SEEING CLEARLY, REMEMBERING SELECTIVELY: MITIGATING MULTI-AGENT HALLUCINATION SNOWBALLING VIA EVIDENTIAL MODULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As Multimodal Large Language Models (MLLMs) continue to advance, they are increasingly integrated into multi-agent systems for collaborative visual reasoning and task completion. However, MLLMs may still generate responses inconsistent with visual or contextual evidence, and such hallucinations can propagate across agents, causing hallucination snowballing and undermining system-level robustness. Existing mitigation methods primarily strengthen individual agents' visual grounding, leaving the reliability of information exchanged between agents insufficiently addressed. In this work, we investigate this problem from an evidential perspective by manipulating agentic visual and memory evidence to trace the influence on latent representations. Our analysis indicates that the generation of hallucination within an agent is linked to the disregard of visual evidence, whereas the propagation of hallucination between agents is associated with reliance on memory evidence supported by cross-layer inconsistencies. Based on these observations, we propose Dual Evidential Adaptive Modulation (DELVE), a training-free framework with two complementary modules. Specifically, Pristine Visual Evidence Reinforcement (PVER) derives dimension-level evidential weights to selectively reinforce latent representations with stable and pristine visual evidence of the initial agent, providing a stronger visual reference throughout agent interactions to mitigate intra-agent hallucination. Conflict-Aware Unreliable Memory Suppression (CUMS) exploits cross-layer directional inconsistency as a proxy for potential memory conflict, selectively attenuating inconsistent memory contributions while retaining coherent upstream information, thereby mitigating inter-agent hallucination propagation. Experiments on six hallucination and visual question answering benchmarks, spanning four MLLMs and three communication topologies, demonstrate improved hallucination mitigation while largely preserving general multimodal capabilities.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have made substantial progress in visual understanding (Feng et al., 2026), reasoning (Zheng et al., 2026), and generation (Fang et al., 2025). These capabilities have motivated their use in multi-agent systems (MAS), where agents exchange information and refine one another's responses to solve complex visual tasks (Qian et al., 2025). Despite this progress, MLLMs remain susceptible to hallucinations, where generated statements may contradict the image or lack support from the available evidence (Zhu et al., 2024; Chen et al.). In multi-agent systems, such errors are not confined to individual agents but may be repeated and amplified over successive inter-agent interactions, resulting in *Multi-agent Hallucination Snowballing (MHS)*. This propagation can substantially undermine system-level reliability and lead to severe consequences, particularly in safety-critical applications (Ma et al., 2026).

Addressing MHS generally requires tackling two coupled challenges: intra-agent hallucination generation and inter-agent hallucination propagation. Extensive prior work has focused primarily on the former. Specifically, inference-time approaches improve visual grounding through techniques such as contrastive decoding (Leng et al., 2024) and activation steering (Su et al., 2025), while training-based approaches fine-tune models to align model responses with visual evidence through

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.

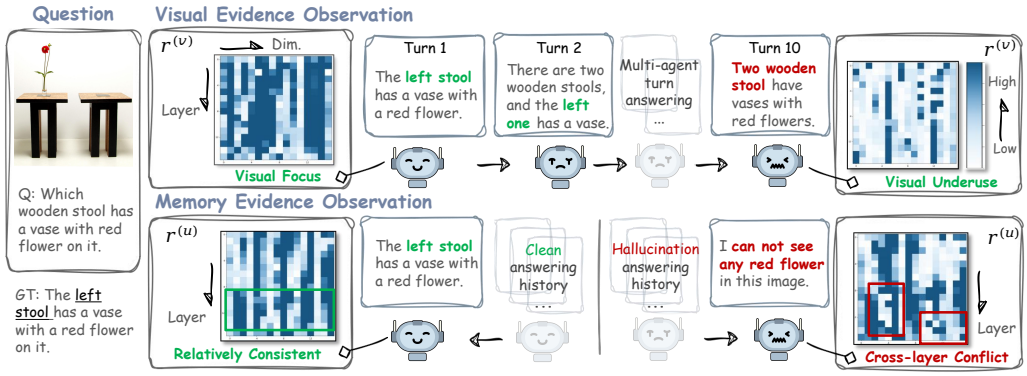


Figure 1: Overview of two key challenges in multimodal multi-agent systems: intra-agent hallucination generation and inter-agent hallucination propagation.

preference optimization (He et al.). These methods address an essential part of MHS, namely preventing an agent from introducing an error. However, these methods mainly focus on improving the robustness of individual agents, with limited ability to handle the propagation of erroneous or conflicting information across agents. Recent work on MHS primarily preserves visual information during interactions to counter excessive reliance on textual context, thus mitigating the hallucinations caused by ignoring visual information. However, stronger visual grounding alone does not explicitly distinguish useful upstream information from potentially misleading memory. Conversely, indiscriminately discounting upstream context can discard useful information that supports collaboration. Therefore, existing methods struggle to jointly exploit visual evidence and reliable upstream context, making it difficult to simultaneously suppress intra-agent hallucination generation and inter-agent hallucination propagation.

In this paper, we revisit these two challenges through an evidential perspective. When answering visual questions, agents rely fundamentally on two external sources of evidence: visual evidence and upstream historical evidence. To reveal the shared evidential causes underlying these challenges, we manipulate each evidence source and trace its relative evidential support (defined in Equation (4)) across the decoder layers during inference. As illustrated in Figure 1, visual support can weaken as interaction history accumulates, even though the image remains available. Consequently, agents become over-reliant on textual priors, leading to intra-agent hallucination generation despite the availability of relevant visual evidence. Meanwhile, upstream errors can contaminate historical evidence with conflicting information, resulting in inconsistent evidential support across layers and driving hallucination propagation to downstream agents. Ultimately, intra-agent hallucination generation and inter-agent hallucination propagation stem from an under-reliance on visual evidence and an over-reliance on unreliable memory evidence, respectively.

To tackle these challenges, we propose a training-free **D**ual **E**videntia**L** Adapti**VE** Modulation framework, termed **DELVE**. DELVE manipulates visual and memory sources to observe their evidential effects, and leverages these insights to dynamically modulate evidence in the latent space, thereby mitigating both intra- and inter-agent hallucinations in MAS. Our DELVE consists of two complementary modules: Pristine Visual Evidence Reinforcement module (PVER) and Conflict-Aware Unreliable Memory Suppression (CUMS). More specifically, PVER quantifies the relative support of visual evidence across dimensions of the latent space. By transforming these measurements into evidential weights, it adaptively reinforces latent representations with pristine visual evidence extracted from the initial agent. Benefiting from this, stable and uncontaminated visual evidence is assigned greater weight in reinforcing latent representations, preserving a robust visual reference for intra-agent hallucination mitigation. Moreover, CUMS employs cross-layer directional inconsistency in memory contributions as a proxy for potential upstream conflicts, enabling reliability assessment without requiring explicit verification of the factual correctness of upstream information. Based on this, CUMS adaptively attenuates potentially conflicting memory contributions while preserving consistent ones, mitigating inter-agent hallucination propagation. In summary, our contributions are as follows:

- We propose a Dual Evidential Adaptive Modulation framework (DELVE), which jointly modulates visual and upstream memory evidence to mitigate hallucination snowballing in MAS.

- We introduce a Pristine Visual Evidence Reinforcement module (PVER) that constructs dimension-level evidential weights to selectively reinforce latent representations with pristine visual evidence, mitigating intra-agent hallucination generation.
- We develop a Conflict-Aware Unreliable Memory Suppression module (CUMS) that exploits cross-layer directional inconsistency to suppress conflicting memory contributions, mitigating inter-agent hallucination propagation.
- We evaluate DELVE on six VQA benchmarks across four MLLMs and three MAS topologies. Comparison experiments, ablations, and analyses of interaction scale support the complementary roles of visual reinforcement and memory suppression in mitigating MHS.

2 Related Work

Hallucination Mitigation in MLLMs. Hallucination mitigation in MLLMs aims to improve the faithfulness of generated responses to visual inputs. Existing hallucination mitigation methods can be broadly categorized into two paradigms: **1)** Training-based approaches mitigate hallucinations by further optimizing MLLMs with carefully designed supervision signals (Jiang et al., 2024; Chen et al., 2024; Xiao et al., 2025) to strengthen the alignment between generated responses and visual evidence. For example, several methods (Zhang et al., 2024b) mitigate hallucinations through supervised fine-tuning on carefully curated data (Li et al., 2026), preference optimization (Yu et al., 2024; He et al.; Yang et al., 2025) over grounded and hallucinated responses, or reinforcement learning (Guo et al., 2026) with grounding-aware reward signals. However, these methods typically require additional data curation and model optimization, resulting in higher training costs and limited flexibility in multi-agent task adaptation. **2)** Training-free approaches (Leng et al., 2024; Huang et al., 2024) instead intervene during inference by perturbing inputs, pruning visual tokens (Zhuang et al., 2025), designing contrastive decoding signals (Xiao et al., 2026), or steering internal activations (Su et al., 2025), avoiding parameter updates and additional training. Most existing methods operate in the logit space, limiting direct control over hallucination-related internal representations.

Hallucination in Long-Horizon LLM Reasoning. Recent studies (Zhang et al., 2024a; Lu et al., 2026; Jiang et al., 2026) have revealed that hallucinations are not merely isolated errors but can be progressively inherited and amplified over multi-turn interactions and long-chain reasoning (Cheng et al., 2025), giving rise to a Hallucination Snowballing effect (HS). Existing methods (Lightman et al., 2024; Zou et al., 2025; Zhang et al., 2026b) to mitigate HS have primarily focused on improving the reliability of contextual information, including context refinement (Xu et al., 2024), intermediate verification (Dhuliawala et al., 2024), self-correction (Madaan et al., 2023), and retrieval-augmented generation (Shuster et al., 2021). More recently, emerging work (Yu et al., 2026) has begun to investigate hallucination in multi-agent systems (Qian et al., 2025), where richer interaction structures and inter-agent communication introduce additional pathways for erroneous information to accumulate and propagate. This approach alleviates this issue by preserving visual information throughout agent interactions, thereby reducing hallucination caused by excessive reliance on text. However, existing methods mainly correct individual agents' erroneous outputs, overlooking the learning and propagation of unreliable evidence.

3 Method

3.1 Latent-Space Evidence Modeling

For a given image-question pair, the prediction of each agent in an MLLM-based Multi-Agent System (MAS) can be modeled as $p(y \mid q, I, U; \mathbf{\Theta})$, where q denotes the question, I denotes the input image, U denotes the memory information accumulated from upstream agents in MAS, and Θ represents the endogenous knowledge encoded in the model parameters. Collectively, I, U, and Θ constitute the generalized information sources involved in model reasoning. From this perspective, hallucinations can be attributed to failures in utilizing these information sources.

Motivated by Evidence Theory (Dempster, 2008), we first treat the controllable external information sources as evidence sources, $\mathcal{S} = \{\mathcal{S}^v, \mathcal{S}^u\}$, where $\mathcal{S}^v \sim I$ and $\mathcal{S}^u \sim U$ denote the visual and memory evidence sources, respectively, with their corresponding information instantiated by I and U. For either evidence source $\mathcal{S}^s \in \mathcal{S}$, we construct a positive input context $\mathcal{C}^{(s,+)}$ and a negative

context $\mathcal{C}^{(s,-)}$ for the generated response by intervening on the corresponding evidence source while keeping the question, prompt, and other conditions fixed. This process can be formulated as:

$$\mathcal{C}^{(s,+)} = \mathcal{C}_0 \oplus \mathcal{S}^{(s)}, \qquad \mathcal{C}^{(s,-)} = \mathcal{C}_0, \tag{1}$$

where $\mathcal{C}_0$ denotes the common background context, and $\oplus$ denotes the incorporation of a controlled evidence source into $\mathcal{C}_0$. Different evidence sources induce distinct observations. Accordingly, following (Li et al., 2025), we perform the observation within the latent representation space of each MLLM decoder layer, *i.e.*,

$$\boldsymbol{H}_l = [\boldsymbol{h}_{l,1}, \boldsymbol{h}_{l,2}, \ldots, \boldsymbol{h}_{l,T}]^\top \in \mathbb{R}^{T\times d}, \quad \text{where} \quad \boldsymbol{h}_{l,t} = F_l\left(\boldsymbol{h}_{l-1,1}, \boldsymbol{h}_{l-1,2}, \ldots, \boldsymbol{h}_{l-1,t}\right). \tag{2}$$

$\boldsymbol{H}_l$ denotes the latent representation at the l-th layer, $\boldsymbol{h}_{l,t}$ represents the latent representation of the t-th token at the l-th layer, F_l is the computation performed by the decoder layer, T is the total number of tokens, and d is the dimensionality of the latent space. Following (Fu et al., 2025; Lin et al., 2026), the last several tokens aggregate preceding visual and historical contexts through causal attention, serving as decision anchors for comparing the two conditions. Based on this, we define the following two evidential metrics in the observations:

Definition 3.1 (Latent evidential observation)**.** Under $\mathcal{C}^{(s,\pm)}$, let $T^{(s,\pm)}$ be the prompt length and K_t the decision anchor size ($K_t \le T^{(s,\pm)}$). The observation $\boldsymbol{z}_l^{(s,\pm)}$ and latent evidence $\boldsymbol{\delta}_l^{(s)}$ can be formulated as:

$$\boldsymbol{z}_l^{(s,\pm)} = \frac{1}{K_t} \sum_{t=T^{(s,\pm)}-K_t+1}^{T^{(s,\pm)}} \boldsymbol{h}_{l,t}^{(s,\pm)}, \qquad \boldsymbol{\delta}_l^{(s)} = \boldsymbol{z}_l^{(s,+)} - \boldsymbol{z}_l^{(s,-)}. \tag{3}$$

The latent evidence $\boldsymbol{\delta}_l^{(s)}$ provides a vanilla measure of the contribution for the effect induced by the evidence source. To account for differences in activation scale across dimensions, we further introduce relative support as follows:

Definition 3.2 (Relative evidential support)**.** For latent dimension j, with $z_{l,j}^{(s,\pm)}$ and $\delta_{l,j}^{(s)}$ denoting the corresponding vector entries, the relative support $r_{l,j}^{(s)}$ can be formulated as:

$$r_{l,j}^{(s)} = \frac{|\delta_{l,j}^{(s)}|}{|z_{l,j}^{(s,+)}| + |z_{l,j}^{(s,-)}| + \varepsilon}, \qquad j = 1, \ldots, d, \tag{4}$$

where ε is a small numerical stabilizer.

The triangle inequality gives $r_{l,j}^{(s)} \in [0, 1)$, and we stack these values into the dimension-wise support scores: $\boldsymbol{r}_l^{(s)} = [r_{l,1}^{(s)}, \ldots, r_{l,d}^{(s)}]$, which characterize how strongly the evidence source influences different latent dimensions at the l-th layer. A larger value indicates stronger evidence-induced influence on the corresponding dimension, whereas a value close to zero indicates little evidence contribution. Guided by the above metrics, we propose a Dual Evidential Adaptive Modulation framework (DELVE), as illustrated in Figure 2, which selectively reinforces pristine visual evidence and attenuates potentially unreliable memory evidence through PVER (Section 3.2) and CUMS (Section 3.3), respectively.

3.2 Pristine Visual Evidence Reinforcement

To mitigate intra-agent hallucinations associated with underutilization of visual information, we propose a Pristine Visual Evidence Reinforcement module (PVER) that selectively reinforces latent components responsive to the visual input. Following Equations (3) and (4), we first compare the paired contexts (q, I, U) and $(q, \varnothing, U)$ to observe the latent visual evidence $\boldsymbol{\delta}_l^{(v)}$ and its relative support $\boldsymbol{r}_l^{(v)}$. These quantities characterize the changes induced by the image under the current question and upstream memory.

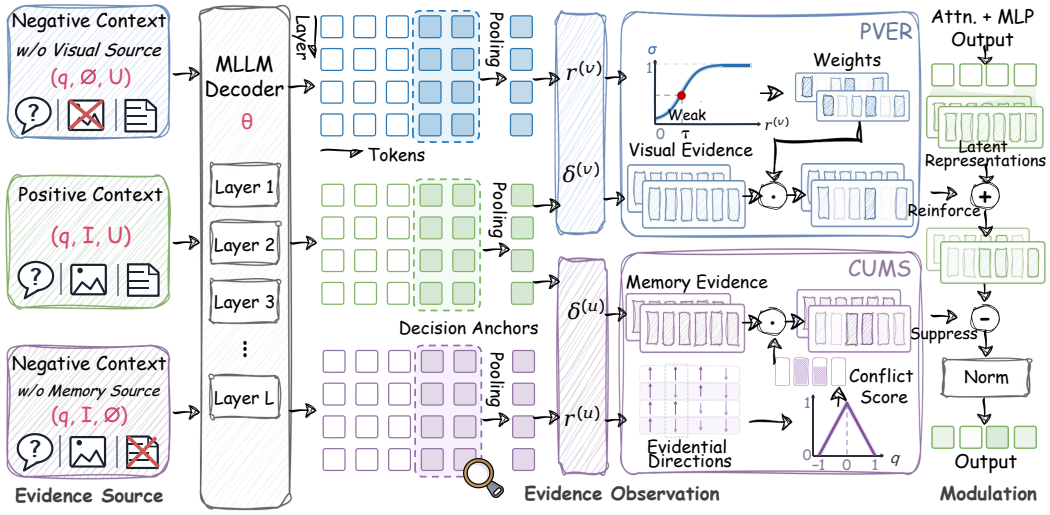


Figure 2: Overview of our proposed DELVE. The full context (*i.e.*, (q, I, U)) and the negative visual and memory contexts (*i.e.*, $(q, \varnothing, U)$ and $(q, I, \varnothing)$) are separately fed into the MLLM to obtain latent evidence $\delta^{(\cdot)}$ and relative evidential support $r^{(\cdot)}$ observed within the decision anchor. PVER then estimates dimension-level visual weights, and CUMS computes cross-layer directional conflict scores. These signals are finally used to refine the output representation at each layer by reinforcing pristine visual cues and suppressing conflicting memory information.

The relative support $r_{l,j}^{(v)}$ quantifies the influence of visual evidence on dimension j relative to its activation magnitudes. We convert $r_{l,j}^{(v)}$ into a soft weight through a sigmoid transformation:

$$g_{l,j} = \sigma\left(\frac{r_{l,j}^{(v)} - \tau}{T_e}\right), \qquad j = 1, \ldots, d, \tag{5}$$

where σ is the *Sigmoid* function, τ is the support threshold, and $T_e > 0$ controls the smoothness of selection. This weighting encourages the reinforcement to rely more on dimensions that exhibit strong visual support, thereby preserving a more stable visual reference and reducing the influence of noisy visual components.

Collecting these weights into $\boldsymbol{g}_l = [g_{l,1}, \ldots, g_{l,d}]$, we then reweight the visual evidence $\boldsymbol{\delta}_l^{(v)}$ to obtain the selected visual evidence:

$$\tilde{\boldsymbol{v}}_l = \boldsymbol{g}_l \odot \boldsymbol{\delta}_l^{(v)}, \tag{6}$$

where $\odot$ denotes the Hadamard product. This weighting adjusts the relative contributions of the visual evidence components, yielding selectively weighted visual evidence for adaptive reinforcement of latent representations.

We then adopt the selected visual evidential direction to reinforce the output representation. Given $\boldsymbol{x}_{l,t}$ at layer l and token t, we apply the following norm-preserving update:

$$\boldsymbol{x}_{l,t}^{(v)} = \|\boldsymbol{x}_{l,t}\|_2 \, \mathcal{N}\left(\mathcal{N}(\boldsymbol{x}_{l,t}) + \lambda(1 + \rho_{l,t})\mathcal{N}(\tilde{\boldsymbol{v}}_l)\right), \quad \rho_{l,t} = \max\{0, \cos\left(\boldsymbol{x}_{l,t}, -\mathcal{N}(\tilde{\boldsymbol{v}}_l)\right)\}, \tag{7}$$

where $\lambda \geq 0$ controls the reinforcement strength, $\|\cdot\|_2$ denotes the Euclidean norm, and $\mathcal{N}(\tilde{\boldsymbol{v}}_l) = \tilde{\boldsymbol{v}}_l / \|\tilde{\boldsymbol{v}}_l\|_2$ for nonzero $\tilde{\boldsymbol{v}}_l$. If either input vector is zero, we leave the output unchanged. The coefficient $\rho_{l,t}$ measures directional opposition through cosine similarity. This update strengthens the selected visual evidence's contribution to the representation, applying greater reinforcement when the representation opposes it.

However, as upstream memory accumulates, it can interfere with the visual evidence $\boldsymbol{\delta}_l^{(v)}$, even when the image itself remains unchanged. To maintain a stable visual reference, we use the initial agent's visual evidence, extracted from $(q, I, \varnothing)$ and $(q, \varnothing, \varnothing)$ before introducing any upstream history. For subsequent agents, we construct pristine visual evidence in MAS, written as:

$$\boldsymbol{v}_l = \boldsymbol{g}_l \odot \boldsymbol{\delta}_l^{(v_0)}, \tag{8}$$

where $\boldsymbol{\delta}_l^{(v_0)}$ is the cached initial visual evidence at layer l. In Equation (7), we replace $\tilde{\boldsymbol{v}}_l$ with cleaner $\boldsymbol{v}_l$ to combine a fixed visual reference with context-dependent dimension selection. Specifically, the initial input provides the visual evidence, whereas the current context determines which components should be emphasized. This design maintains a consistent and clean visual reference throughout agent interactions, mitigating intra-agent hallucinations.

3.3 Conflict-Aware Unreliable Memory Suppression

Although PVER mitigates the risk of intra-agent hallucinations through visual evidence enhancement, potentially contaminated upstream memory may still impact the performance of downstream agents, triggering a cascading MHS. To handle this issue, we present a Conflict-Aware Unreliable Memory Suppression module (CUMS), which adaptively discounts potentially unreliable upstream agentic memory to mitigate inter-agent hallucination propagation. Similarly, we compare the paired contexts (q, I, U) and $(q, I, \varnothing)$ to observe the latent memory evidence $\boldsymbol{\delta}_l^{(u)}$ and the relative support $\boldsymbol{r}_l^{(u)}$ from Equations (3) and (4).

To alleviate the impact of upstream hallucinations, directly verifying the factual correctness of upstream memory is often challenging and computationally expensive. We therefore shift our focus to potential conflicts between hallucinated upstream information and the current visual and textual context. Inspired by layer-wise analyses of contextual knowledge (Ju et al., 2024), CUMS assesses memory reliability through the directional consistency of its latent evidence, using inconsistency as a proxy for potential contextual conflict. Specifically, to capture conflicts more relevant to the final predictions, we consider the last K_l decoder layers, denoted by $\mathcal{L} = \{L - K_l + 1, \ldots, L\}$, and define cross-layer consistency $q_j^{(u)}$ and the corresponding conflict score $\kappa_j^{(u)}$ as:

$$q_j^{(u)} = \frac{\sum_{l \in \mathcal{L}} r_{l,j}^{(u)} \operatorname{sign}(\delta_{l,j}^{(u)})}{\sum_{l \in \mathcal{L}} r_{l,j}^{(u)} + \varepsilon}, \qquad \kappa_j^{(u)} = 1 - |q_j^{(u)}|, \tag{9}$$

where $\operatorname{sign}(\cdot)$ indicates the direction of the activation change induced by memory evidence in each latent dimension. Thus, $q_j^{(u)}$ reflects the cross-layer directional consistency of memory evidence support on the j-th latent dimension, and $\kappa_j^{(u)} \in [0, 1]$ quantifies conflict across evidence. When sufficient memory support is present, a larger $\kappa_j^{(u)}$ indicates stronger cross-layer directional disagreement. More detailed theoretical discussion is provided in our Appendix.

Stacking the dimensional conflict scores into the dimension-wise vector $\boldsymbol{\kappa}^{(u)} = [\kappa_1^{(u)}, \ldots, \kappa_d^{(u)}]$, we construct a suppression vector $\boldsymbol{s}_l^{(u)}$ and jointly modulate the output of each token in the terminal layers, yielding a final output latent representation $\tilde{\boldsymbol{x}}_{l,t}$:

$$\boldsymbol{s}_l^{(u)} = \boldsymbol{\kappa}^{(u)} \odot \boldsymbol{\delta}_l^{(u)}, \qquad \tilde{\boldsymbol{x}}_{l,t} = \frac{\|\boldsymbol{x}_{l,t}\|_2 \big(\boldsymbol{x}_{l,t}^{(v)} - \boldsymbol{s}_l^{(u)}\big)}{\|\boldsymbol{x}_{l,t}^{(v)} - \boldsymbol{s}_l^{(u)}\|_2 + \varepsilon}. \tag{10}$$

This formulation adaptively suppresses contributions induced by unreliable memory information according to the degree of conflict in the memory evidence. As a result, latent dimensions associated with stronger conflict receive larger attenuation, whereas dimensions supported by consistent memory evidence are largely preserved. In this way, the model reduces the impact of contradictory memory history arising from hallucinated upstream agents when it potentially conflicts with clean context and visual semantics, thereby alleviating the propagation of inter-agent hallucinations.

4 Experiment

In this paper, all experiments are conducted on NVIDIA H20 GPU(s). Inspired by prior work (Yu et al., 2026), we evaluate DELVE across three representative MAS topologies: *chained*, *mesh*, and *random*. Among them, the random topology provides greater structural diversity due to its stochastic connectivity and is therefore adopted as the primary setting for further analysis. Unless otherwise specified, the number of MAS interaction turns is fixed to 10. To evaluate the generalizability of DELVE across different MLLMs, we conduct experiments with LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-Next-7B, Qwen2.5-VL-7B (Bai et al., 2025b), and Qwen3-VL-8B (Bai et al., 2025a).

Table 1: Comparison across different MAS structures on POPE, MMHal-Bench, and Hall Bench in terms of performance metrics (*e.g.*, F1(↑), Acc(↑), *etc.*). Specifically, POPE additionally includes hallucination snowballing metrics (*i.e.*, HS (↓) and ET(↓)). Better results are highlighted in **bold**.

MAS Structure	Method	POPE-R			POPE-P			POPE-A			POPE			MMHal	Hall Bench		
		F1	HS	ET	F1	HS	ET	F1	HS	ET	F1	HS	ET	Score	aAcc	qAcc	fAcc
Chained	LLaVA-1.5-7B	86.4	9.93	1.41	81.0	14.17	2.06	77.1	17.70	2.69	81.5	13.93	2.05	2.35	44.5	7.7	12.1
	w/ Ours	**86.7**	**7.43**	**1.27**	**84.8**	**8.60**	**1.45**	**80.2**	**13.38**	**2.16**	**83.9**	**9.80**	**1.63**	**2.56**	**49.2**	**12.3**	**19.4**
	LLaVA-NeXT-7B	77.3	11.10	1.73	76.1	12.87	1.98	75.3	13.31	2.05	76.2	12.43	1.92	2.30	42.4	9.7	13.3
	w/ Ours	**81.1**	**9.19**	**1.57**	**80.9**	**9.08**	**1.60**	**77.5**	**11.12**	**1.73**	**79.8**	**9.80**	**1.63**	**2.49**	**45.6**	**11.2**	**15.3**
	Qwen2.5-VL-7B	85.1	6.89	1.21	84.9	**7.43**	**1.29**	83.2	8.04	1.40	84.4	7.45	1.30	3.06	63.8	31.2	34.1
	w/ Ours	**86.8**	**6.58**	**1.14**	**85.4**	7.47	**1.29**	**84.8**	**7.88**	**1.36**	**85.7**	**7.31**	**1.26**	**3.11**	**64.2**	**33.2**	**40.5**
	Qwen3-VL-8B	87.5	6.14	1.15	85.7	7.55	1.30	85.5	7.99	1.36	86.2	7.23	1.27	2.82	**61.9**	29.0	**37.0**
	w/ Ours	**88.1**	**6.01**	**1.03**	**87.1**	**6.92**	**1.22**	**86.4**	**7.45**	**1.30**	**87.2**	**6.79**	**1.18**	**3.07**	61.5	**30.1**	34.1
Mesh	LLaVA-1.5-7B	**86.3**	11.42	1.66	77.9	13.02	2.03	74.1	18.19	2.81	79.4	14.21	2.17	2.34	43.2	11.2	13.3
	w/ Ours	85.8	**7.49**	**1.31**	**84.2**	**8.50**	**1.48**	**81.0**	**12.29**	**2.03**	**83.7**	**9.43**	**1.61**	**2.60**	**48.4**	**13.2**	**19.4**
	LLaVA-NeXT-7B	81.1	9.57	1.57	**78.2**	**13.65**	2.11	78.4	13.97	2.19	79.2	12.40	1.96	2.28	43.4	9.7	14.5
	w/ Ours	**84.2**	**8.13**	**1.38**	**78.2**	13.83	**2.09**	**80.1**	**13.43**	**2.06**	**80.8**	**11.80**	**1.84**	**2.76**	**45.4**	**10.3**	**16.2**
	Qwen2.5-VL-7B	86.4	6.92	1.23	83.9	7.60	1.36	84.2	7.91	1.42	84.8	7.48	1.34	2.72	60.9	28.6	33
	w/ Ours	**87.8**	**6.06**	**1.07**	**85.6**	**7.39**	**1.31**	**86.3**	**6.98**	**1.23**	**86.6**	**6.81**	**1.20**	**2.86**	**63.7**	**31.7**	**38.7**
	Qwen3-VL-8B	**90.1**	**4.96**	**0.90**	87.5	**6.53**	1.17	86.2	7.24	1.29	87.9	**6.24**	**1.12**	2.67	60.8	**28.4**	32.7
	w/ Ours	89.8	5.20	0.93	**87.7**	6.54	**1.16**	**86.7**	**7.20**	**1.28**	**88.1**	6.31	**1.12**	**2.76**	**61.2**	27.0	**34.1**
Random	LLaVA-1.5-7B	86.1	8.88	1.35	78.2	12.81	2.13	77.1	14.94	2.45	80.5	12.21	1.98	2.38	40.4	8.4	11.0
	w/ Ours	**88.0**	**7.22**	**1.17**	**83.7**	**11.29**	**1.72**	**79.2**	**13.32**	**2.21**	**83.6**	**10.61**	**1.70**	**2.40**	**48.3**	**12.8**	**20.2**
	LLaVA-NeXT-7B	80.4	8.88	1.57	78.2	9.87	1.74	77.7	10.42	1.82	78.8	9.72	1.71	1.90	43.4	11.2	14.2
	w/ Ours	**81.1**	**8.76**	**1.52**	**80.1**	**9.30**	**1.61**	**79.9**	**9.53**	**1.69**	**80.4**	**9.20**	**1.61**	**2.38**	**45.7**	**11.7**	**17.3**
	Qwen2.5-VL-7B	86.4	6.93	1.32	83.9	7.67	1.36	84.1	7.89	1.40	84.8	7.50	1.36	**2.86**	62.9	31.6	37.3
	w/ Ours	**87.7**	**6.15**	**1.09**	**86.0**	**7.11**	**1.26**	**85.3**	**7.55**	**1.33**	**86.3**	**6.94**	**1.23**	2.83	**64.8**	**34.7**	**40.8**
	Qwen3-VL-8B	90.1	**5.12**	**0.89**	87.9	6.91	1.19	**86.6**	7.62	1.31	**88.2**	6.55	**1.13**	2.69	60.5	27.7	33
	w/ Ours	**90.2**	5.24	0.92	**88.1**	**6.76**	**1.17**	86.4	**7.38**	**1.29**	**88.2**	**6.46**	**1.13**	**2.75**	**61.0**	**28.6**	**35.3**

We conduct experiments on a wide range of multimodal hallucination benchmarks, including three POPE sets (*i.e.*, Random (POPE-R), Popular (POPE-P), and Adversarial (POPE-A)) (Li et al., 2023), MMHal-Bench (Sun et al., 2024), and HallusionBench (Hall Bench) (Guan et al., 2024), as well as several general Visual Question Answering (VQA) benchmarks, including 14 MME subsets (Fu et al., 2026), MMBench (Liu et al., 2024b), and MM-Vet (Yu et al., 2023). For a fair comparison, all evaluations requiring an external LLM judge are conducted using the GPT-5.2 API. In addition to the standard metrics reported by each benchmark, we further report the Hallucination Snowballing (HS) score (Yu et al., 2026) and the average number of Erroneous Turns (ET) to provide a more comprehensive understanding of hallucination severity in MAS. Due to space constraints, implementation details, MAS configurations, base MLLMs, benchmark descriptions, evaluation metrics, and supplementary experiments are deferred to our Appendix.

4.1 Comparative Experiment

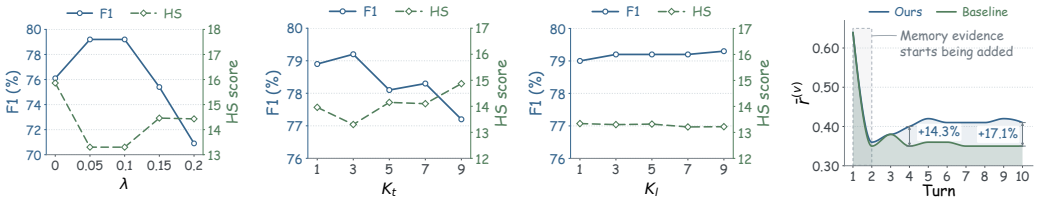
Table 2: Comparison on MME, MMBench, and MM-Vet.

Method	MME			MMBench			MM-Vet
	MME-P	MME-C	MME-Total	V-Acc	C-Acc	P-Acc	Score
LLaVA-1.5-7B	1182.7	260.0	1442.7	67.4	43.2	63.8	28.9
w/ Ours	**1197.6**	**274.3**	**1471.9**	**69.6**	**46.9**	**66.5**	**30.3**
LLaVA-NeXT-7B	1047.0	250.4	1297.4	71.3	58.4	70.6	38.3
w/ Ours	**1291.1**	**251.8**	**1543.0**	**71.3**	**60.7**	**71.0**	**40.1**
Qwen2.5-VL-7B	1629.3	646.1	2275.4	86.1	82.0	85.3	67.7
w/ Ours	**1673.7**	**664.6**	**2338.3**	**86.4**	**82.9**	**85.8**	**68.2**
Qwen3-VL-8B	**1669.7**	685.0	**2354.6**	89.0	84.7	87.7	**71.9**
w/ Ours	1636.2	**685.4**	2321.5	**89.9**	**85.7**	**88.2**	71.0

In this section, we first apply our DELVE to four base MLLMs under chained, mesh, and random MAS topologies on hallucination benchmarks, as shown in Table 1. In addition, we apply our DELVE to four base MLLMs on general VQA benchmarks and compare our DELVE with current state-of-the-art hallucination mitigation methods for single MLLMs (*i.e.*, LLaVA-RLHF (Sun et al., 2024), RLSF-V (He et al.), VISTA (Li et al., 2025), SHIELD (Huang et al., 2026), and Saliency (Zhang et al., 2026a)) on hallucination benchmarks under the representative random topology. The results are shown in Tables 2 and 3. From these experimental results, we can obtain the following observations: **1)** Different MLLMs exhibit varying degrees of performance degradation under different MAS settings. This highlights the detrimental effects of MHS and reveals its varying impact across different connection topologies. **2)** Our DELVE can effectively mitigate MHS, with particularly notable improvements on the more challenging POPE-A, MME, and

Table 3: Comparison with single-MLLM hallucination mitigation methods on POPE, MMHal-Bench, and Hall Bench under random topologies. † indicates the use of additional fine-tuning data.

Methods	POPE				MMHal-Bench	Hall Bench		
	Random	Popular	Adversarial	Avg. F1	Score	aAcc	qAcc	fAcc
LLaVA-RLHF† (2024)	87.1 -0.9	82.6 -1.1	78.7 -0.5	82.8 -0.8	2.38 -0.02	28.3 -20.0	6.6 -6.2	9.5 -10.7
RLSF-V† (2026)	86.7 -1.3	83.3 -0.4	82.1 +2.9	84.0 +0.4	2.27 -0.13	44.9 -3.4	10.8 -2.0	15.9 -4.3
Baseline	86.1 -1.9	78.2 -5.5	77.1 -2.1	80.5 -3.1	2.38 -0.02	40.4 -7.9	8.4 -4.4	11.0 -9.2
VISTA (2025)	86.1 -1.9	82.8 -0.9	77.5 -1.7	82.1 -1.5	2.44 +0.04	44.2 -4.1	9.7 -3.1	13.2 -7.0
SHIELD (2026)	85.9 -2.1	78.4 -5.3	77.7 -1.5	80.7 -2.9	2.36 -0.04	44.9 -3.4	14.3 +1.5	18.8 -1.4
Saliency (2026)	84.8 -3.2	77.3 -6.4	74.9 -4.3	79.0 -4.6	2.31 -0.09	42.9 -5.4	9.0 -3.8	12.7 -7.5
Ours	88.0	83.7	79.2	83.6	2.40	48.3	12.8	20.2



(a) Parameter λ (b) Parameter K_t (c) Parameter K_l (d) Comparison of $r^{(v)}$

Figure 3: (a)–(c) present the sensitivity analyses of λ, K_t, and K_l in our DELVE, evaluated in terms of F1 ($\uparrow$) and HS score ($\downarrow$), respectively. (d) compares the average visual evidence support $r^{(v)}$ between baseline and DELVE across multiple agentic turns on MMHal-Bench.

the open-ended generation benchmark MMHal-Bench. This improvement can be attributed to the complementary intra- and inter-agent hallucination mitigation mechanisms introduced by PVER and CUMS. **3)** Our method effectively mitigates hallucinations while preserving base models' original capabilities on general VQA benchmarks, with slight performance improvements observed on most benchmarks. **4)** Compared with hallucination mitigation methods for individual MLLMs, DELVE shows strong robustness across most benchmarks, particularly on more challenging ones (*e.g.*, Hall Bench). It also remains competitive with LLaVA-RLHF and RLSF-V despite their additional 10K preference samples and larger-scale SFT data, demonstrating the effectiveness of PVER for intra-agent hallucination mitigation and CUMS for suppressing inter-agent propagation.

4.2 Ablation Study and Parameter Analysis

In this section, we conduct an ablation study to assess the contribution of each component to our DELVE. Specifically, we ablate PVER and CUMS, corresponding to the intra-agent and inter-agent hallucination mitigation mechanisms, respectively. The results in Table 4 lead to the following observations. **1)** Removing PVER causes a substantial increase in hallucinations, indicating that strengthened visual evidence can effectively prevent hallucinations from occurring. **2)** Combining CUMS with PVER further improves performance across all three evaluations, suggesting that suppressing potentially unreliable upstream memory complements visual reinforcement in mitigating hallucination snowballing in multi-agent systems.

Table 4: Ablation study of our method on three benchmarks. ✓ stands for use.

PVER	CUMS	Random		Chained
		POPE-R	POPE-A	MMHal-Bench
		86.1	77.1	2.35
✓		87.7	77.8	2.50
	✓	87.0	76.1	2.44
✓	✓	**88.0**	**79.2**	**2.56**

Additionally, to evaluate the sensitivity of DELVE to different trade-off parameter settings, we plot the F1 and HS scores against varying values of λ, K_t, and K_l on POPE-A, using LLaVA-1.5 under the MAS topology, as shown in Figure 3 (a)-(c). The experimental results lead to the following observations: **1)** When λ is too small, *i.e.*, $\lambda < 0.05$, PVER provides insufficient visual enhancement to counteract hallucinations. Conversely, an excessively large $\lambda \geq 0.15$ may strongly alter the hidden representations, interfering with normal answer generation and reducing performance. **2)** Overly large K_t values degrade performance. Specifically, when $K_t > 7$, excessive pooling may dilute answer-relevant information. **3)** DELVE maintains relatively stable performance over a moderate range of parameter values, particularly for K_l. To strike a balance between performance and efficiency, we set $K_l = 3$ in our final configuration.

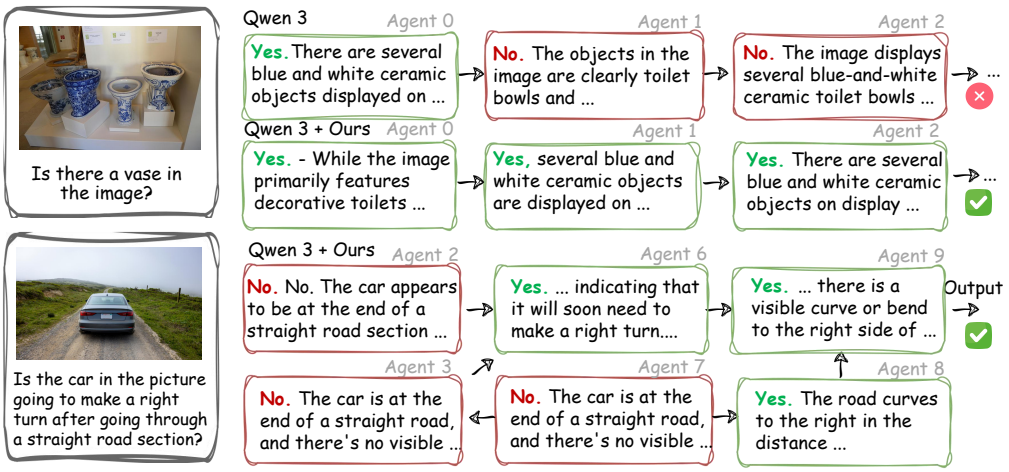
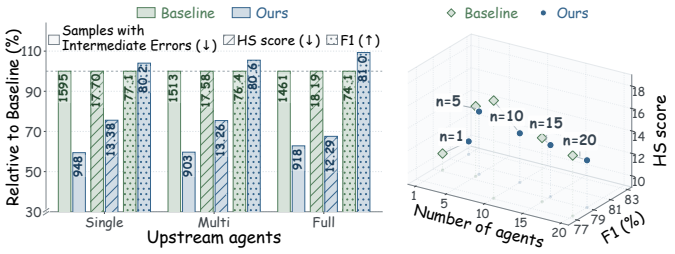


Figure 4: Case study of MHS mitigation with DELVE under chain and random topologies.

4.3 Visualization Analysis



(a) Different upstream scales (b) Different system scales

Figure 5: Robustness of DELVE under different MAS scales. (a) compares the baseline and DELVE under different upstream-agent scales (*Single*: 1, *Multi*: 5, *Full*: 9). (b) compares their performance across system scales (1, 5, 10, 15, and 20 agents).

To provide a comprehensive analysis of DELVE, we conduct a series of visualization experiments. First, we evaluate the effectiveness of PVER on MMHal-Bench by comparing the average visual evidence support $r^{(v)}$ of LLaVA-1.5 and DELVE across agentic turns 1-10, computed over 32 consecutive dimensions (Figure 3(d)), together with qualitative examples in Figure 4. We further evaluate MAS on POPE-A under varying communication densities and system sizes (Figure 5). Specifically, we cap the number of upstream agents per agent at 1 (*Single*), 5 (*Multi*), or 9 (*Full*), and report intermediate-error counts, HS, and F1. We also vary the number of agents from 1 to 20 under the chained topology and report the corresponding HS and F1 scores. The results reveal three key observations: **1)** As memory evidence accumulates across agentic turns, reliance on visual information progressively decreases, with a pronounced drop after two turns. PVER reinforces stable visual evidence, improving visual evidence support by over 17% relatively and reducing excessive reliance on language priors. **2)** Beyond mitigating intra-agent hallucinations, DELVE is robust to contaminated upstream information: downstream agents remain resilient even when hallucinated information is propagated from upstream agents. **3)** Across different system configurations, DELVE consistently achieves stronger performance than the baseline. These results demonstrate its effectiveness in mitigating MHS under varying communication densities and system scales.

5 Conclusion

In this paper, we propose a novel training-free Dual Evidential Adaptive Modulation framework, namely DELVE, to mitigate hallucination snowballing in multimodal multi-agent systems. DELVE integrates two complementary modules: Pristine Visual Evidence Reinforcement (PVER) and Conflict-Aware Unreliable Memory Suppression (CUMS). Specifically, PVER selectively reinforces latent representations with pristine visual evidence to alleviate intra-agent hallucination generation, and CUMS adaptively suppresses conflicting upstream memory contributions based on cross-layer directional inconsistency to mitigate inter-agent hallucination propagation. Extensive experiments on various VQA benchmarks demonstrate the effectiveness of DELVE across four base models and three multi-agent topologies.

6 USAGE OF LARGE LANGUAGE MODELS

We claim that ChatGPT was used as general-purpose assistive tools for grammar checking, and text polishing of drafts. Additionally, we used it to help identify content missing from the main text and to draft an initial structure for our appendix. It was used only to support this preliminary organization and all subsequent content was substantially revised, completed, and independently written by the authors. It were not used for retrieval and discovery (*e.g.*, finding related work), research ideation, methodological design, data collection, analysis, or writing substantive content that could qualify as a contribution. The authors take full responsibility for the contents of the paper, including any text influenced by these tools, and no LLM is listed as an author.

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*, 2025a.

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

Sarah Calderwood, Kevin McAreavey, Weiru Liu, and Jun Hong. Context-dependent combination of sensor information in dempster–shafer theory for bdi. *Knowledge and Information Systems*, 51 (1):259–285, 2017.

Kaiqi Chen, Yang Qin, Changhao He, Xi Peng, and Peng Hu. Doubt: Decoupled object-level understanding and bridging via vmf-based trustworthiness for hallucination detection in mllms. In *Forty-third International Conference on Machine Learning*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14239–14250. IEEE, 2024.

Zhangtao Cheng, Shichong Li, Yichen Xin, Bin Chen, Ting Zhong, and Fan Zhou. Precision through progression: Empowering temporal knowledge graph reasoning with knowledge-guided chain of thought. *Knowledge-Based Systems*, pp. 114448, 2025.

Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pp. 57–72. Springer, 2008.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Rongyao Fang, Chengqi Duan, Kun Wang, Hao Li, Linjiang Huang, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Hongsheng Li, et al. Puma: Empowering unified mllm with multi-granular visual generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15447–15457. IEEE, 2025.

Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, Xiaomin Song, and Peng Hu. Robust cross-modal alignment learning for cross-scene spatial reasoning and grounding. *Advances in Neural Information Processing Systems*, 38:49055–49075, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, and Qing Gu. Token prepending: A training-free approach for eliciting better sentence embeddings from llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3168–3181, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Peizheng Guo, Jingyao Wang, Wenwen Qiang, Jiahuan Zhou, Changwen Zheng, and Gang Hua. Copo: Causal-oriented policy optimization for hallucinations of mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11051–11063, 2026.

Changhao He, Shuhao Yan, Shuxian Li, Xi Peng, and Peng Hu. Rlsf-v: Mitigating hallucinations in mllms via fuzzy semantic self-feedback. In *Forty-third International Conference on Machine Learning*.

Yibo Hu, Yuzhe Ou, Xujiang Zhao, Jin-Hee Cho, and Feng Chen. Multidimensional uncertainty-aware evidential neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7815–7822, 2021.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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)*, pp. 13418–13427. IEEE, 2024.

Yiyang Huang, Liang Shi, Yitian Zhang, Yi Xu, and Yun Fu. Shield: Suppressing hallucinations in lvlm encoders via bias and vulnerability defense. In *International Conference on Learning Representations*, volume 2026, pp. 125123–125143, 2026.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27026–27036. IEEE, 2024.

Yue Jiang, Xue Jiang, Lihua Zhang, Zhiqiang Wang, Yuhang Lu, Peng Wang, Bo Han, Feng Zheng, and Dingkang Yang. Mm-snowball: Evaluating and mitigating hallucination snowballing in multimodal multi-turn dialogue. *arXiv preprint arXiv:2606.00622*, 2026.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8235–8246, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Wenyun Li, Zheng Zhang, Dongmei Jiang, and Xiangyuan Lan. Bolster hallucination detection via prompt-guided data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37618–37626, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Ailiang Lin, Zhuoyun Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. Causal2vec: Improving decoder-only llms as embedding models through a contextual token. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22769–22788, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Haolang Lu, Minghui Pan, Guoshun Nan, Jialin Zhuang, Zijie Zhao, ZhongXiang Sun, Kun Wang, Yang Liu, et al. Streaming hallucination detection in long chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21157–21183, 2026.

Ting Ma, Xiufeng Huang, Benlei Cui, Xiaowen Xu, Shikai Qiu, Ruijie Jian, Hongxing Li, Guanghui Wang, Longtao Huang, Haiwen Hong, et al. Yuvion llm: An adversarially-aware large language model for content and ai safety. *arXiv preprint arXiv:2606.27632*, 2026.

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

Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, et al. Scaling large language model-based multi-agent collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 41488–41505, 2025.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803, 2021.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Qing Li, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974, 2025.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pp. 13088–13110, 2024.

Yisong Xiao, Xianglong Liu, QianJia Cheng, Zhenfei Yin, Siyuan Liang, Jiapeng Li, Jing Shao, Aishan Liu, and Dacheng Tao. Genderbias-vl: Benchmarking gender bias in vision language models via counterfactual probing: Y. xiao et al. *International Journal of Computer Vision*, 133 (12):8332–8355, 2025.

Yisong Xiao, Aishan Liu, Yongxin Huang, Zonghao Ying, Shiji Zhao, Tianlin Li, Yong Han, Jian Yang, and Xianglong Liu. Mitigating bias in large vision-language models via counterfactual ensemble decoding. *arXiv preprint arXiv:2608.21415*, 2026.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *International Conference on Learning Representations*, volume 2024, pp. 43478–43502, 2024.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10610–10620. IEEE, 2025.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xinlei Yu, Chengming Xu, Guibin Zhang, Yongbo He, Zhangquan Chen, Zhucun Xue, Jiangning Zhang, Yue Liao, Xiaobin Hu, Yu-Gang Jiang, et al. Visual multi-agent system: Mitigating hallucination snowballing via visual flow. In *International Conference on Learning Representations*, volume 2026, pp. 43581–43606, 2026.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 59670–59684, 2024a.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Barrett Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 15062–15083, 2026a.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog grounding large language models to holistic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14227–14238. IEEE, 2024b.

Zhaohan Zhang, Chengzhengxu Li, Xiaoming Liu, Chao Shen, Ziquan Liu, and Ioannis Patras. Confidence should be calibrated more than one turn deep. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38578–38594, 2026b.

Pengcheng Zheng, Chaoning Zhang, Mingzheng Cui, Guo Chen, Qigan Sun, Jiaxin Huang, Jiaquan Zhang, Tae-Ho Kim, Caiyan Qin, Yazhou Ren, et al. Towards visual chain-of-thought reasoning: A comprehensive survey. 2026.

Jiawei Zhu, Yishu Liu, Huanjia Zhu, Hui Lin, Yuncheng Jiang, Zheng Zhang, and Bingzhi Chen. Combating visual question answering hallucinations via robust multi-space co-debias learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 955–964, 2024.

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4189–4199. IEEE, 2025.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *Forty-second International Conference on Machine Learning (ICML)*, 2025.

Supplementary Appendix

Seeing Clearly, Remembering Selectively:
Mitigating Multi-Agent Hallucination Snowballing via Evidential Modulation

In this appendix, we provide supplementary information and further discussion on our method analysis and the implementation details of DELVE in Section A and Section B, respectively. Moreover, we provide additional experimental details and results in Section C.

A Further Discussion of the Method

Following the main-text notation, we interpret the latent memory evidence used by CUMS in terms of directional uncertainty and reliability discounting.

Directional evidence as a belief assignment. For CUMS in Section 3.3, consider the binary frame $\Omega = \{+, -\}$, representing the sign of a nonzero latent memory evidence component $\delta_{l,j}^{(u)}$. Using the terminal-layer set $\mathcal{L}$ and the relative evidential support $r_{l,j}^{(u)}$ defined in the main text, define the accumulated support for each direction as:

$$E_j^+ = \sum_{\substack{l \in \mathcal{L} \\ \delta_{l,j}^{(u)} > 0}} r_{l,j}^{(u)}, \qquad E_j^- = \sum_{\substack{l \in \mathcal{L} \\ \delta_{l,j}^{(u)} < 0}} r_{l,j}^{(u)}. \tag{11}$$

Since $\delta_{l,j}^{(u)} = 0$ implies $r_{l,j}^{(u)} = 0$, we have $E_j^+ + E_j^- = \sum_{l \in \mathcal{L}} r_{l,j}^{(u)}$. Define an auxiliary basic belief assignment m_j by:

$$m_j(\{+\}) = \frac{E_j^+}{E_j^+ + E_j^- + \varepsilon}, \quad m_j(\{-\}) = \frac{E_j^-}{E_j^+ + E_j^- + \varepsilon}, \quad m_j(\Omega) = \frac{\varepsilon}{E_j^+ + E_j^- + \varepsilon}, \tag{12}$$

where $m_j(\varnothing) = 0$ and $\varepsilon > 0$ is the same stabilizer as in the main text. These masses are nonnegative and sum to one. The singleton masses summarize support for the two directions, and $m_j(\Omega)$ is the residual uncertainty mass. This auxiliary assignment describes directional support, rather than the probability that an upstream statement is correct.

Separating directional conflict from ignorance. Substitution gives $q_j^{(u)} = m_j(\{+\}) - m_j(\{-\})$ and the following exact identity for $\kappa_j^{(u)}$:

$$\kappa_j^{(u)} = \underbrace{m_j(\Omega)}_{\text{vacuity}} + \underbrace{2\min\{m_j(\{+\}), m_j(\{-\})\}}_{\text{binary dissonance}} = \frac{\varepsilon + 2\min(E_j^+, E_j^-)}{E_j^+ + E_j^- + \varepsilon}. \tag{13}$$

The second term is the binary specialization of subjective-logic dissonance (Hu et al., 2021), quantifying opposing committed support. It is zero when either singleton mass is zero. Thus, $\kappa_j^{(u)}$ combines directional conflict and lack of support. Balanced opposing support yields $\kappa_j^{(u)} = 1$, whereas one-sided support much larger than ε makes it approach zero. With no support, $\kappa_j^{(u)} = 1$, but $s_{l,j}^{(u)} = 0$ for all $l \in \mathcal{L}$ because the corresponding latent evidence components are zero. The vacuity is relative to the numerical regularizer ε, rather than a calibrated measure of predictive uncertainty.

Connection to reliability discounting. Classical reliability discounting reduces committed belief and transfers the removed mass to ignorance (Calderwood et al., 2017). We assume that corresponding latent dimensions are comparable across the selected terminal layers, allowing cross-layer directional agreement to guide the retention of memory evidence. Taking $1 - \kappa_j^{(u)} = |q_j^{(u)}|$ as the retention factor yields the following discounted belief assignment:

$$\begin{aligned} m_j^{\text{disc}}(\{a\}) &= (1 - \kappa_j^{(u)}) m_j(\{a\}), \qquad a \in \Omega, \\ m_j^{\text{disc}}(\Omega) &= \kappa_j^{(u)} + (1 - \kappa_j^{(u)}) m_j(\Omega). \end{aligned} \tag{14}$$

Algorithm 1 Inference process of DELVE

Input: Image-question pair (I, q); agents $\{\mathcal{A}_i\}_{i=1}^{N}$ indexed in MAS execution order; decision-anchor size K_t; number of terminal layers $1 \le K_l \le L$; hyperparameters $\lambda \ge 0$, τ, $T_e > 0$, and $\varepsilon > 0$.

Define $\mathcal{L} = \{L - K_l + 1, \dots, L\}$.
Use prompt-only forward passes with DELVE modulation disabled for evidence extraction.
for $i = 1, 2, \dots, N$ **do**
  Retrieve upstream memory U_i according to the MAS protocol, with $U_1 = \varnothing$.
  // Latent-Space Evidence Modeling.
  Construct visual contexts $\mathcal{C}^{(v,+)} = (q, I, U_i)$ and $\mathcal{C}^{(v,-)} = (q, \varnothing, U_i)$.
  Pool the last K_t prompt-token states in each context ($1 \le K_t \le \min\{T^{(v,+)}, T^{(v,-)}\}$) to obtain $\boldsymbol{z}_l^{(v,\pm)}$. And compute latent visual evidence $\boldsymbol{\delta}_l^{(v)}$ and relative evidential support $\boldsymbol{r}_l^{(v)}$ for $l = 1, \dots, L$ via Equations (3) and (4).
  if $i = 1$ **then**
    Cache the pristine latent visual evidence $\boldsymbol{\delta}_l^{(v_0)} \leftarrow \boldsymbol{\delta}_l^{(v)}$ for every layer l, using $(q, I, \varnothing)$ and $(q, \varnothing, \varnothing)$.
  end if
  // Pristine Visual Evidence Reinforcement.
  Compute $g_{l,j} = \sigma((r_{l,j}^{(v)} - \tau)/T_e)$ and $\boldsymbol{g}_l = [g_{l,1}, \dots, g_{l,d}]$ for every layer l.
  Form the weighted pristine visual vector $\boldsymbol{v}_l = \boldsymbol{g}_l \odot \boldsymbol{\delta}_l^{(v_0)}$ via Equation (8).
  // Conflict-Aware Unreliable Memory Suppression.
  Initialize $\boldsymbol{s}_l^{(u)} = \mathbf{0}$ for $l = 1, \dots, L$.
  if $U_i \neq \varnothing$ **then**
    Construct memory contexts $\mathcal{C}^{(u,+)} = (q, I, U_i)$ and $\mathcal{C}^{(u,-)} = (q, I, \varnothing)$.
    Pool the last K_t prompt-token states in each context ($1 \le K_t \le \min\{T^{(u,+)}, T^{(u,-)}\}$); compute latent memory evidence $\boldsymbol{\delta}_l^{(u)}$ and relative evidential support $\boldsymbol{r}_l^{(u)}$ for $l \in \mathcal{L}$ via Equations (3) and (4).
    Compute cross-layer directional consistency $q_j^{(u)}$ and conflict scores $\kappa_j^{(u)}$ via Equation (9).
    Form $\boldsymbol{\kappa}^{(u)} = [\kappa_1^{(u)}, \dots, \kappa_d^{(u)}]$ and $\boldsymbol{s}_l^{(u)} = \boldsymbol{\kappa}^{(u)} \odot \boldsymbol{\delta}_l^{(u)}$ for $l \in \mathcal{L}$.
  end if
  During autoregressive decoding, reuse $\boldsymbol{v}_l$ in Equation (7), replacing $\tilde{\boldsymbol{v}}_l$ with $\boldsymbol{v}_l$, to obtain $\boldsymbol{x}_{l,t}^{(v)}$ from each layer output $\boldsymbol{x}_{l,t}$.
  For $l \in \mathcal{L}$, apply $\boldsymbol{s}_l^{(u)}$ and the normalization in Equation (10) to obtain $\tilde{\boldsymbol{x}}_{l,t}$. Otherwise, set $\tilde{\boldsymbol{x}}_{l,t} = \boldsymbol{x}_{l,t}^{(v)}$.
  Generate the agent response $\hat{y}_i$ from the modulated representations and pass it downstream according to the MAS topology.
end for

Output: Final agent answer $\hat{y}_N$.

CUMS follows a corresponding principle at the latent-evidence level. Specifically, its suppression vector satisfies $\delta_{l,j}^{(u)} - s_{l,j}^{(u)} = |q_j^{(u)}| \delta_{l,j}^{(u)}$. Since $|q_j^{(u)}| \le 1$, this identity describes a contraction of each latent evidence component, with weaker directional agreement yielding a smaller retained fraction. In the hidden-state update, CUMS subtracts the resulting suppression vector $\boldsymbol{s}_l^{(u)}$ from $\boldsymbol{x}_{l,t}^{(v)}$ before final normalization. The shared retention factor thus provides a reliability-discounting interpretation of the latent evidence suppression.

Additionally, to facilitate a comprehensive understanding of the overall procedure, we provide a pipeline algorithm describing the execution of our method, as shown in Algorithm 1.

B SUPPLEMENTARY IMPLEMENTATION DETAILS

In this section, we provide additional implementation details of DELVE, covering the multi-agent communication topologies and upstream memory management. We further analyze its computational overhead and compare its inference runtime with the baseline model and Saliency.

B.1 DETAILED INTRODUCTION TO MULTI-AGENT SYSTEMS

Communication topologies. Our MAS follows the experimental settings of (Yu et al., 2026). Since the released code does not provide a complete MAS implementation, we build our system upon MacNet (Qian et al., 2025), one of the earliest and most widely adopted MAS. We adopt a decentralized design with peer generators and consider three communication patterns: sequential forwarding (*Chained*), dense connectivity (*Mesh*), and randomly sampled links (*Random*). Interactions follow the topological order of a directed dependency graph. A *reviewer (critic)* evaluates the response transmitted along an edge and provides feedback, which the downstream *generator (actor)* uses to revise its answer to the image-question pair (Yu et al., 2026).

Upstream memory control. We distinguish the local dialogue used for refinement from the historical memory passed between agents. Local memory retains the exchanges needed within an interaction, whereas downstream historical memory U receives only the finalized outputs of immediately connected upstream agents. Earlier information is therefore carried forward implicitly through these outputs, without repeatedly appending complete upstream dialogues or their intermediate analyses. When an agent has multiple predecessors, their finalized outputs jointly form its upstream context.

This output-based memory mechanism preserves contextual continuity while limiting context growth. Under the dense-network analysis of (Qian et al., 2025), it reduces the sink agent's context-token cost from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$, where n denotes the number of agent nodes, assuming bounded message lengths and a fixed number of local refinement turns. This complexity concerns the context processed by the sink agent, rather than the total computation of the entire MAS. Unless otherwise specified, we evaluate each sample over ten MAS interaction turns.

B.2 COMPUTATIONAL OVERHEAD OF DELVE

DELVE introduces additional computation through prompt-level evidence extraction and token-level latent-space modulation. For each agent context, a fixed set of source-conditioned forward passes estimates the visual and memory evidence before answer generation. These passes process the input prompts without generating auxiliary answers, and their computational cost depends on the input length. The resulting modulation vectors are then reused throughout the response. Compared with decoding methods that require auxiliary forward passes at every generation step, DELVE avoids repeated auxiliary model evaluations during decoding. Specifically, PVER and CUMS operate on existing hidden states through element-wise operations, inner products, and normalization, adding $\mathcal{O}(Ld)$ arithmetic operations per generated token, where L denotes the number of decoder layers and d denotes the hidden dimension. Generation retains a single autoregressive stream, without additional attention or feed-forward network evaluations for the interventions at each decoding step.

The following experimental results compare the inference time of our method with the baseline and a recent training-free approach on the same set of POPE samples, as shown in Table 5. Although our method incurs additional overhead, its inference time remains below $2\times$ (1.61) that of the baseline. It is also more than $4\times$ faster than the existing training-free *Saliency* (Zhang et al., 2026a). This efficiency gain mainly comes from directly operating on latent representations, avoiding the repeated full decoding required by prior approaches.

C ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

In this section, we discuss the reproducibility considerations for ViF and provide additional details on our experimental configurations, evaluation benchmarks, and metrics. We also present supplementary visualizations and qualitative cases.

Table 5: Runtime comparison of different methods. Lower is better for all metrics. Slowdown is measured relative to baseline's per-sample runtime.

Method	Round (s) ↓	Sample (s) ↓	Agent turn (s) ↓	Slowdown ↓
Baseline	146.254	14.086	0.387	1.000×
Ours	232.628	22.705	1.232	1.612×
Saliency	981.798	97.613	8.627	6.930×

Table 6: Default fixed parameter settings of our DELVE.

Parameter	Description	Default value
τ	Relative support threshold	0.02
T_e	Soft-weighting temperature	0.15
ε	Numerical stabilizer	10^{-6}

C.1 Discussion of ViF

ViF (Yu et al., 2026) addresses hallucination snowballing through visual information relay and attention reallocation, making it a relevant comparison for our study. However, the publicly available resources did not allow us to establish a faithful reproduction of its reported results under our evaluation protocol. Specifically, the released repository describes a two-stage procedure involving pretraining and instruction tuning, but its training scripts instantiate a placeholder VLM rather than the backbones evaluated in the paper. We could not identify a complete reproduction setup covering the evaluated backbones, the corresponding training-data specification, and trained checkpoints. Obtaining numerical results would therefore require additional implementation and training choices whose consistency with the original experiments could not be verified. We consequently omit this comparison to avoid attributing results from an unvalidated reimplementation to ViF.

Our work pursues a fully training-free approach to the same broader challenge. In contrast to the training workflow described in the public ViF repository, DELVE keeps all model parameters frozen and requires neither additional training data nor optimization of auxiliary modules. Its two complementary interventions are designed to jointly mitigate intra-agent hallucination generation and inter-agent hallucination propagation: PVER reinforces pristine visual evidence to support visual grounding, while CUMS selectively attenuates upstream memory contributions according to their cross-layer directional inconsistency. This design aims to reduce the influence of unreliable memory while retaining useful information for collaboration.

C.2 More implementation details in Experiments

Unless otherwise specified, we use the default configurations summarized in Table 6. We use greedy decoding for all base MLLMs. The support threshold τ and gating temperature T_e are held fixed at their reported values, while ε serves as a numerical stabilizer. To control variability arising from graph sampling, all compared methods use the same fixed directed graph under the default Random setting. The complete edge list is provided below. This ensures that comparisons are conducted under identical communication dependencies.

Fixed Random Communication Topology in Our Experiments

```
# Format: directed edge list ("from->to")
# Nodes: 0--9
# Edges: 24

[
  "0->1",
  "2->4", "2->5", "2->6", "2->8",
  "3->0", "3->2", "3->6",
  "4->0",
```

```
    "5->0", "5->4", "5->6",
    "6->1", "6->9",
    "7->3", "7->5", "7->6", "7->8",
    "8->1", "8->4", "8->5", "8->9",
    "9->0", "9->4"
]
```

C.3 Evaluation Benchmarks

We evaluate our method on three hallucination-oriented benchmarks, including POPE (Li et al., 2023), MMHal-Bench (Sun et al., 2024), and HallusionBench (Guan et al., 2024), as well as three general-purpose multimodal benchmarks, MME (Fu et al., 2026), MMBench (Liu et al., 2024b), and MM-Vet (Yu et al., 2023). The details of each benchmark and the corresponding evaluation metrics are described below.

POPE-R. POPE (Li et al., 2023) evaluates object hallucination through binary object-existence questions. The Random set (*i.e.*, POPE-R) contains 3,000 questions constructed from 500 images, with three positive and three negative questions for each image. For negative questions, the queried objects are randomly sampled from object categories absent from the corresponding image. We report the F1 score:

$$\mathrm{F1} = \frac{2\,\mathrm{Precision}\,\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{15}$$

where "yes" is treated as the positive class.

POPE-P. The Popular split (*i.e.*, POPE-P) also consists of 3,000 binary questions over 500 images, with an equal number of positive and negative questions. Different from POPE-R, negative objects are sampled from categories that frequently occur in the overall dataset but are absent from the queried image, making the benchmark more susceptible to language and object-frequency priors. We also report the F1 score following the same protocol as POPE-R.

POPE-A. The Adversarial split (*i.e.*, POPE-A) contains another 3,000 questions over 500 images and constitutes the most challenging POPE setting. Its negative objects are selected from absent categories that frequently co-occur with objects actually present in the image, thereby deliberately inducing object-association hallucinations. We again use the F1 score as the primary evaluation metric.

MMHal-Bench. MMHal-Bench (Sun et al., 2024) contains 96 challenging image-question pairs collected from OpenImages and spans eight question categories. It evaluates free-form multimodal responses with an LLM-based judge that jointly considers informativeness and hallucination. Each response receives a score $s_i \in \{0, \ldots, 6\}$, where scores 0-2 indicate hallucinated answers and 3-6 indicate non-hallucinated answers with increasing informativeness. We report the mean MMHal score,

$$\mathrm{Score} = \frac{1}{N}\sum_{i=1}^{N} s_i, \qquad N = 96, \tag{16}$$

where a higher score indicates a more informative and better-grounded response.

HallusionBench. HallusionBench (Guan et al., 2024) is a human-curated benchmark for diagnosing both visual illusion and language hallucination. It contains 346 figures and 1,129 yes/no questions organized into carefully constructed control groups, covering diverse visual and reasoning scenarios. We report all-question accuracy (**aAcc**), question-pair accuracy (**qAcc**), and figure accuracy (**fAcc**). Let $b_M(I, q) \in \{0, 1\}$ indicate whether model correctly answers question q for image I. The all-question accuracy is:

$$\mathrm{aAcc} = \frac{1}{|\mathcal{V}|}\sum_{(I,q)\in\mathcal{V}} b_M(I, q). \tag{17}$$

The stricter qAcc requires all members of a corresponding question pair to be answered correctly, while fAcc requires all questions associated with the same figure to be correct:

$$\text{qAcc} = \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} \mathbb{I}\left[\bigwedge_{(I,q) \in Q} b_M(I,q) = 1 \right], \tag{18}$$

$$\text{fAcc} = \frac{1}{|\mathcal{I}|} \sum_{F \in \mathcal{I}} \mathbb{I}\left[\bigwedge_{(I,q) \in F} b_M(I,q) = 1 \right]. \tag{19}$$

Higher values indicate better performance for all three metrics.

MME. MME (Fu et al., 2026) is a comprehensive multimodal evaluation benchmark covering 14 subtasks, including 10 perception tasks and four cognition tasks. Each image is associated with a pair of manually constructed yes/no questions, one with a positive answer and the other with a negative answer. For each subtask, MME defines:

$$S_t = \text{Acc}_t + \text{Acc}_t^+, \tag{20}$$

where Acc is conventional question-level accuracy and Acc^+ counts an image as correct only when both associated questions are answered correctly. We report MME-P, MME-C, and MME-Total. To be specific,

$$\text{MME-P} = \sum_{t \in \mathcal{T}_P} S_t, \qquad \text{MME-C} = \sum_{t \in \mathcal{T}_C} S_t, \tag{21}$$

and

$$\text{MME-Total} = \text{MME-P} + \text{MME-C}, \tag{22}$$

where the maximum scores are 2,000, 800, and 2,800, respectively.

MMBench. MMBench (Liu et al., 2024b) is a general-purpose multiple-choice benchmark covering 20 fine-grained multimodal abilities in perception and reasoning. In our experiments, we only use the English *Dev Set*, which contains 1,164 questions, as the official evaluation service for the test set is no longer available. We report vanilla accuracy (V-Acc), circular accuracy (C-Acc), and pass accuracy (P-Acc).

Vanilla accuracy is the conventional single-pass multiple-choice accuracy,

$$\text{V-Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = y_i]. \tag{23}$$

For a question with K_i answer choices, circular accuracy evaluates the model under all K_i cyclic permutations of the choices and regards the question as correct only when every permutation is answered correctly:

$$\text{C-Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\bigwedge_{k=1}^{K_i} \hat{y}_{i,k} = y_{i,k} \right]. \tag{24}$$

We additionally report pass accuracy, where each evaluation under a particular option permutation is treated as an individual pass. It is computed as the fraction of correctly answered passes among all evaluated passes:

$$\text{P-Acc} = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K_i} \mathbb{I}[\hat{y}_{i,k} = y_{i,k}]}{\sum_{i=1}^{N} K_i}. \tag{25}$$

MM-Vet. MM-Vet (Yu et al., 2023) evaluates the integrated multimodal capabilities of large vision-language models using 218 open-ended questions over 200 images. The benchmark covers 6 core capabilities, *i.e.*, recognition, OCR, knowledge, language generation, spatial awareness,

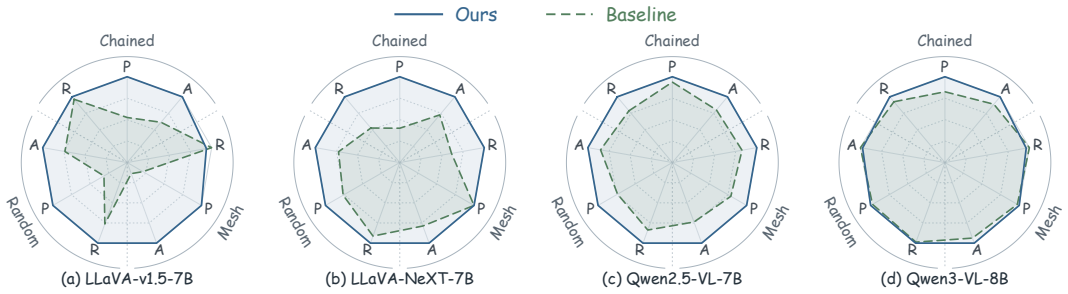


Figure 6: F1 comparison between Baseline and Ours across four backbone models and three MAS structures. R, P, and A denote POPE-R, POPE-P, and POPE-A, respectively. On each spoke, the outer boundary represents the higher F1 of the two methods, with an 8-percentage-point radial range and 2-percentage-point grid intervals.

and mathematics, as well as 16 combinations of these capabilities. Because the answers are open-ended, MM-Vet employs an external evaluator to assign each response a correctness score according to the reference answer. We report the average MM-Vet score over all questions:

$$\text{MM-Vet} = \frac{1}{N}\sum_{i=1}^{N} s_i, \qquad N = 218, \tag{26}$$

which is conventionally reported on a 0-100 scale, with higher values indicating better integrated multimodal understanding.

C.4 Additional Evaluation MHS Metrics

We report three additional metrics to characterize hallucinations during reasoning: the Hallucination Snowballing (HS) score, the average number of erroneous turns (ET), and the number of samples with intermediate error. HS evaluates hallucination severity and propagation in multi-agent systems, whereas the latter two metrics assess intra-agent hallucinations from complementary perspectives.

Hallucination Snowballing (HS) score. HS measures the extent of hallucination snowballing by jointly considering hallucination severity and propagation distance. Following (Yu et al., 2026), it is computed as:

$$\text{HS} = \frac{1}{A}\sum_{a=1}^{A} \frac{h_a}{1+\exp\left(D/2 - d_a\right)}, \tag{27}$$

where A is the number of agents, d_a denotes the hallucination propagation distance associated with agent a, D is the total propagation distance, and h_a denotes hallucination severity assessed by an external judge on a 100-point scale. The sigmoid weighting assigns larger weights to longer propagation distances, yielding higher scores for more severe hallucinations that propagate farther.

Average number of Erroneous Turns (ET). ET measures how frequently hallucinations occur across the ten reasoning turns of each sample.

Number of samples with intermediate error. This metric measures how many samples encounter at least one hallucination during reasoning, including those whose errors are subsequently corrected. Each sample is counted once if any reasoning turn contains a hallucination, regardless of the number of erroneous turns or the correctness of the final answer. ET captures the frequency of hallucinations within samples, while the number of samples with intermediate error captures their occurrence across samples. Lower values indicate better performance for all three metrics.

C.5 Additional Experiments

In this section, we provide additional experiments to further evaluate our DELVE. Specifically, we first report performance comparisons across different benchmark subsets (*e.g.*, POPE), depicted

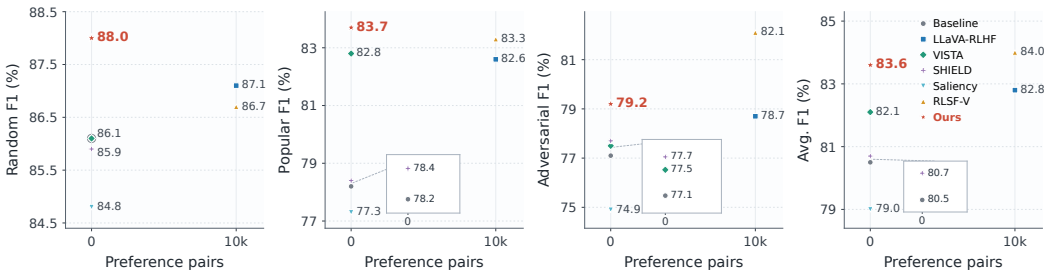


Figure 7: Comparison with existing hallucination mitigation methods for single MLLMs on POPE. From left to right, the panels show F1 scores on the *Random*, *Popular*, and *Adversarial* subsets, followed by the average F1 score.

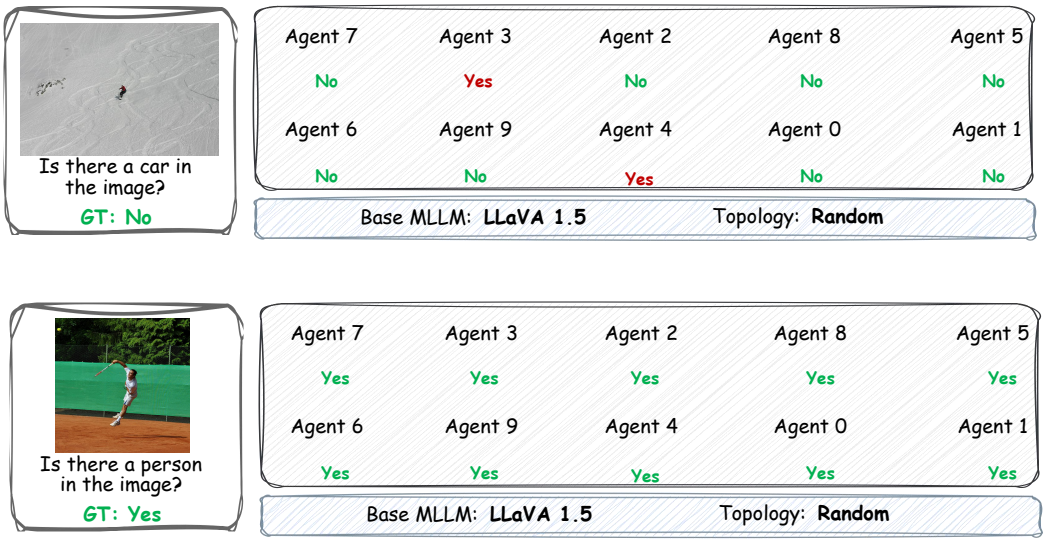


Figure 8: Additional cases of MHS mitigation by DELVE using LLaVA-1.5-7B.

in Figure 6. We then compare our method with existing hallucination mitigation methods for single MLLMs on POPE, as illustrated in Figure 7. Notably, the x-axis denotes the number of preference data pairs required by existing methods for fine-tuning. Finally, to provide a more comprehensive understanding, we present additional qualitative cases of hallucination mitigation on VQA tasks, as shown in Figure 8 to Figure 13. Overall, the results show that our DELVE achieves superior hallucination mitigation with minimal training cost, demonstrating its effectiveness in alleviating MHS.

D Ethics Statement

This work adheres to the ICLR 2027 Code of Ethics. Our study focuses on hallucination mitigation in multimodal multi-agent systems using existing benchmarks and involves no human-subject studies or animal experiments.

E Reproducibility Statement

We follow the ICLR 2027 reproducibility guidelines. Implementation and evaluation details are provided in Sections 4.1 of the main paper and Appendix B, Appendix C.3, and Appendix C.2 . We will release our code upon acceptance.

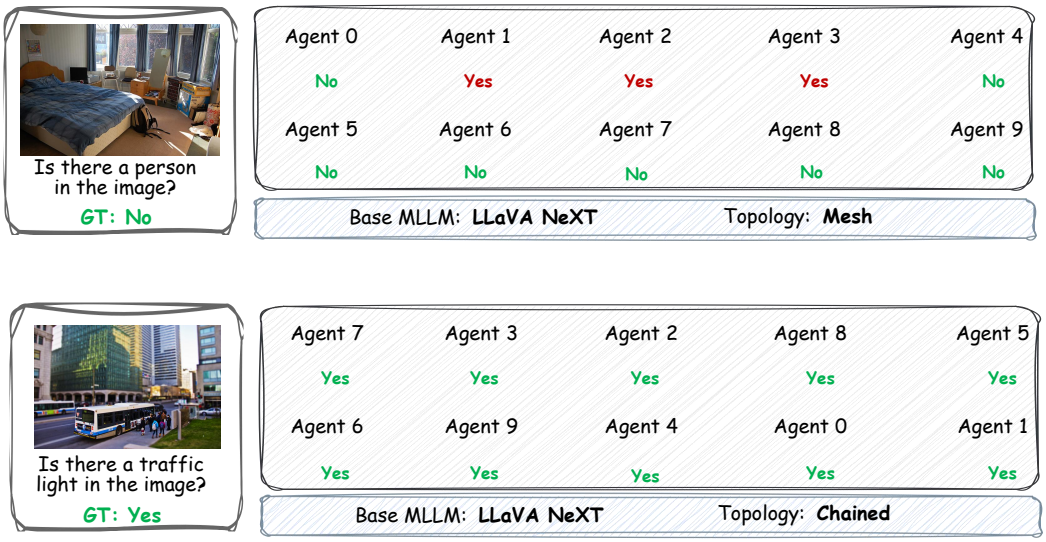


Figure 9: Additional cases of MHS mitigation by DELVE using LLaVA-NeXT-7B.

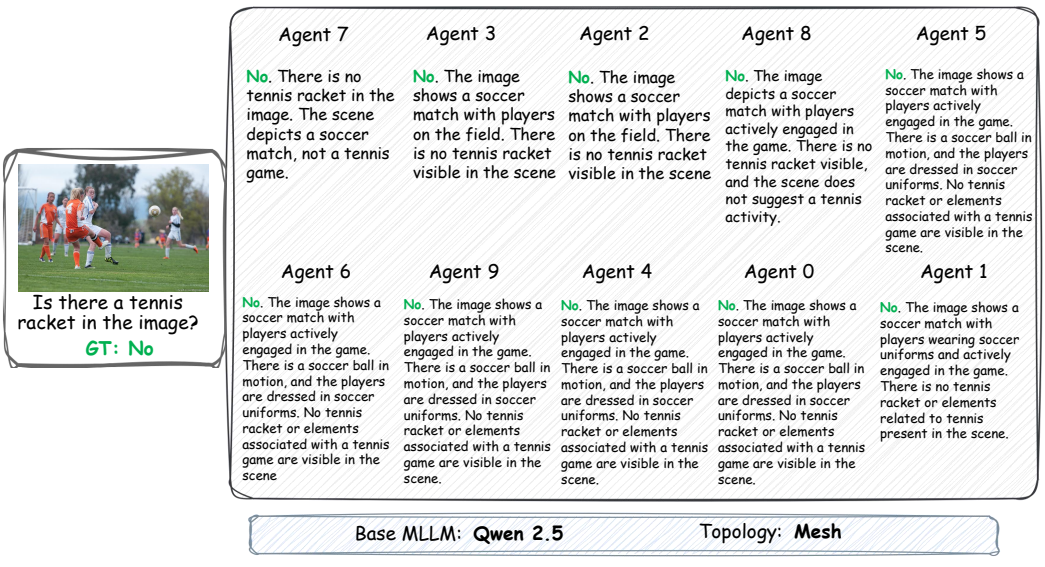


Figure 10: Additional cases of MHS mitigation by DELVE using Qwen2.5-VL-7B.

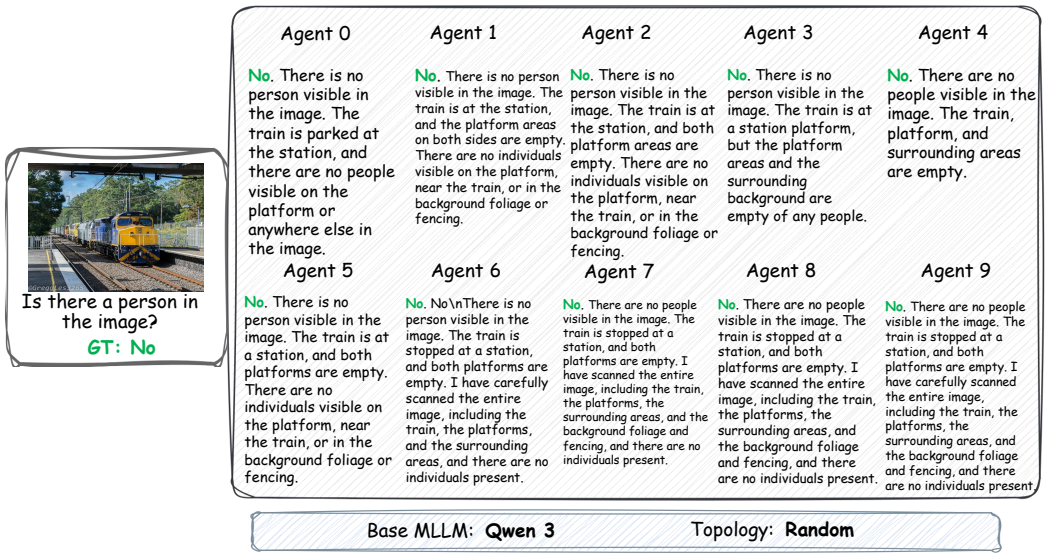


Figure 11: Additional cases of MHS mitigation by DELVE using Qwen3-VL-8B.

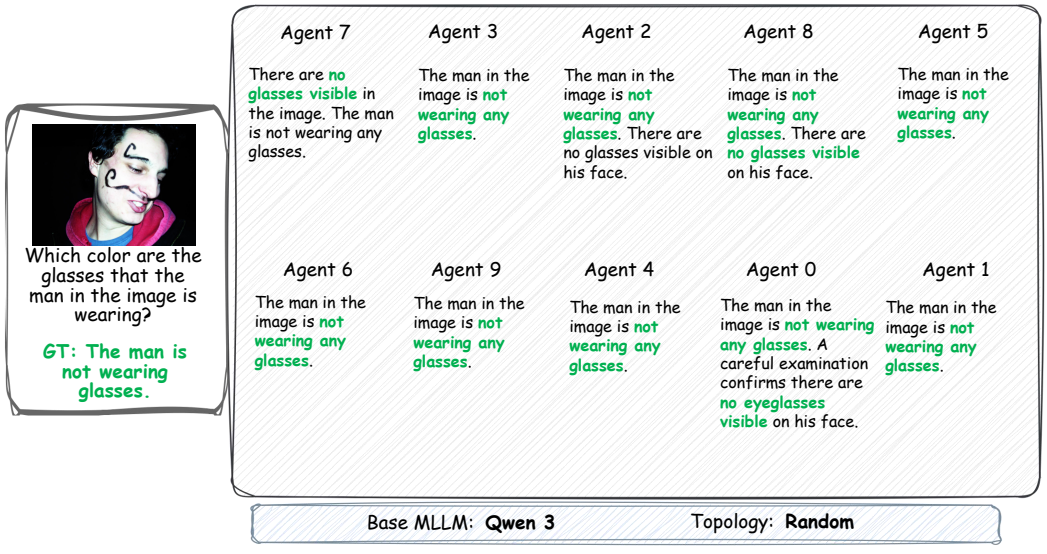


Figure 12: Additional cases of MHS mitigation by DELVE using Qwen3-VL-8B.

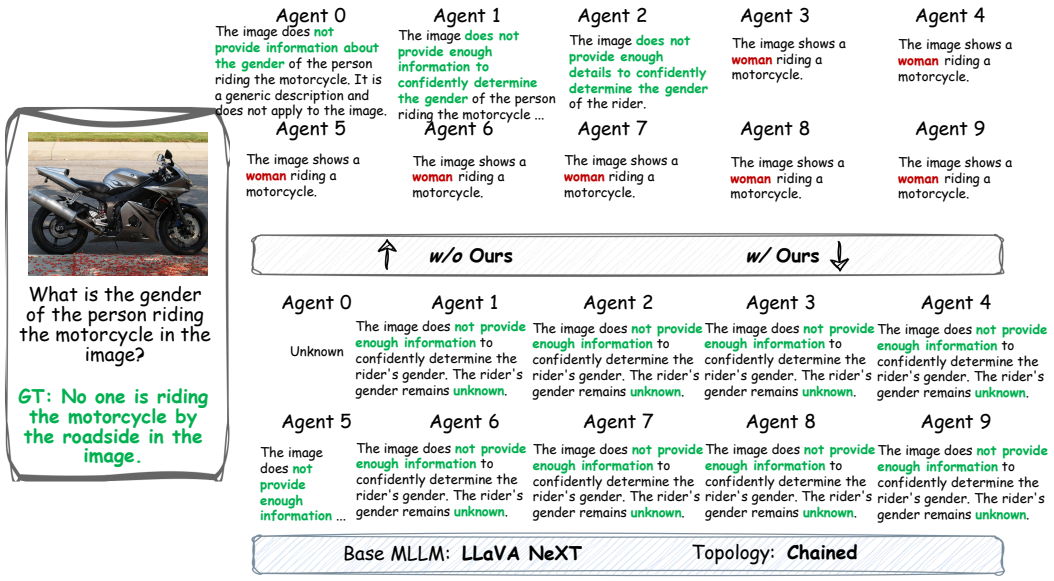


Figure 13: Additional comparisons between the baseline and DELVE using LLaVA-NeXT-7B.