KaHou Tam, Zhanting Zhou, Haicheng Liao, Jing Xiong, Zhijiang Guo, Li Li, and Cheng zhong Xu. 2026e. [A survey on federated fine-tuning of large language models](#). *Transactions on Machine Learning Research*. Survey Certification.
- Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, and Shouling Ji. 2026a. [When agents "misremember" collectively: Exploring the mandela effect in LLM-based multi-agent systems](#). In *The Fourteenth International Conference on Learning Representations*.
- Naen Xu, Changjiang Li, Tianyu Du, Minxi Li, Wenjie Luo, Jiacheng Liang, Yuyuan Li, Xuhong Zhang, Meng Han, Jianwei Yin, et al. 2024. Copyrightmeter: Revisiting copyright protection in text-to-image models. *arXiv preprint arXiv:2411.13144*.
- Naen Xu, Jiayi Sheng, Changjiang Li, Chunyi Zhou, Yuyuan Li, Tianyu Du, Jun Wang, Zhihui Fu, Jinbao Li, and Shouling Ji. 2026b. ["I see what you did there": Can large vision-language models understand multimodal puns?](#) In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9786–9805, San Diego, California, United States. Association for Computational Linguistics.
- Naen Xu, Jinghuai Zhang, Changjiang Li, Hengyu An, Chunyi Zhou, Jun Wang, Boyu Xu, Yuyuan Li, Tianyu Du, and Shouling Ji. 2026c. Bridging the copyright gap: Do large vision-language models recognize and respect copyrighted content? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35949–35957.
- Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. 2025. [VideoEraser: Concept erasure in text-to-video diffusion models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5954–5983, Suzhou, China. Association for Computational Linguistics.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.
- Shaowei Zhang and Deyi Xiong. 2025. [Debate4MATH: Multi-agent debate for fine-grained reasoning in math](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824, Vienna, Austria. Association for Computational Linguistics.
- Jingyuan Zhao, Wenyi Zhao, Bo Deng, Zhenghong Wang, Feng Zhang, Wenxiang Zheng, Wanke Cao, Jinrui Nan, Yubo Lian, and Andrew F Burke. 2024. Autonomous driving system: A comprehensive survey. *Expert Systems with Applications*, 242:122836.
- Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2025. [Beyond multimodal hallucinations: Enhancing llms through hallucination-aware direct preference optimization](#). In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.
- Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. [Analyzing and mitigating object hallucination in large vision-language models](#). In *International Conference on Learning Representations*, volume 2024, pages 56969–56998.
- Deyao Zhu, jun chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. [Minigt-4: Enhancing vision-language understanding with advanced large language models](#). In *International Conference on Learning Representations*, volume 2024, pages 18378–18394.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Sink Token Identification

As highlighted by recent literature (Sun et al., 2024a), certain channels within the token representations of LLMs tend to produce exceptionally massive activation values. These specific tokens act as attention black holes—monopolizing a substantial portion of the attention budget while contributing minimally to the actual semantic context. To quantify this artifact, we adopt the scoring mechanism introduced by VAR (Kang et al., 2025). For any given token representation $\mathbf{x} \in \mathbb{R}^D$, its sink propensity is evaluated via a scalar score $\phi(\mathbf{x})$. This metric isolates the maximum normalized magnitude across a pre-defined set of model-specific outlier dimensions, denoted as $\mathcal{D}_{sink}$:

$$\phi(\mathbf{x}) = \max_{d \in \mathcal{D}_{sink}} \left| \frac{\mathbf{x}[d]}{\sqrt{\frac{1}{D} \sum_{k=1}^D \mathbf{x}[k]^2}} \right|, \quad (11)$$

where $\mathbf{x}[d]$ represents the activation value at the d -th channel. Utilizing this scoring function, we can dynamically filter out sink tokens at any specific layer l by applying a predetermined threshold δ . Formally, the index set of identified sink tokens $\tilde{\mathcal{I}}^l$ is formulated as:

$$\tilde{\mathcal{I}}^l = \{j \in \mathcal{I} \mid \phi(\mathbf{x}_j^{l-1}) \geq \delta\}. \quad (12)$$

Here, $\mathbf{x}_j^{l-1}$ signifies the hidden state of the j -th token fed into layer l . Through this formulation, we can effectively identify sink tokens and subsequently reclaim their monopolized attention budget. Notably, consistent with the empirical configuration in VAR (Kang et al., 2025), the detection threshold δ is set to 20.

B Implementation Details

We set the key-cues filtering threshold to $\beta = 0.7$, while the attention reclamation ratio is fixed at $\gamma = 0.5$ following VAR (Kang et al., 2025). Unless otherwise specified, EVIAMAD operates under a default tri-agent configuration ($\mathcal{A}_G, \mathcal{A}_L, \mathcal{A}_C$) spanning two debate rounds. All experiments are conducted on NVIDIA H800 GPUs, with reported results averaged over three independent runs to ensure statistical reliability.

C More Experimental Results

Model Generalization Analysis. To further validate the architectural generalizability of EVIAMAD, we conduct experiments on distinct

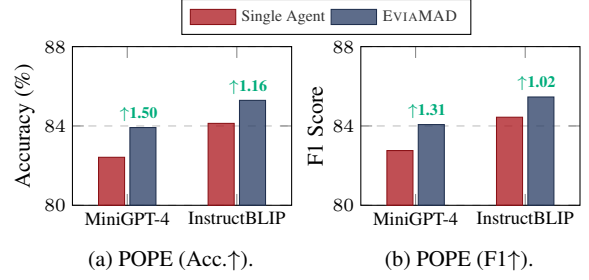


Figure 8: Performance evaluation on different MLLMs.

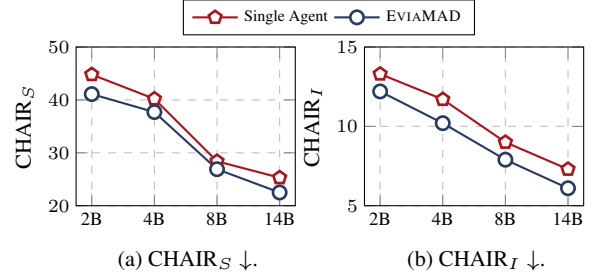


Figure 9: More scaling results on CHAIR.

MLLMs, specifically MiniGPT-4 (Zhu et al., 2024) and InstructBLIP (Dai et al., 2023). As illustrated in Figure 8, our framework consistently outperforms the standard single-agent baseline across both models. On the POPE benchmark, EVIAMAD achieves absolute accuracy improvements of 1.50% (from 82.42% to 83.92%) on MiniGPT-4 and 1.16% (from 84.13% to 85.29%) on InstructBLIP. This robust enhancement is equally reflected in the F1 scores, securing gains of 1.31% and 1.02%, respectively. These consistent performance boosts across entirely different model families confirm that our evidence-anchored debate mechanism serves as a universal, plug-and-play framework for hallucination mitigation.

More Scaling Results. To further investigate the scalability of our method in mitigating hallucinations, we evaluate EVIAMAD on the CHAIR benchmark across four parameter scales (comprising the Qwen3-VL series at 2B, 4B, and 8B, alongside InternVL3 at 14B). As depicted in Figure 9, while the single-agent baseline exhibits a natural decrease in hallucination rates as model capacity grows, EVIAMAD consistently drives these metrics even lower at every scale. Specifically, our framework secures absolute reductions in the CHAIR_S score by 3.7%, 2.5%, 1.5%, and 2.8% for the 2B, 4B, 8B, and 14B models, respectively. Parallel improvements are observed in the CHAIR_I metric, with consistent drops ranging from 1.1%

Method	POPE		CHAIR	
	Acc.↑	F1↑	CHAIR _S ↓	CHAIR _T ↓
InternVL3-8B				
+ EVIAMAD	89.2	89.3	27.3	8.1
w/o Roles	86.4	87.0	33.5	10.1
w/o Analyzer	88.6	88.7	30.1	9.2
w/o Enhance.	88.1	88.4	28.6	8.8
Qwen3-VL-8B				
+ EVIAMAD	89.5	89.9	26.9	7.9
w/o Roles	86.5	86.4	30.5	10.8
w/o Analyzer	89.2	89.2	27.5	8.6
w/o Enhance.	89.0	88.8	28.0	8.7

Table 5: Additional ablation study of EVIAMAD.

to 1.5%. These scale-invariant reductions confirm that our evidence-anchored framework provides an orthogonal structural enhancement, effectively suppressing hallucinations beyond what mere parameter scaling can achieve.

More Ablation Results. Table 5 provides additional ablation results on stronger MLLMs, including InternVL3-8B and Qwen3-VL-8B. The consistent performance drops after removing role specialization, the Conflict Analyzer, or attention enhancement further confirm that each component contributes to the robustness of EVIAMAD. In particular, removing role specialization causes the largest degradation on both models, indicating that asymmetric perceptual stances are essential for exposing visual blind spots. Removing the Analyzer or the enhancement module also weakens hallucination mitigation, showing that reliable evidence extraction and targeted reflection are both necessary for factual convergence.

Reversing Visual Attention Decay. To confirm that our attention reallocation mechanism successfully defends against context-induced cognitive inertia, we track the proportion of attention mass allocated to visual tokens across all layers. Under standard MAD, the rapidly expanding textual history of peer exchanges relentlessly hijacks the finite attention budget. Consequently, its layer-wise visual attention ratio in Round 2 suffers a severe structural collapse compared to its healthy state in Round 1 (grey dashed lines in Figure 10). This decay explains precisely why unguided debate progressively decouples from the image and degenerates into ungrounded linguistic guessing. In stark contrast, EVIAMAD actively reclaims the attention budget from uninformative sink tokens and strategically reinvests it into the bimodal evidence anchor. As illustrated in Figure 10, this realloca-

tion not only completely reverses the decay back to the initial Round 1 levels but actually surpasses them, establishing a robust upper envelope of visual focus. Particularly within the deep reasoning layers (shaded Peak Zones), EVIAMAD significantly amplifies the visual attention mass, elevating the peak ratio from 21.9% to 25.5% on LLaVA-1.5-7B, and from 25.9% to nearly 30% (29.8%) on InternVL3-8B. This sustained and enhanced visual focus ensures that the final consensus remains strictly anchored in physical evidence rather than drifting into text-induced hallucinations.

Detailed Results on MME. To provide a granular understanding of the performance gains, we present a comprehensive breakdown of the MME perception and cognition tasks in Tables 6 and 7, respectively. As demonstrated, EVIAMAD consistently improves model performance across every single subtask without any negative transfer. On the perception front (Table 6), explicitly anchoring debates in bimodal evidence strictly enhances fine-grained visual perception. For instance, on LLaVA-1.5-7B, highly sensitive metrics like Existence and Color jump by 3.9 and 3.5 points. Crucially, this robust grounding scales flawlessly to advanced models, securing steady gains for Qwen3-VL-8B across challenging dimensions like OCR and Posters. Furthermore, Table 7 validates that these benefits organically translate into superior higher-order reasoning. Across all four cognitive dimensions—including demanding tasks like Code Reasoning and Numerical Calculation—EVIAMAD universally outscores the single-agent baselines. This comprehensive sweep confirms that our framework fundamentally elevates both low-level visual perception and high-level cognitive deduction.

D Prompt Templates for POPE

In this section, we detail the specific system prompts utilized for evaluating our framework on the POPE benchmark, as illustrated in Figures 11 and 12. To ensure the multi-agent system operates autonomously and robustly, we meticulously engineer these templates to enforce strict role-playing and standardized output formats. Specifically, the prompts for the Global ($\mathcal{A}_G$) and Local ($\mathcal{A}_L$) Observers are designed to establish perceptual asymmetry, preventing homogenized observations. Concurrently, the instruction for the Conflict Analyzer ($\mathcal{A}_C$) strictly constrains its output to a parsable

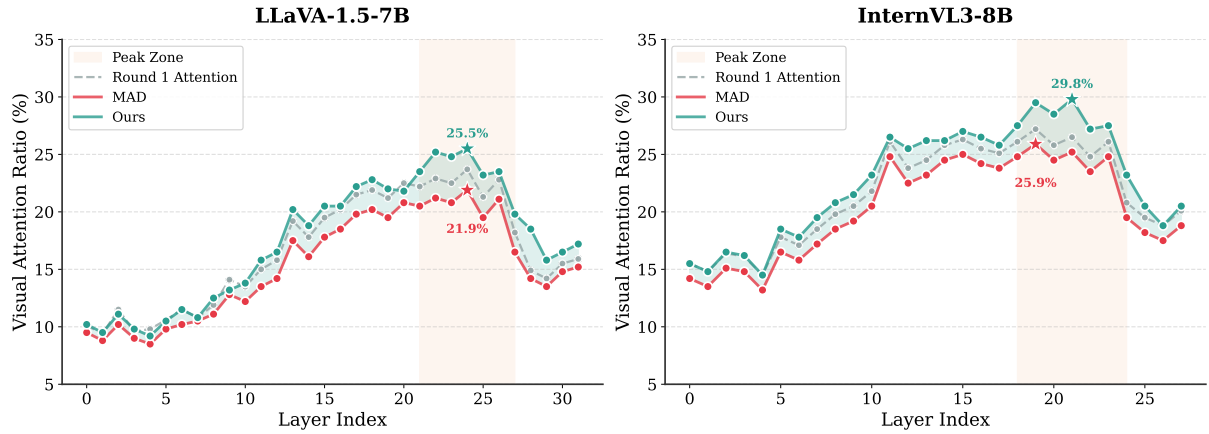


Figure 10: **Dynamics of layer-wise visual attention.** Compared to the initial healthy state in Round 1 (grey dashed line), standard MAD (red) suffers severe context-induced visual attention decay during Round 2. In stark contrast, EVIAMAD (green) completely reverses this structural collapse, actively reclaiming the attention budget to establish a robust upper envelope, particularly within the deep reasoning layers (shaded Peak Zones).

Base Agent	Perception										Total↑
	Exist.↑	Count↑	Position↑	Color↑	OCR↑	Posters↑	Artwork↑	Scene↑	Landmark↑	Celebrity↑	
LLaVA-1.5-7B	157.2	121.9	113.4	151.0	114.9	119.6	122.5	131.9	127.6	83.3	1243.3
+EVIAMAD	161.1	124.8	116.1	154.5	117.5	123.7	124.4	135.1	130.4	84.9	1272.5
LLaVA-NeXT	158.7	123.0	114.4	152.3	115.8	122.1	122.3	133.0	128.7	83.7	1254.0
+EVIAMAD	162.3	125.7	116.9	155.7	118.4	126.1	125.0	135.1	131.6	85.5	1282.3
Qwen2.5-VL-7B	160.5	124.7	115.1	154.7	117.8	123.0	123.8	134.5	129.6	84.7	1268.4
+EVIAMAD	163.4	126.4	117.6	155.6	119.9	124.6	124.5	136.8	131.1	85.9	1285.8
Gemma3-12B	173.7	132.2	122.9	162.1	121.7	128.9	128.2	140.7	136.6	87.5	1334.5
+EVIAMAD	175.2	133.4	124.7	165.9	122.4	130.7	125.7	141.9	137.8	88.8	1346.5
InternVL3-8B	175.1	135.7	126.2	168.0	127.8	134.9	135.0	146.7	142.0	92.3	1383.7
+EVIAMAD	177.4	137.4	127.8	170.2	129.5	136.6	136.1	148.6	143.8	93.5	1400.9
Qwen3-VL-8B	177.6	137.5	128.0	170.4	129.6	136.4	136.8	148.8	144.0	93.6	1402.7
+EVIAMAD	179.2	138.8	129.1	171.9	130.8	139.1	137.2	150.1	145.3	94.5	1416.0

Table 6: Detailed performance breakdown on the perception tasks of the MME benchmark. EVIAMAD yields consistent improvements across various fine-grained visual perception subtasks compared to the baselines. The best results are highlighted in **bold**.

Base Agent	Cognition				Total↑
	Code Reasoning↑	Commonsense Reasoning↑	Numerical Calculation↑	Text Translation↑	
LLaVA-1.5-7B	68.1	72.3	59.6	55.3	255.3
+EVIAMAD	69.7	74.0	60.9	56.6	261.2
LLaVA-NeXT	68.7	72.9	60.1	55.8	257.5
+EVIAMAD	70.2	74.6	61.2	57.0	263.0
Qwen2.5-VL-7B	69.9	73.2	60.8	56.3	260.2
+EVIAMAD	70.6	74.6	61.9	57.3	264.4
Gemma3-12B	74.7	80.1	64.4	59.8	279.0
+EVIAMAD	75.9	81.5	65.8	60.6	283.8
InternVL3-8B	75.7	80.5	66.3	61.5	284.0
+EVIAMAD	77.3	81.5	67.1	62.3	288.2
Qwen3-VL-8B	76.8	81.6	67.2	62.4	288.0
+EVIAMAD	77.5	82.4	67.5	63.0	290.4

Table 7: Detailed performance breakdown on the cognition subset of the MME benchmark. Compared to the baselines, EVIAMAD yields consistent improvements across various high-level visual reasoning and cognitive subtasks. The best results are highlighted in **bold**.

JSON schema, enabling the automated extraction of critical visual coordinates for subsequent attention reallocation.

Agent1 Prompt	Agent2 Prompt
You are Agent 1, a Global Observer. Analyze the entire image comprehensively and answer the following question: [Question]. To ensure a thorough and systematic response, first internally observe the overall scene, break down spatial relationships, and weigh the visual evidence. Then, strictly distill your findings and output ONLY the following format: Format: Answer: [yes/no] Confidence: [0.0-1.0] Reasoning: [1-2 sentences highlighting the critical visual evidence from your global observation]	You are Agent 2, a Local Observer. Focus your attention meticulously on the specific local regions, cropped details, or fine-grained elements of the image to answer the following question: [Question]. To ensure a precise and detail-oriented response, first internally scrutinize the localized visual features, object attributes, and subtle cues that might be overlooked from a global perspective. Then, strictly distill your findings and output ONLY the following format: Format: Answer: [yes/no] Confidence: [0.0-1.0] Reasoning: [1-2 sentences highlighting the critical fine-grained visual evidence]

Figure 11: Prompt templates for the Global Observer ($\mathcal{A}_G$) and Local Observer ($\mathcal{A}_L$) on the POPE benchmark. These instructions are explicitly designed to enforce perceptual asymmetry, compelling the agents to respectively extract macro-level context and fine-grained visual details.

Agent3 Prompt
You are Agent 3, a Conflict Analyzer. Your primary task is to compare the responses from Agent 1 (Global Observer) and Agent 2 (Local Observer) for the following question: [Question]. Agent 1 Response: [Agent 1 Output] Agent 2 Response: [Agent 2 Output] Internally analyze the provided responses to identify any logical contradictions or significant discrepancies in their answers or confidence levels. Perform the following reasoning steps: 1. Determine if the conflict is a "direct contradiction" (answers differ) or "confidence discrepancy" (answers match but confidence varies greatly). 2. Extract visually grounded tokens (objects, attributes, spatial terms). 3. Identify the bounding box [x1, y1, x2, y2] that encloses the specific visual region most critical to resolving the conflict. 4. Formulate a short attention guidance instruction focusing on that region and the key tokens. Output ONLY a valid JSON object with NO extra text, strictly adhering to the following schema: <pre>{ "Has_Conflict": true/false, "Conflict_Description": "Brief description of the disagreement.", "Bounding_Box": [x1, y1, x2, y2], "Attention_Guidance": "Focus on [Conflict_Description], [Bounding_Box], and pay attention to visually grounded tokens." }</pre>

Figure 12: Prompt template for the Conflict Analyzer ($\mathcal{A}_C$) on the POPE benchmark. The prompt strictly instructs the arbiter to identify logical discrepancies and output a structured JSON format containing critical spatial bounding boxes and attention guidance to direct subsequent attention reallocation.

E More Discussion

Contribution Analysis. Although EVIAMAD builds upon several foundational concepts, it fundamentally redesigns each component and orchestrates them into a self-consistent framework for evidence-anchored debate. First, in terms of role-based prompting, EVIAMAD does not merely encourage linguistic diversity among agents. Instead, it explicitly bifurcates the observers into macroscopic and detail-oriented roles, thereby breaking shared visual blind spots through complementary perceptual views. Second, for disagreement analysis, the Conflict Analyzer is not a passive summarizer of textual debates. It serves as an active factual arbiter that converts subjective linguistic disputes into deterministic bimodal evidence anchors. Third, the cropping operation compels models to re-examine the disputed visual regions directly, effectively dismantling the martingale trap caused

by ungrounded peer exchange. Finally, unlike existing attention steering methods that perform generalized redistribution, EVIAMAD reallocates attention specifically toward the verified bimodal evidence anchor. These coordinated designs transform loosely connected techniques into a unified mechanism that grounds multi-agent deliberation in physically verifiable evidence.

Analysis of Conflict Analyzer. We further evaluate the localization quality of the Conflict Analyzer on POPE by comparing its predicted bounding boxes against the ground-truth annotations. As shown in Table 8, the evaluation covers 500 samples and yields a 0.652 mean Intersection over Union (mIoU), with a median IoU of 0.671 and a standard deviation of 0.148. The predicted boxes also achieve a 72.4% Hit Rate under the $\text{Acc@IoU} \geq 0.5$ criterion, while maintaining 58.2% and 39.6% Hit Rates at stricter IoU thresholds of

Metric	Value
Number of samples	500
Mean IoU (mIoU)	0.652
Median IoU	0.671
Standard deviation	0.148
Hit Rate (IoU ≥ 0.5)	72.40%
Hit Rate (IoU ≥ 0.6)	58.20%
Hit Rate (IoU ≥ 0.7)	39.60%

Table 8: Localization quality of the Conflict Analyzer.

0.6 and 0.7, respectively. Beyond localization accuracy, the Analyzer is robust in practice: most detection errors are false-positive conflicts caused by strict linguistic comparisons, which merely trigger a redundant reflection round rather than a destructive false-correction loop. Moreover, our dual-input reflection design serves as a fail-safe mechanism. Since agents receive both the cropped patch and the full original image, they are never deprived of the complete visual context even when a predicted box is imperfect. These properties make the Analyzer a reliable source of spatially grounded evidence for subsequent cropping and attention reallocation.

Necessity of Multi-Agent Design. A hallucinating single agent often cannot self-diagnose its own visual blind spots, because the same incomplete perception that causes the error also constrains the agent’s ability to recognize it. EVIAMAD addresses this limitation by engineering asymmetric perceptual stances among agents. The Global Observer is encouraged to capture holistic scene-level context, while the Local Observer is directed toward fine-grained visual details. Their structured disagreement exposes latent inconsistencies that would otherwise remain hidden in a single-agent pipeline. The Conflict Analyzer then transforms these inconsistencies into verifiable evidence anchors, ensuring that the subsequent reflection stage is driven by cross-agent perceptual checks rather than by self-confirmation from a single model.

Overhead Analysis. We conduct a quantitative profiling of the computational costs introduced by EVIAMAD. On average, the token consumption per query is 2,243 for EVIAMAD, compared with 714 for the single-agent baseline. Table 9 further reports the per-sample latency and TFLOPs across representative MLLMs. As expected, the multi-agent debate process introduces additional computation. Nevertheless, EVIAMAD is designed to explore the feasibility of leveraging

Model	Method	Latency (ms)	TFLOPs
LLaVA-1.5-7B	Single-Agent	56.0	14.2
	EVIAMAD	168.3	42.3
InternVL3-8B	Single-Agent	68.2	16.5
	EVIAMAD	204.6	48.9
Qwen3-VL-8B	Single-Agent	65.0	16.2
	EVIAMAD	195.5	47.5

Table 9: Latency and FLOPs comparison per sample.

evidence-anchored multi-agent debate to mitigate MLLM hallucinations. Improving computational efficiency, such as through adaptive debate triggering, lightweight analyzers, and early stopping, remains an important direction for future work.

F Literature Survey

Existing MLLM hallucination mitigation methods are largely built around a single-agent paradigm. Training-based approaches align the model with hallucination-aware preference data or supervised signals (Sun et al., 2024b; Zhao et al., 2025), but they typically require additional optimization and carefully curated annotations. Post-hoc correction methods verify and revise generated responses after inference (Zhou et al., 2024; Yin et al., 2024), improving factuality at the cost of extra verification pipelines and latency. Inference-time interventions provide a training-free alternative by manipulating decoding behaviors or attention patterns (Huang et al., 2024; Leng et al., 2024), yet they still operate within the perceptual boundary of a single model. Despite their effectiveness, these methods largely overlook the wisdom of crowds: a single agent may fail to expose its own blind spots, whereas structured multi-agent interaction can surface disagreements and transform them into verifiable evidence for grounded reflection.

G The Use of Large Language Models

We declare that Large Language Models (LLMs) were exclusively utilized for linguistic refinement and text polishing during the preparation of this manuscript. The core ideas, methodology, experimental designs, and scientific conclusions are entirely the original work of the human authors. The authors have carefully verified all AI-assisted language edits and assume full responsibility for the originality, accuracy, and integrity of the complete contents presented herein.