WHEN DELIMITERS MISLEAD: VISUAL-TEXT REGROUNDING FOR MULTI-IMAGE HALLUCINATION MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) are increasingly capable of reasoning over multiple images, but this capability also introduces new hallucination risks. In this paper, we reveal that delimiter tokens designed to separate multiple images are associated with two distinct internal patterns on the visual and textual sides: *Visual-Delimiter Leakage* and *Text-Delimiter Sink*, both of which contribute to multi-image hallucination. To address these issues, we propose a training-free method called Delimiter-Aware Regrounding (DAR). DAR consists of two modules corresponding to the two internal patterns: Delimiter Ownership Correction (DOC) mitigates *Visual-Delimiter Leakage* by restoring visual-token ownership, while Visual Evidence Redirection (VER) strengthens image-specific visual grounding to alleviate *Text-Delimiter Sink*. Extensive experiments on multi-image benchmarks across multiple MLLMs show that DAR effectively enhances multi-image reasoning while preserving the structural role of delimiter tokens in separating visual inputs. Moreover, DAR provides a new perspective on hallucination mitigation in broader multi-input scenarios, such as video reasoning.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have shown strong capabilities in visual understanding and multimodal reasoning (Wang et al., 2024; Chen et al., 2024a;b; Zhu et al., 2024; Tian et al., 2025), yet hallucination remains a persistent challenge (Guan et al., 2024; Leng et al., 2024). In single-image scenarios, hallucination refers to generated content that is not faithfully supported by the given image, reflecting a mismatch between textual generation and visual evidence (Bai et al., 2024). Prior studies have extensively investigated this problem and commonly attribute it to insufficient visual grounding and overreliance on language priors (Leng et al., 2024; He et al., 2025). Beyond the single-image setting, real-world visual reasoning increasingly involves multi-image scenarios, where MLLMs are required to jointly process multiple visual inputs (Chen et al., 2026; Yin et al., 2026b; Li et al., 2024). Despite this growing demand, multi-image hallucination remains relatively under-explored compared with single-image hallucination. More importantly, directly extending the conventional view of single-image hallucination to the multi-image setting is insufficient. Multi-image reasoning requires visual evidence to remain reliably associated with its source image, which we refer to as the *image ownership relation*. When this ownership relation is not well maintained, visual evidence from one image may leak into the reasoning process for another image, leading to *cross-image information leakage*. Such leakage can entangle visual information across different images and increase the risk of generating hallucinated answers (Park et al., 2025). In multi-image inputs, visual tokens from different images are concatenated into a shared context. To preserve the ownership relation, they are typically separated by special delimiter tokens. For example, in Qwen3.5, each image is bracketed by `<|vision_start|>` and `<|vision_end|>`. These delimiter tokens are essential structural elements in multi-image reasoning. They mark image boundaries and function as implicit image identity cues, thereby helping maintain the image ownership relation and facilitating intra-image interactions (Lee et al., 2026).

However, delimiter tokens also introduce unintended interactions that contribute to multi-image hallucination, manifesting as two phenomena. ① *Visual-Delimiter Leakage*: visual tokens attend to delimiter tokens of other images and thus weaken their ownership relation with their source image,

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.

potentially causing cross-image information leakage. ❷ *Text-Delimiter Sink*: text tokens allocate a disproportionately large fraction of their image-related attention to delimiter tokens relative to visual tokens, indicating strong delimiter reliance and comparatively weaker direct visual grounding. To further analyze these delimiter-associated internal patterns, we introduce two internal diagnostic scores. ❶ **Visual-Delimiter Leakage Score (VDLS)** measures how much delimiter-related attention from visual tokens is assigned to non-own delimiter tokens. ❷ **Text-Delimiter Sink Score (TDSS)** measures the extent to which image-related attention is absorbed by delimiter tokens. These scores provide diagnostic signals for the two internal patterns and guide the subsequent intervention. To address the delimiter-associated multi-image hallucination, we propose a training-free method called **Delimiter-Aware Regrounding (DAR)**. DAR follows a simple principle: delimiter tokens should serve as image-identity anchors for calibrating visual ownership, while reasoning should remain grounded in the corresponding visual evidence. ✣ On the visual side, we propose **Delimiter Ownership Correction (DOC)** to adaptively reallocate the attention mass that visual tokens assign to wrong delimiter tokens back to their own delimiter tokens, thereby consolidating visual ownership while preserving the total delimiter-attention budget. ✣ On the text side, we propose **Visual Evidence Redirection (VER)**. It first identifies delimiter-reliant heads based on TDSS. For the selected heads, VER constructs local visual evidence, and then redirects the contribution of delimiter tokens toward the corresponding visual evidence. DOC and VER form a complementary framework that mitigates multi-image hallucination by addressing delimiter-associated patterns on both the visual and textual sides. Our work forms a *closed loop* for multi-image hallucination mitigation:

- ➤ OBSERVATION → We study multi-image hallucination from a delimiter-centered perspective and observe two internal phenomena in MLLMs (Figure 1 left): *Visual-Delimiter Leakage* and *Text-Delimiter Sink*.
- ➤ DIAGNOSIS → We introduce two internal diagnostic scores, VDLS and TDSS, and conduct a comprehensive *empirical diagnostic study*, revealing how delimiter tokens affect multi-image reasoning.
- ➤ INTERVENTION → Based on the diagnostic findings, we propose DAR, a diagnostic-guided intervention method that consists of DOC and VER.
- ➤ MITIGATION → We conduct extensive evaluations across multiple MLLMs and benchmarks, demonstrating that DAR consistently reduces multi-image hallucination. We further extend DAR to video reasoning, showing its potential for broader multi-input scenarios.

2 RELATED WORK

Multi-Image Hallucination. Hallucination in MLLMs has been widely studied, with most existing efforts focusing on the single-image setting, where mitigation methods can be broadly categorized into intervention-based (Liu et al., 2025a; Yin et al., 2026a; Chen et al., 2025), decoding-based (Leng et al., 2024; Woo et al., 2024), and training-based methods (Liu et al., 2024; Yu et al., 2025). Compared with single-image hallucination, multi-image hallucination remains less explored (Min et al., 2026). Existing multi-image hallucination mitigation methods mainly focus on image-level attention allocation or interaction across different images (Liu et al., 2025b; Yang et al., 2026). Recently, several studies have begun to focus on cross-image information leakage, where visual evidence from one image incorrectly interferes with other images. FOCUS (Park et al., 2025) addresses cross-image information leakage through focused decoding, which requires multiple image-wise forward passes with non-target images masked by noise. DTS (Lee et al., 2026) scales the hidden states of image delimiter tokens to suppress undesired cross-image interactions. Building on this delimiter-centered perspective, we further investigate delimiter interactions on both the visual and textual sides, preserving their image-tagging role while selectively recalibrating visual ownership and text-side grounding.

Attention Sink. Attention sink refers to the phenomenon that certain tokens receive disproportionately large attention weights, even when they are not semantically important (Su et al., 2026; Xiao et al., 2024; Gu et al., 2025). Recent work on visual attention sink shows that MLLMs may allocate high attention to certain irrelevant visual tokens, often due to abnormal activation patterns in specific hidden dimensions (Kang et al., 2025). Such visual sink tokens may attract attention while contributing limited meaningful visual evidence, suggesting that attention concentration does not

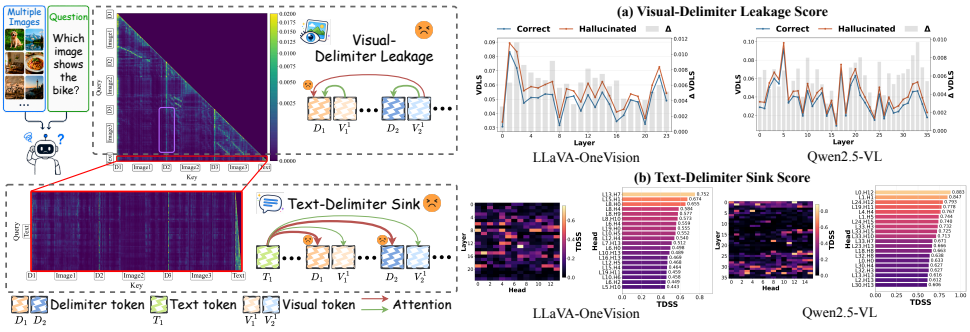


Figure 1: **Left:** *Observation* of two attention patterns associated with image delimiter tokens in multi-image reasoning. Thicker arrows indicate larger attention weights. **Right:** Empirical diagnostic experiments on Mantis. (a) Hallucinated samples exhibit higher VDLS than correct samples at each layer, and the gray bars denote the VDLS gap between the two groups. (b) Text-Delimiter Sink is particularly evident in specific layers (L) and attention heads (H).

necessarily imply useful grounding. These findings motivate a broader view that attention sink can emerge in different token types, such as delimiter tokens. Our work differs from prior attention sink studies in both setting and mechanism. We study delimiter tokens in multi-image MLLMs, where sink behavior interacts with visual-token ownership and image-specific visual grounding.

3 Empirical Diagnosis of Delimiter-Associated Patterns

Delimiter tokens are introduced as structural anchors to organize multiple images, but their actual effects inside MLLMs are more complex. In this section, we investigate delimiter-associated attention patterns related to unreliable multi-image reasoning on the visual and textual sides.

3.1 How Delimiters Affect Visual Tokens

We first analyze the influence of delimiters on visual tokens. For multi-image inputs, the image ownership of each visual token is determined by the input layout. A visual token $v \in V_i$ that belongs to Image i should be associated with the delimiter tokens for Image i. However, we observe that such correspondence is not always stable. Some visual tokens assign substantial attention to delimiters of other images instead of consistently attending to their own image delimiter. This phenomenon indicates that the internal ownership boundary of visual tokens can become ambiguous. When visual tokens are influenced by non-own delimiters, information across different images may become entangled. We refer to this phenomenon as *Visual-Delimiter Leakage*. To quantify this phenomenon, we propose the **Visual-Delimiter Leakage Score (VDLS)**. For a visual token $v \in V_i$, let $\mathcal{D}(v)$ denote the set of causally visible image delimiter tokens, and let $\mathcal{D}_i(v)$ denote its own-delimiter set. We first compute the total delimiter attention mass $\rho_d(v) = \sum_{d\in\mathcal{D}(v)} A_{v,d}$. Then, we compute the proportion of delimiter attention assigned to its own delimiter token $q_v = \frac{\sum_{d\in\mathcal{D}_i(v)} A_{v,d}}{\rho_d(v)+\epsilon}$. A smaller q_v indicates that more delimiter-related attention is assigned to non-own delimiters. To avoid over-interpreting tokens that barely attend to any delimiter, we further introduce a delimiter-signal reliability term:

$$\eta_v = \frac{\rho_d(v)}{\rho_d(v) + \mu_d + \epsilon}, \tag{1}$$

where μ_d denotes the average delimiter attention mass over visual tokens. The token-level Visual-Delimiter Leakage is then defined as $\mathrm{VDL}(v) = \eta_v\,(1 - q_v)$. Finally, VDLS is computed by averaging VDL at layer ℓ:

$$\mathrm{VDLS}^{\ell} = \mathbb{E}_{i,v\in V_i}[\mathrm{VDL}^{\ell}(v)]. \tag{2}$$

A higher VDLS means that visual tokens assign more delimiter-related attention to non-own image delimiters, indicating stronger Visual-Delimiter Leakage.

Diagnostic Analysis of Visual-Delimiter Leakage. We compute VDLS on Mantis and compare hallucinated samples with correct ones. Following our task-level operational definition, incorrect predictions are treated as hallucination-related failures for diagnostic analysis. As shown in Figure 1(a), hallucinated samples consistently exhibit higher VDLS than correct samples. This observation can be summarized as:

✤ **Finding 1:** Incorrect responses are associated with stronger *Visual-Delimiter Leakage*.

Unlike single-image hallucination, multi-image hallucination is further associated with unstable image ownership, where visual evidence may fail to remain reliably bound to its source image. When visual tokens fail to maintain a clear association with their own image delimiters, image-specific evidence is more likely to leak across image boundaries, resulting in incorrect responses.

3.2 How Delimiters Affect Text Tokens

As illustrated in Figure 1, some attention heads allocate a substantial fraction of image-related text-token attention to delimiter tokens rather than directly to visual tokens. We refer to this relative dominance of delimiter-mediated routing over direct visual-token attention as *Text-Delimiter Sink*, whereby the model identifies the relevant image at the boundary level but remains weakly grounded in its internal visual evidence. To quantify this phenomenon, we define the **Text-Delimiter Sink Score (TDSS)**. Since attention sink often emerges in specific attention heads (Kang et al., 2025), we compute TDSS at the head level. Specifically, for each layer ℓ and head h, we first compute the total attention mass assigned from text tokens to delimiter tokens:

$$M_{\text{del}}^{\ell,h} = \sum_{t\in\mathcal{T}} \sum_{i=1}^{N} \sum_{d\in\mathcal{D}_i(t)} A_{t,d}^{\ell,h}, \tag{3}$$

where $\mathcal{D}_i(t)$ denotes the text-side delimiter set for Image i, $\mathcal{T}$ denotes the set of text tokens used for diagnosis, N denotes the number of images, and $A_{t,d}^{\ell,h}$ denotes the attention weight from text token t to delimiter token d in layer ℓ and head h. We further compute the total attention mass assigned from text tokens to visual tokens:

$$M_{\text{vis}}^{\ell,h} = \sum_{t\in\mathcal{T}} \sum_{i=1}^{N} \sum_{v\in V_i} A_{t,v}^{\ell,h}. \tag{4}$$

where V_i denotes the set of visual tokens of Image i, and $A_{t,v}^{\ell,h}$ denotes the attention weight from text token t to visual token v in head h and layer l. The head-level TDSS is defined as the fraction of image-related attention assigned to delimiter tokens:

$$\text{TDSS}^{\ell,h} = \frac{M_{\text{del}}^{\ell,h}}{M_{\text{del}}^{\ell,h} + M_{\text{vis}}^{\ell,h} + \epsilon}. \tag{5}$$

A higher $\text{TDSS}^{\ell,h}$ indicates that text tokens allocate a larger proportion of image-related attention to delimiter tokens, suggesting a stronger Text-Delimiter Sink and weaker grounding in visual tokens.

Diagnostic Analysis of Text-Delimiter Sink. We compute TDSS for each attention head to examine its head-level distribution. As shown in Figure 1(b), Text-Delimiter Sink is widely observed across attention heads, while its severity varies substantially from head to head. In LLaVA-OneVision, the highest-TDSS heads mainly appear in middle layers, with L13.H7 reaching 0.752, followed by L15.H1 with 0.674 and L8.H0 with 0.655. In Qwen2.5-VL, high-TDSS heads appear in different layers, including shallow layers such as L0.H12 with 0.883 and L1.H1 with 0.847, as well as deeper layers such as L24.H12 with 0.793. These observations reveal:

✤ **Finding 2:** *Text-Delimiter Sink* is widespread but uneven in severity across heads and layers.

This finding suggests that delimiter-associated grounding failure on the text side should be addressed in a head-aware manner. An effective mitigation strategy should identify and calibrate high-TDSS heads, while avoiding uniform intervention over delimiter-irrelevant heads.

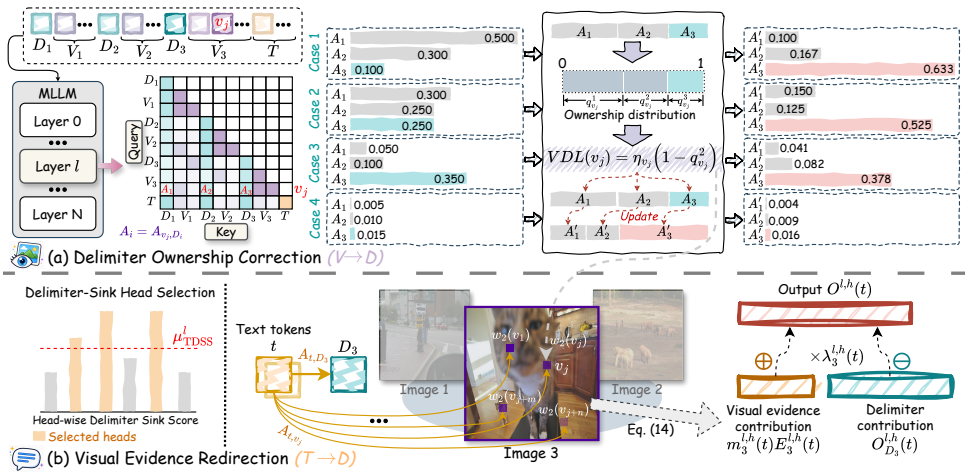


Figure 2: Overview of DAR. (a) DOC estimates the ownership distribution of each visual token over image delimiter tokens and adjusts attention weights to reduce Visual-Delimiter Leakage. The four cases demonstrate that DOC adaptively adjusts its intervention strength according to the severity of Visual-Delimiter Leakage. (b) VER identifies high-TDSS heads with severe delimiter sink and redirects their delimiter contribution toward the corresponding visual evidence.

4 Diagnostic-Guided Intervention

Guided by the diagnostic findings in the previous section, we propose Delimiter-Aware Regrounding (DAR), a diagnostic-guided intervention method that targets multi-image hallucinations. The overview of DAR is illustrated in Figure 2. The core idea of DAR is to use the proposed diagnostic scores as intervention signals, enabling image-specific regrounding. Specifically, DAR consists of two complementary modules: ① **Delimiter Ownership Correction (DOC)** mitigates Visual-Delimiter Leakage by reallocating delimiter attention from non-own delimiters to source-image delimiter tokens. ② **Visual Evidence Redirection (VER)** alleviates Text-Delimiter Sink by transferring delimiter contributions to image-specific visual evidence.

4.1 Delimiter Ownership Correction

When Visual-Delimiter Leakage occurs, image-specific visual evidence may leak across image boundaries, which may increase cross-image interference and weaken image-specific visual grounding. To address this issue, DOC uses VDL as an intervention signal to correct visual-token ownership by adaptively reallocating delimiter attention from non-own image delimiter tokens to source-image delimiter tokens.

Correction Strength. For a visual token $v \in V_i$, $\mathrm{VDL}(v)$ measures the degree of Visual-Delimiter Leakage, indicating how strongly its attention is assigned to non-own image delimiters. DOC uses this diagnostic signal as the ownership correction strength:

$$g_i(v) = \mathrm{VDL}(v). \tag{6}$$

Well-anchored visual tokens yield low leakage values and therefore receive only weak correction from DOC. In contrast, tokens that allocate substantial delimiter attention to other images exhibit larger leakage values and receive stronger correction.

Ownership Correction. DOC operates on the delimiter tokens visible to each visual token. For $v \in V_i$, we denote the own-delimiter set as $\mathcal{D}_i(v)$, and the non-own delimiter set as $\mathcal{D}_{\neg i}(v) = \mathcal{D}(v) \setminus \mathcal{D}_i(v)$. We first compute the attention mass assigned to non-own delimiters: $M_{\neg i}(v) = \sum_{d \in \mathcal{D}_{\neg i}(v)} A_{v,d}$, where $A_{v,d}$ is the attention weight from token v to token d. The removed mass is reassigned to the own delimiter tokens. To preserve the relative structure among the delimiters, we

distribute the reassigned mass according to their original attention proportions:

$$\alpha_{v,d} = \frac{A_{v,d} + \epsilon}{\sum_{u \in \mathcal{D}_i(v)} (A_{v,u} + \epsilon)}, d \in \mathcal{D}_i(v). \tag{7}$$

The corrected attention weights are defined as

$$A'_{v,d} = \begin{cases} A_{v,d}\,(1 - g_i(v))\,, & d \in \mathcal{D}_{\neg i}(v), \\ A_{v,d} + g_i(v) M_{\neg i}(v) \alpha_{v,d}, & d \in \mathcal{D}_i(v). \end{cases} \tag{8}$$

This correction preserves the total delimiter-attention budget and avoids introducing excessive delimiter reliance. Meanwhile, DOC does not fully eliminate attention to non-own image delimiter tokens, thereby preserving necessary cross-image contextual interactions.

Ownership Prior. After ownership correction, DOC derives an ownership prior. This prior guides VER to select visual tokens with more reliable ownership when building visual evidence, preventing delimiter-token contributions from being redirected to visual tokens with potential leakage risk. We compute the updated own-delimiter confidence $q'_v = \frac{\sum_{d \in \mathcal{D}_i(v)} A'_{v,d}}{\sum_{d \in \mathcal{D}(v)} A'_{v,d} + \epsilon}$ and the ownership prior:

$$\omega_i(v) = 1 - \eta_v\,(1 - q'_v)\,. \tag{9}$$

We omit the layer and head indices for clarity. Here, η_v is the delimiter-evidence gate defined in Eq. (1). As v's attention to delimiter tokens vanishes, $\eta_v \to 0$ and $\omega_i(v) \to 1$; this limiting value means that no ambiguity penalty is applied, rather than that ownership is certain. The prior thus transfers DOC's corrected ownership information to VER when delimiter evidence is available.

4.2 Visual Evidence Redirection

TDSS-based Head Selection. As discussed in Finding 2, Text-Delimiter Sink exhibits head-level variation. Therefore, VER reuses TDSS as the signal to select high-sink attention heads for intervention. Specifically, we first compute the layer-wise average TDSS:

$$\mu^{\ell}_{\mathrm{TDSS}} = \frac{1}{H} \sum_{h=1}^{H} \mathrm{TDSS}^{\ell,h}, \tag{10}$$

where H is the number of heads in layer ℓ. The head-selection gate is defined as:

$$\chi_{\ell,h} = \mathbb{I}\left[\mathrm{TDSS}^{\ell,h} > \mu^{\ell}_{\mathrm{TDSS}}\right]. \tag{11}$$

Heads with TDSS values above the layer-wise average are regarded as delimiter-sink heads, and redirection is performed only in these heads.

Visual Evidence Construction. For a text token t and Image i, VER constructs image-local visual evidence from the visual tokens in V_i. First, we compute the image-local visual relevance to identify the visual tokens most relevant to t:

$$r^{(i,\ell,h)}_{t,v} = \mathrm{softmax}_{u \in V_i}\left(S^{\ell,h}_{t,u}\right)_v, \quad v \in V_i, \tag{12}$$

where $S^{\ell,h}_{t,u}$ denotes the attention logit. Since the normalization is restricted to V_i, this relevance captures the visual grounding preference of the text token t within Image i. We then combine this local relevance with the corrected ownership prior from DOC:

$$\pi^{(i,\ell,h)}_{t,v} = \frac{r^{(i,\ell,h)}_{t,v}\,\omega^{\ell,h}_i(v)}{\sum_{u \in V_i} r^{(i,\ell,h)}_{t,u}\,\omega^{\ell,h}_i(u) + \epsilon}. \tag{13}$$

The resulting $\pi^{(i,\ell,h)}_{t,v}$ is an ownership-conditioned evidence posterior. The visual evidence for Image i is then constructed by aggregating visual tokens that are relevant to text token t and have reliable image ownership:

$$E^{\ell,h}_i(t) = \sum_{v \in V_i} \pi^{(i,\ell,h)}_{t,v}\,\mathbf{z}^{\ell,h}_v, \tag{14}$$

where $\mathbf{z}^{\ell,h}_v$ is the value vector of token v.

Table 1: Results on multi-image benchmarks across different MLLMs, reported as accuracy (%). DTS results are reproduced using the official implementation under the default evaluation settings.

Model	Scale	Method	Mantis	BLINK	MIRB	QBench2	Model	Scale	Method	Mantis	BLINK	MIRB	QBench2
Qwen3.5	2B	Baseline	61.29	50.83	53.15	57.20	InternVL3	1B	Baseline	47.00	42.14	38.70	51.00
		+ DTS	58.06	45.85	52.22	57.60			+ DTS	48.85	41.61	39.11	49.40
		+ Ours	**63.59**	**52.16**	**53.87**	**62.60**			+ Ours	**50.69**	**42.56**	**41.58**	**51.30**
	9B	Baseline	74.65	65.33	59.55	81.00		8B	Baseline	67.28	53.97	50.15	75.70
		+ DTS	72.35	63.18	60.47	78.70			+ DTS	60.37	52.34	51.39	74.50
		+ Ours	**76.04**	**65.97**	**61.82**	**81.70**			+ Ours	**68.20**	**54.55**	**52.22**	**76.20**
Qwen2.5-VL	3B	Baseline	63.59	46.03	55.73	62.50	LLaVA-OneVision	0.5B	Baseline	40.09	37.22	31.68	**51.70**
		+ DTS	57.60	44.87	55.62	61.50			+ DTS	37.79	37.74	29.51	49.80
		+ Ours	**65.90**	**46.98**	**56.14**	**63.00**			+ Ours	**41.94**	**38.05**	**32.51**	**51.70**
	7B	Baseline	68.20	53.76	55.73	68.00		7B	Baseline	58.53	49.97	46.65	73.80
		+ DTS	61.75	51.08	49.54	65.50			+ DTS	56.68	49.18	46.85	**74.00**
		+ Ours	**69.59**	**54.29**	**56.66**	**68.60**			+ Ours	**60.83**	**52.45**	**47.68**	**74.00**

Contribution Redirection. After constructing image-specific visual evidence, VER performs contribution-level redirection in selected heads, transferring delimiter-token output contributions to the corresponding visual evidence contributions. The redirection weight is defined as:

$$\lambda_i^{\ell,h}(t) = \frac{m_i^{\ell,h}(t)}{m_i^{\ell,h}(t) + b_i^{\ell,h}(t) + \epsilon}, \tag{15}$$

where $m_i^{\ell,h}(t) = \sum_{d \in \mathcal{D}_i(t)} A_{t,d}^{\ell,h}$ is the text-delimiter attention mass. $b_i^{\ell,h}(t) = \sum_{v \in V_i} A_{t,v}^{\ell,h}$ is the text-visual attention mass. When delimiter sink becomes more severe, $\lambda_i^{\ell,h}(t)$ becomes larger, leading VER to apply stronger redirection from delimiter-token contributions toward visual evidence. The original delimiter contribution is $O_{\mathcal{D}_i}^{\ell,h}(t) = \sum_{d \in \mathcal{D}_i(t)} A_{t,d}^{\ell,h} \mathbf{z}_d^{\ell,h}$. VER redirects this contribution toward visual evidence:

$$O'^{\ell,h}(t) = O^{\ell,h}(t) + \chi_{\ell,h} \sum_{i=1}^{N} \lambda_i^{\ell,h}(t) \Big(m_i^{\ell,h}(t) E_i^{\ell,h}(t) - O_{\mathcal{D}_i}^{\ell,h}(t) \Big), \tag{16}$$

where $O^{\ell,h}(t)$ and $O'^{\ell,h}(t)$ denote the original and updated head outputs for text token t.

5 Hallucination Mitigation Evaluation

5.1 Evaluation Setup

Evaluation Benchmarks. We evaluate DAR on four multi-image benchmarks. Mantis (Jiang et al., 2024) evaluates general multi-image instruction-following and reasoning abilities, including comparison, co-reference, and temporal understanding. BLINK (Fu et al., 2024) focuses on core visual perception abilities, with tasks involving fine-grained perception and multi-view reasoning. MIRB (Zhao et al., 2024) evaluates relational multi-image understanding, covering perception, visual world knowledge, reasoning, and multi-hop reasoning across images. QBench2 (Zhang et al., 2024) assesses low-level visual perception and quality-oriented understanding in multi-image settings. We further use VidHalluc (Li et al., 2025b) to evaluate the effectiveness on video hallucination.

Implementation Details. We adopt four representative MLLMs as backbones, including Qwen3.5 (Team, 2026), Qwen2.5-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), and LLaVA-OneVision (Li et al., 2025a). We use accuracy as the evaluation metric for multi-image benchmarks and follow the official evaluation metrics for VidHalluc.

5.2 Evaluation Results

Mitigation on Multi-Image Hallucination. Table 1 reports the main results across four multi-image benchmarks, and Figure 3 presents the case study. DAR improves the baseline in 31 out of 32 settings and matches the baseline in the remaining setting, demonstrating consistent effectiveness

Table 2: Ablation study of DAR.

Method	DOC	VER	Mantis	BLINK
Baseline	✗	✗	63.59	46.03
V_1	✓	✗	64.98	46.55
V_2	✗	✓	64.52	46.24
V_3 **(Ours)**	✓	✓	**65.90**	**46.98**

Table 3: Results of Qwen3.5 on VidHalluc.

Method	TSH	STH			ACH		
	Acc. ↑	Cls. ↑	Desc. ↑	MCC ↑	Avg. ↑	BQA ↑	MCQ ↑
Baseline	63.83	95.95	11.86	95.91	62.31	70.45	92.32
+ **Ours**	**66.33**	**96.39**	**13.04**	**96.35**	**63.05**	**71.02**	**92.45**

across different MLLMs and evaluation scenarios. DAR achieves notable gains across benchmarks, with improvements of up to 3.69 points on Mantis, 2.48 points on BLINK, 2.88 points on MIRB, and 5.40 points on QBench2. The results indicate that DAR is beneficial not only for cross-image reasoning but also for broader multi-image understanding tasks involving fine-grained perception. Moreover, as a training-free method, DAR requires no additional training and remains computationally efficient. Detailed computational overhead analyses are provided in ***Appendix*** C.

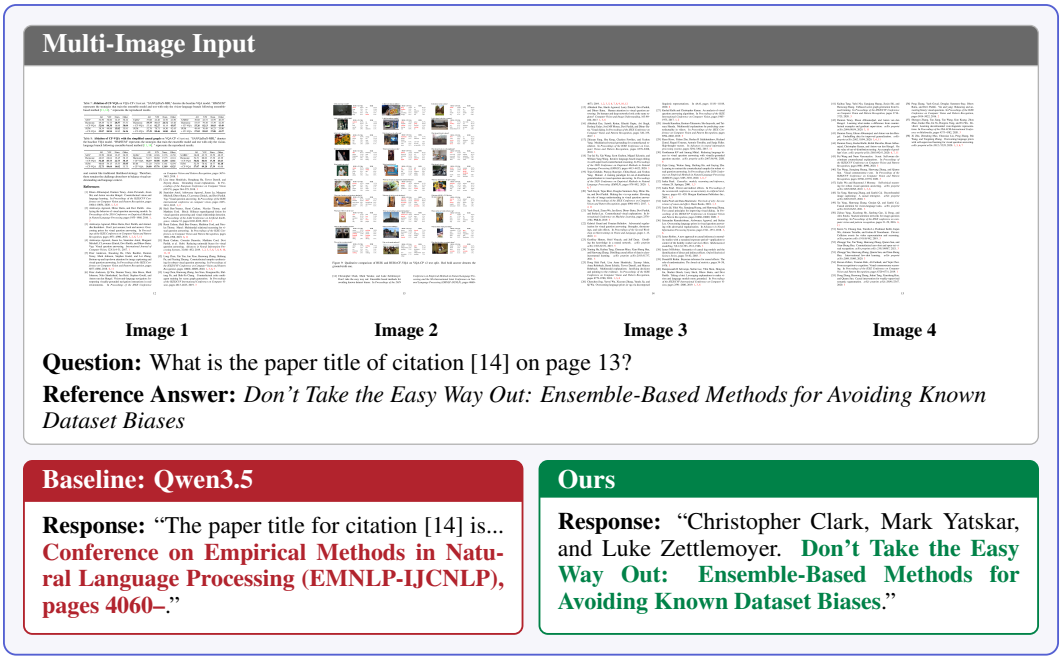


Figure 3: Qualitative comparison on MIRB. Our method correctly identifies the cited paper title from the multi-page input. More case studies are provided in ***Appendix*** F.

Extension to Video Hallucination Mitigation. To further assess the applicability of DAR beyond multi-image inputs, we evaluate it on the video hallucination benchmark VidHalluc using Qwen3.5-2B. In the video input, the sequence contains multiple timestamped visual segments, where timestamp tokens such as `<0.5 seconds>` and `<1.0 seconds>` are used as delimiter tokens for different temporal segments. As shown in Table 3, DAR consistently improves performance across temporal sequence (TSH), scene transition (STH), and action hallucination (ACH) evaluations, demonstrating its effectiveness in mitigating video hallucination.

5.3 Further Analysis of Proposed Method

Ablation Study. We first study the contribution of DOC and VER. To isolate their effects, we compare the baseline model, DOC only, VER only, and the full DAR framework. As shown in Table 2, both DOC and VER improve over the baseline, indicating that visual ownership correction and text-side evidence redirection are individually effective. DOC mainly improves the reliability of image-specific visual evidence by reducing delimiter-level ownership ambiguity. VER improves text-side grounding by steering delimiter-based routing toward visual tokens. The full DAR frame-

Table 4: Results of different head selection strategies in VER. Random heads are selected with the same selection ratio as the TDSS-based strategy.

Strategy	Mantis	BLINK	MIRB
All heads	62.21	46.55	53.87
Random heads	61.75	46.40	53.56
TDSS-based head selection	**65.90**	**46.98**	**56.14**

Table 5: Analysis of the ownership prior. The variant without $\omega_i(v)$ constructs visual evidence using only image-local visual relevance.

Variant	Mantis	BLINK	MIRB
w/o ownership prior	61.29	46.55	53.15
w/ ownership prior	**65.90**	**46.98**	**56.14**

work achieves the best overall performance, suggesting that the two modules are complementary rather than redundant.

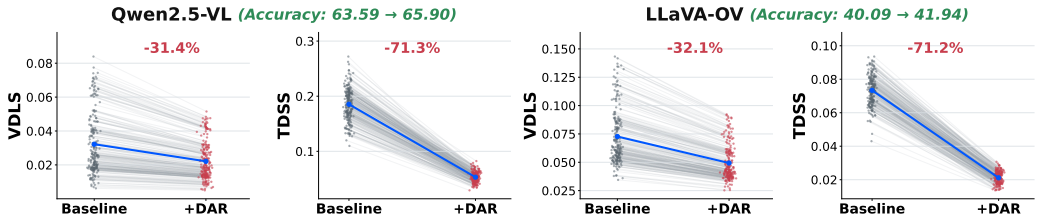


Figure 4: Comparison of VDLS and TDSS before and after applying DAR on Mantis. Each gray/red dot denotes an individual sample, and blue dots indicate the mean.

Reduction of Visual-Delimiter Leakage and Text-Delimiter Sink. We further compare VDLS and TDSS before and after applying DAR. As shown in Figure 4, VDLS decreases by 31.4% on Qwen2.5-VL and 32.1% on LLaVA-OneVision, while TDSS decreases by 71.3% and 71.2%, respectively. The paired sample-level trajectories further show that most samples move toward lower VDLS and TDSS after applying DAR. These results confirm that DAR modifies the targeted internal attention patterns in the intended direction, substantially reducing both VDLS and TDSS.

Effectiveness of TDSS-based Head Selection. Table 4 evaluates different head selection strategies. Applying VER to all heads does not consistently improve performance, achieving 62.21, 46.55, and 53.87 on Mantis, BLINK, and MIRB, respectively, because it also perturbs delimiter-irrelevant heads. Random head selection also yields limited gains. In contrast, the TDSS-based strategy achieves the best results across all three benchmarks. These results show that TDSS can effectively identify delimiter-sink heads, validating the necessity of TDSS-based intervention.

Effectiveness of Ownership Prior. We further verify the effectiveness of the ownership prior $\omega_i(v)$ in VER. As shown in Table 5, incorporating $\omega_i(v)$ improves performance across benchmarks. This suggests that text-to-visual relevance alone is insufficient for reliable visual evidence construction. A visual token may appear relevant to the text token, but it can still be unreliable if its image ownership is unstable. The corrected ownership prior helps VER select visual tokens that are both relevant and ownership-consistent.

6 Conclusion

In this paper, we provide a finer-grained analysis of delimiter-mediated interactions in multi-image MLLMs by distinguishing two complementary internal patterns: *Visual-Delimiter Leakage* on the visual side and *Text-Delimiter Sink* on the textual side. To systematically analyze these phenomena, we introduce VDLS and TDSS to quantify delimiter-associated patterns on the visual and textual sides. Based on these diagnostics, we propose DAR, with two complementary modules: DOC and VER. Extensive experiments on multi-image benchmarks demonstrate the effectiveness and generalizability of DAR. Moreover, experiments on video hallucination suggest that DAR can also benefit broader multi-input scenarios. Limitations and future work are discussed in ***Appendix*** B.

AI USAGE STATEMENT

In this work, generative AI tools were used to assist with language editing, including improving grammar, sentence structure, clarity, and readability, as well as limited coding and visualization support. Generative AI was not used to formulate the core research ideas, design the proposed method, conduct experimental analysis, or derive scientific conclusions. All AI-assisted content was carefully reviewed and verified by the authors. The authors take full responsibility for the accuracy, validity, and integrity of all content presented in this work.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The study does not involve human subjects, animal experiments, or the collection of personally identifiable or sensitive information. All datasets, pretrained multimodal large language models, and evaluation benchmarks used in our experiments are publicly available and were used in accordance with their respective licenses and terms of use. Our study focuses on analyzing and mitigating hallucinations in multimodal large language models and does not introduce additional data collection involving human participants.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed method, experimental setup, evaluation protocols, and implementation details throughout the paper and appendix. The model configurations, intervention settings, evaluation benchmarks, and major hyperparameters used in our experiments are reported accordingly. All pretrained multimodal large language models and benchmarks evaluated in this work are publicly available. Additional implementation details are provided in the supplementary material.

REFERENCES

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

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

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26540–26550, 2024a.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4209–4221, 2025.

Xi Chen, Mingkang Zhu, Shaoteng Liu, Xiaoyang Wu, Xiaogang Xu, Yu Liu, Xiang Bai, and Hengshuang Zhao. Mico: Multi-image contrast for reinforcement visual reasoning. *Advances in Neural Information Processing Systems*, 38:10786–10805, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv Preprint arXiv:2412.05271*, 2024b.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3488–3501, 2025.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *13th International Conference on Learning Representations, ICLR 2025*, pp. 13547–13574. International Conference on Learning Representations, ICLR, 2025.

Minyoung Lee, Yeji Park, Dongjun Hwang, Yejin Kim, Seong Joon Oh, and Junsuk Choe. Enhancing multi-image understanding through delimiter token scaling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7QFf05KrOm.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025b.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv Preprint arXiv:2407.07895*, 2024.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 34583–34610, 2025b.

Joonki Min, Chaeyun Kim, Hyungwook Choi, Yejin Kim, Kihyun Kim, Yohan Jo, and Joonseok Lee. Fine-grained multi image object hallucination benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18295–18305, 2026.

Yeji Park, Minyoung Lee, Sanghyuk Chun, and Junsuk Choe. Mitigating cross-image information leakage in lvlms for multi-image tasks. *arXiv Preprint arXiv:2508.13744*, 2025.

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, et al. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *arXiv Preprint arXiv:2604.10098*, 2026.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10599–10609, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv Preprint arXiv:2409.12191*, 2024.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in large vision language models. *arXiv Preprint arXiv:2405.17821*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Xiaochen Yang, Hao Fang, Jiawei Kong, Yaoxin Mao, Bin Chen, and Shu-Tao Xia. Looking back and forth: Cross-image attention calibration and attentive preference learning for multi-image hallucination mitigation. *arXiv Preprint arXiv:2603.07048*, 2026.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. *arXiv Preprint arXiv:2602.21704*, 2026a.

Jianghao Yin, Qingbin Li, Kun Sun, Cheng Ding, Jie Wang, Qin Chen, Jie Zhou, Nan Wang, Changqing Li, Pei Wu, et al. Mimic human cognition, master multi-image reasoning: A meta-action framework for enhanced visual understanding. *arXiv Preprint arXiv:2601.07298*, 2026b.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19985–19995, 2025.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-bench $^{+}$+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10404–10418, 2024.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv Preprint arXiv:2406.12742*, 2024.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

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

A Qualitative Analysis

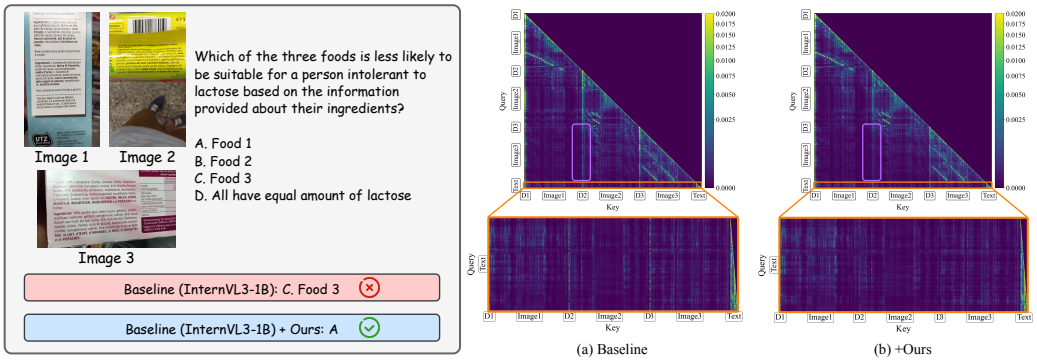


Figure 5: Qualitative results and attention visualization before and after applying DAR. DAR corrects the baseline answer and alleviates both visual-delimiter leakage and text-delimiter sink by reducing attention to non-own delimiters and strengthening text-side grounding in image-specific visual evidence.

Figure 5 presents a qualitative example illustrating how DAR modifies internal attention patterns and mitigates multi-image hallucination. The model is asked to identify which food is less suitable for a person with lactose intolerance based on three ingredient images. While the correct answer is Food 1, the baseline InternVL3 model incorrectly predicts Food 3.

The attention visualization reveals two delimiter-associated internal patterns in the baseline. First, visual tokens from Image 3 allocate substantial attention to the delimiter of Image 1. This attention pattern is consistent with ambiguous delimiter-based ownership between Images 1 and 3 and coincides with the incorrect prediction of Food 3. Second, the zoomed text-to-image attention region shows that text tokens concentrate excessive attention on delimiter positions, indicating a clear text-delimiter sink and insufficient grounding in the actual visual evidence.

After applying DAR, the attention assigned by Image 3 tokens to the delimiter of Image 1 is substantially reduced. This suggests that DOC restores the ownership of visual tokens by re-anchoring them to their source-image delimiters, thereby suppressing cross-image leakage. Meanwhile, VER redirects delimiter-dominated text attention toward image-specific visual tokens, alleviating the text-delimiter sink and strengthening visual evidence grounding. Consequently, DAR corrects the prediction from Food 3 to the correct answer, Food 1. This example demonstrates that DAR mitigates multi-image hallucination by jointly correcting visual-delimiter leakage and text-delimiter sink.

Table 6: Computational overhead of DAR.

Benchmark	Method	Accuracy (%)	Peak GPU Memory (GB)	Latency (s/sample)	Throughput (tokens/s)
Mantis	Baseline	61.29	11.194	1.083	2.617
	+DAR	63.59	11.194	1.292	2.198
BLINK	Baseline	50.83	11.313	0.880	6.521
	+DAR	52.16	11.314	1.033	5.554

B Limitations and Future Work

Although DAR shows consistent effectiveness across multiple MLLMs and benchmarks, several aspects remain worth further exploration. First, our experiments mainly cover representative multi-image benchmarks and a video hallucination benchmark. Future work may evaluate DAR on more diverse multi-input scenarios, such as long-video understanding, interleaved image-text documents, and multimodal agent trajectories, where the input structure can be more complex.

Second, DAR currently focuses on delimiter-related attention patterns, including visual-delimiter leakage and text-delimiter sink. While these patterns provide clear and actionable diagnostic signals,

future work may further combine them with other internal signals, such as hidden-state dynamics or output-logit behaviors, to obtain a more comprehensive understanding of multi-image hallucination.

C Computational Overhead

Table 6 reports the computational overhead of DAR on Qwen3.5-2B using GeForce RTX 3090 GPUs. On Mantis, DAR improves accuracy from 61.29% to 63.59%, yielding an absolute gain of 2.30 percentage points, while the peak GPU memory remains unchanged at 11.194 GB. The average end-to-end latency increases moderately from 1.083 to 1.292 seconds per sample, and the generation throughput changes from 2.617 to 2.198 tokens per second. On BLINK, DAR improves accuracy from 50.83% to 52.16%, corresponding to an absolute gain of 1.33 percentage points. Meanwhile, the peak GPU memory remains nearly identical, increasing only from 11.313 to 11.314 GB. The latency increases from 0.880 to 1.033 seconds per sample, while the throughput changes from 6.521 to 5.554 tokens per second.

The main advantage of DAR is that it provides consistent accuracy improvements with virtually no additional peak GPU memory. This favorable memory efficiency stems from its training-free design. Although its inference-time operations introduce a modest increase in latency and a corresponding decrease in throughput, the overall computational cost remains limited. These results demonstrate that DAR offers an effective and efficient method for mitigating multi-image hallucination, achieving improved performance without increasing the model size or requiring additional training.

D Pseudocode of DAR

Algorithm 1 summarizes the inference-time procedure of DAR. For each target layer, DOC first corrects visual-token ownership by reallocating attention from non-own delimiters to the corresponding image delimiters, and then derives an ownership prior. VER subsequently computes TDSS to identify delimiter-sink heads and redirects their delimiter-token contributions toward ownership-conditioned, image-specific visual evidence. DAR operates entirely during inference without introducing additional trainable parameters or modifying the original decoding process.

E More Implementation Details

We follow the official evaluation protocol of VidHalluc. TSH is evaluated using accuracy, while ACH is evaluated using the accuracy of binary question answering (BQA) and multiple-choice questions (MCQ). For STH, we report the classification score (Cls.), description score (Desc.), Matthews correlation coefficient (MCC). The classification score is derived from the normalized MCC, while the description score measures the semantic similarity between the predicted and reference scene descriptions. For benchmarks without explicit hallucination annotations, we use accuracy as a task-level proxy for hallucination-related failures, where a lower error rate indicates fewer unsupported or incorrect multi-image predictions.

F Case Study

We present some representative cases from MANTIS, BLINK, MIRB, and QBench2 using Qwen3.5-2B. The selected examples cover cross-image counting, shared-state recognition, point correspondence, visual analogy, and image-quality comparison. Incorrect decisions or missing final answers are highlighted in red, while the corrected outputs are highlighted in green. We further present three representative cases from VidHalluc, covering open-ended scene transition hallucination (STH), temporal sequence hallucination (TSH), and action hallucination (ACH). Each video is represented by eight uniformly sampled keyframes arranged in temporal order from left to right and top to bottom. These cases qualitatively show that our method more reliably preserves the association between each visual observation and its source image or temporal segment and then aggregates the relevant evidence across images and temporal segments.

Algorithm 1: Delimiter-Aware Regrounding (DAR)

Input: Visual token sets $\{V_i\}_{i=1}^{N}$; delimiter sets $\mathcal{D}(x)$ and $\mathcal{D}_i(x)$; text token set $\mathcal{T}$; target layers; attention weights, attention logits, and value states
Output: Updated head outputs
foreach *target layer* ℓ **do**
  foreach *attention head* h **do**
    `// Stage 1: Delimiter Ownership Correction`
    foreach *image* i *and visual token* $v \in V_i$ **do**
      Compute token-level ownership statistics and correction strength using Eqs. (1) and (6)
      Compute redistribution weights and update delimiter attention using Eqs. (7)–(8)
      Derive the corrected ownership prior using Eq. (9)
    `// Stage 2: Visual Evidence Redirection`
    Compute head-level TDSS using Eqs. (3)–(5)
  Compute the layer-wise average TDSS using Eq. (10)
  foreach *attention head* h **do**
    Select delimiter-sink heads using Eq. (11)
    if h *is selected* **then**
      foreach *text token* $t \in \mathcal{T}$ *and image* i **do**
        Compute image-local visual relevance and the ownership-conditioned evidence posterior using Eqs. (12)–(13)
        Construct image-specific visual evidence using Eq. (14)
        Compute delimiter and visual attention masses and the redirection weight using Eq. (15)
        Redirect delimiter contribution toward image-specific visual evidence using Eqs. (16)
return $O'^{\ell,h}(t)$;

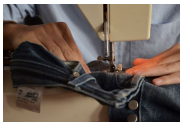
Case 1: MIRB – Visual Analogy

Multi-Image Input


Source 1


Source 2


Query


A


B


C


D

Question: Given the source pair, which candidate best completes the analogous relation for the query image?
Output Format: Answer with a single word. Do not explain.
Reference Answer: *A*

Baseline: Qwen3.5

Response: B

Ours

Response: A

Case 2: MANTIS – Shared Motion State

Multi-Image Input



Image 1



Image 2

Question: Are the birds in motion or static in both images?
Output Format: Answer with only the final answer. Do not explain.
Reference Answer: *motion*

Baseline: Qwen3.5

Response: "In the first image, the seagull... is actively flying... In the second image, the bird... is also shown with..." **[truncated; no final answer]**

Ours

Response: **motion**

Case 3: MANTIS – Cross-Image Counting

Multi-Image Input



Image 1



Image 2

Question: How many more bowls are there in Image 2 than in Image 1?
Output Format: Answer with only the final answer. Do not explain.
Reference Answer: *4*

Baseline: Qwen3.5

Response: "Let's analyze the images... Image 1 shows a stack of four bowls... Image..." **[truncated; no final answer]**

Ours

Response: **4**

Case 4: BLINK – Functional Correspondence

Multi-Image Input



Reference



Candidates

Question: For the action "Brush/Dust," which labeled point in the second image functionally corresponds to the reference point in the first image?
Choices: (A), (B), (C), or (D)
Output Format: Answer with exactly one option from the choices. Use only the format (A), (B), (C), (D), or (E). Do not explain.
Reference Answer: *(C)*

Baseline: Qwen3.5

Response: **(A)**

Ours

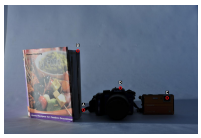
Response: **(C)**

Case 5: BLINK – Visual Correspondence

Multi-Image Input



Reference View



Changed View

Question: Which labeled point in the second image corresponds to the reference point after the camera position or lighting changes?
Choices: (A), (B), (C), or (D)
Output Format: Answer with exactly one option from the choices. Use only the format (A), (B), (C), (D), or (E). Do not explain.
Reference Answer: *(D)*

Baseline: Qwen3.5

Response: (C)

Ours

Response: (D)

Case 6: QBench2 – Motion-Blur Comparison

Multi-Image Input



Image 1



Image 2

Question: Which region is most severely affected by motion blur?
A. The table in Image 1 **B.** The racer in Image 2
C. The background in Image 1 **D.** The background trees in Image 2
Output Format: Answer with the option's letter from the given choices directly.
Reference Answer: *D*

Baseline: Qwen3.5

Response: C

Ours

Response: D

Case 7: QBench2 – Color-Richness Comparison

Multi-Image Input



Image 1

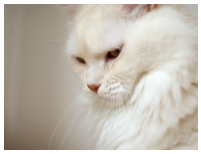


Image 2

Question: Compared with Image 1, how is the color richness of Image 2?
A. richer **B.** similar **C.** less rich
Output Format: Answer with the option's letter from the given choices directly.
Reference Answer: *C*

Baseline: Qwen3.5

Response: B

Ours

Response: C

Case 8: VidHalluc-ACH – Action Hallucination

Video Input



Uniformly sampled frames in temporal order

Question: Is the primary action in the video herding sheep?
Output Format: Answer with only Yes or No. Do not provide any explanation.
Reference Answer: *Yes*

Baseline: Qwen3.5

Response: No

Ours

Response: Yes

Case 9: VidHalluc-STH – Open-Ended Scene Change

Video Input



Uniformly sampled frames in temporal order

Question: A scene change is defined as a significant transition in the overall environment or location within the video. Watch the video and determine whether a scene change occurs.
Output Format: If a scene change occurs, respond with "Scene change: Yes, Locations: from [location] to [location2]." Otherwise, respond with "Scene change: No, Locations: None."
Reference Answer: *Scene change: Yes, Locations: from in a room to in the street.*

Baseline: Qwen3.5

Response: Scene change: Yes, Locations: **from a fan to a fence**

Ours

Response: Scene change: Yes, Locations: **from indoor to outdoor**

Case 10: VidHalluc-TSH – Temporal Action Ordering

Video Input



Uniformly sampled frames in temporal order

Question: Below are two actions in the video: Action A, rinsing the car; and Action B, brushing rims. Sort the two actions in the order in which they occur.

Output Format: Return the detected action or actions in temporal order. If only one action is detected, return only that action.

Reference Answer: *Action B. brushing rims; Action A. rinsing the car*

Baseline: Qwen3.5

Response: **Action A. rinsing the car; Action B. brushing rims**

Ours

Response: **Action B. brushing rims; Action A. rinsing the car**