ATTENTION AT REST STAYS AT REST: BREAKING VISUAL INERTIA TO MITIGATE RELATION HALLUCINATIONS

Anonymous authors
Paper under double-blind review

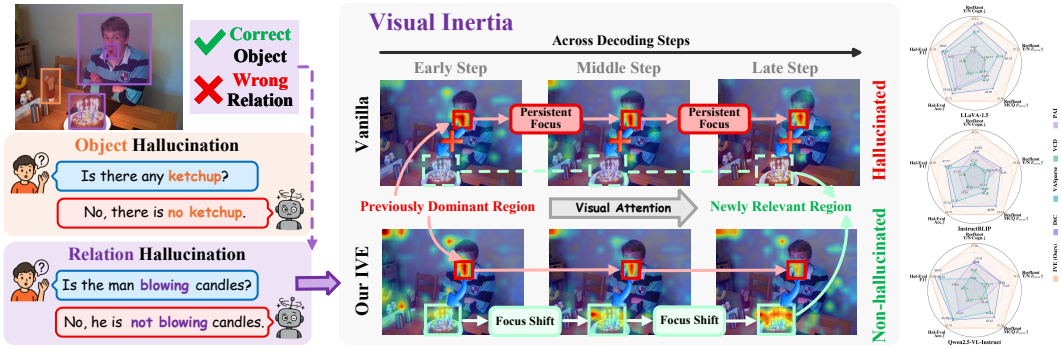


Figure 1: **Overview of Visual Inertia in relation hallucinations. Left:** Object hallucination concerns entity presence, whereas relation hallucination can occur despite correct object recognition. **Middle:** Visual attention remains focused on previously dominant regions across decoding rather than shifting toward newly relevant evidence, illustrating visual inertia. **Right:** Radar comparison of IVE and competing methods across three MLLM backbones on Reefknot and Hal-Eval.

ABSTRACT

While multimodal large language models demonstrate strong entity-level perception, faithfully grounding relational interactions between objects remains a persistent challenge. Although conventional visual grounding techniques attempt to resolve hallucinations by amplifying visual attention, strengthening overall visual signals fails to reliably correct relational errors. Tracing visual attention in relation descriptions reveals that correct responses tend to dynamically shift focus across regions, whereas hallucinated responses often linger on previously dominant evidence, exhibiting an undesirable *visual inertia*. Further analysis shows that relation-prediction performance steadily deteriorates as more previous-step visual attention is carried into the current decoding step. We therefore introduce Inertia-aware Visual Excitation (IVE), an MLLM decoding method that dynamically recalibrates visual values using token-level attention history. By contrasting current attention against recent moving averages, IVE separates emergent tokens with rising relevance from persistently dominant inertia tokens, selectively reinforcing newly needed evidence while mildly attenuating contributions from repeatedly attended regions. Across three MLLMs and decoding strategies, IVE reduces relation hallucinations while preserving broader multimodal performance.

1 INTRODUCTION

Multimodal large language models (MLLMs) Liu et al. (2024b;a); Li et al. (2023a); Zhu et al. (2024); Bai et al. (2025) combine visual and textual information to generate natural-language responses. Object-level grounding requires the entities mentioned in a response to be present in the image. Reliable relation prediction goes further, requiring the model to combine evidence from the recognized entities and determine how they are related. For example, recognizing the man, cake, and candles does not by itself determine whether the man is blowing out the candles (Figure 1, left). An MLLM may nevertheless predict such an image-unsupported relation, producing a *relation hallucination* Zheng et al. (2025); Jiang et al. (2024); Wu et al. (2025); Qin et al. (2026); Wu et al. (2026).

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.

A common strategy for improving visual grounding is to strengthen or recalibrate visual signals during decoding Huang et al. (2024); Yin et al. (2025); Liu et al. (2024c); Yu et al. (2025); Zhuang et al. (2025). This is an intuitive remedy when a model overlooks an object. Relation prediction, however, requires evidence from multiple regions and may not improve simply because the image receives more attention. Under the same visual-attention amplification, performance improves steadily on the object-level POPE benchmark Li et al. (2023b) but changes only modestly and non-monotonically on the relation-focused Reefknot benchmark Zheng et al. (2025) (Section 2). This contrast motivates examining how visual evidence is used across decoding, beyond overall visual attention strength.

We therefore trace visual attention in relation descriptions derived from Reefknot, comparing how visual focus evolves in correct and hallucinated predictions (Section 2). Regions that become dominant early in hallucinated predictions often remain dominant at middle and later output positions, whereas correct predictions show larger shifts in the emphasized regions. This contrast appears in individual examples and in group-level comparisons of attention displacement across output positions. The persistent reuse of previously dominant visual evidence reflects a form of *visual inertia* (Figure 1, middle). To test how stronger persistence affects relation prediction, we progressively carry more previous-step visual attention into the current step. Performance steadily deteriorates as carryover increases, indicating that stronger temporal persistence can impair relation prediction.

Guided by this temporal pattern, we introduce Inertia-aware Visual Excitation (IVE), which couples reinforcement of newly relevant evidence with conditional attenuation of persistently dominant contributions. Recent attention history distinguishes emergent tokens with rising attention from inertia tokens that remain strongly attended without a sufficient relative rise. For inertia tokens, attenuation reflects how frequently they have previously occupied this role. These roles are reassessed as generation proceeds, allowing a previously dominant token to be reinforced when its attention rises again relative to history. The method recalibrates visual values during decoding without additional training or parameter updates.

Across three MLLM backbones, IVE consistently improves relation prediction on Reefknot Zheng et al. (2025) and Hal-Eval Jiang et al. (2024). Results on MMRA Wu et al. (2026) extend this benefit to a multi-image relational setting. IVE also remains effective across decoding strategies while preserving performance on broader multimodal benchmarks.

2 From Attention Strength to Temporal Persistence

Many hallucination-mitigation methods strengthen visual attention to improve grounding Yin et al. (2025); Liu et al. (2024c); Yu et al. (2025). For relation prediction, however, the image must provide evidence not only for the entities mentioned in a response but also for how they are related. A model may attend strongly to the relevant entities while continuing to emphasize evidence that supports their recognition rather than the predicted relation.

2.1 More Visual Attention Is Not Enough

To examine whether stronger visual attention is sufficient to improve relation prediction, we vary the amplification factor on LLaVA-1.5 Liu et al. (2024a) following PAI Liu et al. (2024c). We apply the same settings to POPE Li et al. (2023b), which evaluates object hallucination, and Reefknot Zheng et al. (2025), which evaluates relation hallucination.

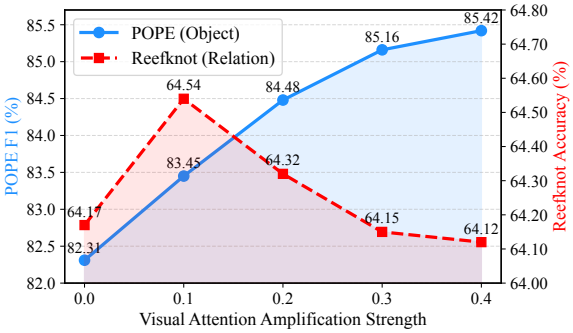


Figure 2: **Visual-attention amplification.** POPE improves steadily, while Reefknot gains only under mild amplification before returning to its baseline range.

POPE F1 improves steadily by 3.11 points, whereas Reefknot gains only under mild amplification before returning to its baseline range (Figure 2). Stronger visual attention thus benefits object-level evaluation without consistently improving relation prediction. This contrast motivates examining how visual attention evolves during generation, beyond its overall strength.

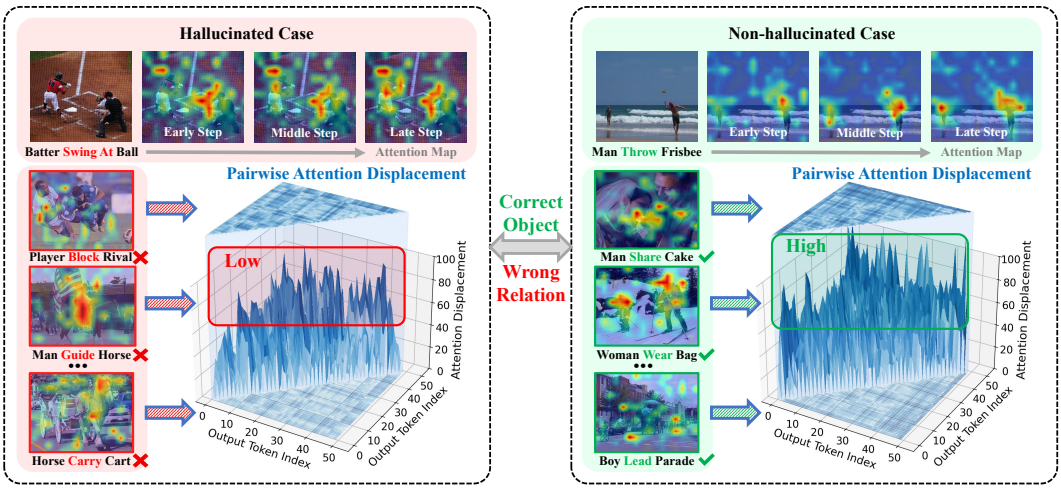


Figure 3: **Visual-attention persistence in Reefknot-derived relation descriptions.** Early, middle, and late attention maps illustrate hallucinated (left) and correct (right) descriptions. The surfaces show each group's mean pairwise index-based attention displacement. Lower values indicate more similar visual-attention distributions across output positions, with the hallucinated group showing more persistent visual focus.

2.2 Visual Evidence Remains Persistent Across Decoding

We analyze visual-attention trajectories in relation descriptions derived from Reefknot, using the generation and annotation protocol in Appendix A.2. Across early, middle, and late attention maps (Figure 3), attention shifts between regions in the correct example but stays on similar regions in the hallucinated example. This continued reliance on previously dominant visual evidence illustrates visual inertia across decoding.

To characterize persistence beyond these snapshots, we analyze 1,000 descriptions under the same Vanilla configuration, with 500 correct, non-hallucinated descriptions and 500 hallucinated descriptions. For each response, we compare attention distributions across all pairs of analyzed output positions, capturing whether visual focus remains similar over both short and long intervals.

To compare visual-token allocation, we aggregate over N_h heads and normalize the visual slice. With $\mathbf{A}_{t,\mathcal{V}}^{(h)}$ denoting attention from query t to visual tokens $\mathcal{V}$ in head h of the analyzed layer, we obtain

$$\mathbf{p}_t = \frac{\sum_{h=1}^{N_h} \mathbf{A}_{t,\mathcal{V}}^{(h)}}{\left\| \sum_{h=1}^{N_h} \mathbf{A}_{t,\mathcal{V}}^{(h)} \right\|_1}. \tag{1}$$

For prediction n, the resulting pairwise displacement matrix is

$$\mathbf{M}_{i,j}^{(n)} = W_{\text{idx}}\left(\mathbf{p}_i^{(n)}, \mathbf{p}_j^{(n)}\right), \qquad 1 \le i, j \le 50, \tag{2}$$

where W_{idx} is the one-dimensional Wasserstein distance on row-major patch indices a, b with cost $|a - b|$, measuring displacement in index units. Larger values indicate greater attention change, whereas extended low-value regions indicate similar distributions across output positions. Averaging per-prediction matrices elementwise within each group yields the surfaces in Figure 3.

Correct descriptions exhibit larger displacement across output-position pairs, whereas hallucinated descriptions retain more similar indexed attention distributions. In these descriptions, relation errors are associated with more persistent visual focus. A prediction-level analysis in Appendix A.2 likewise finds higher relation accuracy among trajectories with larger consecutive attention displacement.

2.3 Increasing Temporal Persistence Degrades Relation Prediction

Hallucinated descriptions in the preceding analysis tend to retain similar visual focus across decoding positions. We examine how relation prediction changes when previous-step attention is explicitly

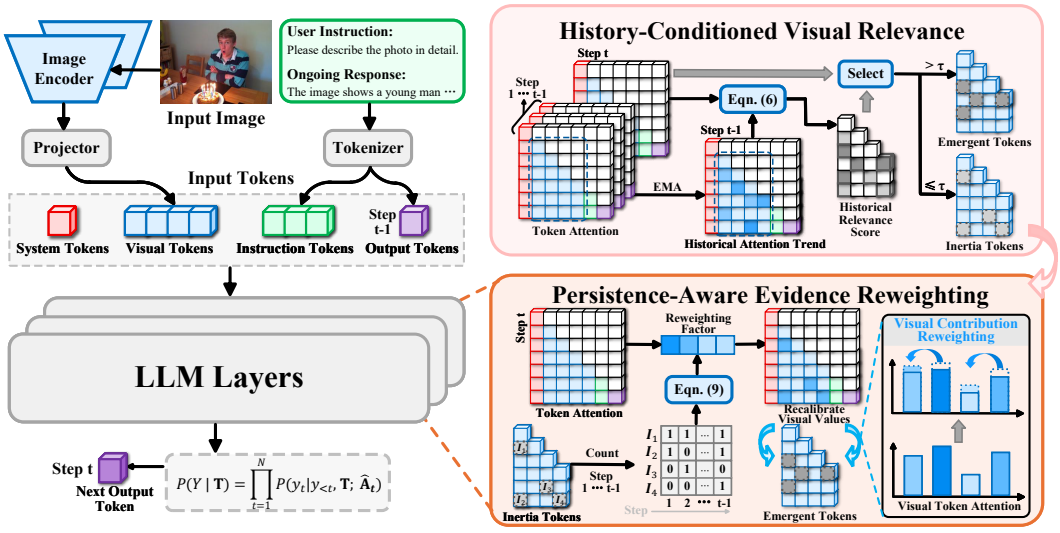


Figure 4: **Overview of IVE.** The method compares current visual attention with recent history to identify emergent and inertia tokens. Recalibration relative to their original visual values reinforces newly relevant evidence and mildly attenuates persistently dominant contributions during decoding.

carried forward. On LLaVA-1.5 and Reefknot, we progressively mix previous-step attention into the current distribution, with λ controlling the previous step's contribution:

$$\mathbf{p}_t^{\text{carry}}(\lambda) = \text{Norm}(\mathbf{p}_t + \lambda \mathbf{p}_{t-1}) \,. \tag{3}$$

Carrying forward previous attention can also change the distribution's concentration. We therefore pair each setting with an entropy-matched reference built solely from current-step attention, comparing carryover and current-step reshaping at matched concentration (Appendix A.4).

As carryover strengthens, R_{score} steadily declines, with a larger drop than under the entropy-matched reference at every nonzero strength (Table 1). The performance gap widens as carryover increases, accompanied by a sharper reduction in consecutive attention displacement.

Table 1: **Progressive carryover on Reefknot Y/N.** The reference matches carryover entropy using current-step attention only. D_{attn} measures consecutive displacement.

	Carryover		Matched reference	
λ	D_{attn}	R_{score} ↑	D_{attn}	R_{score} ↑
0.00	41.52	64.39	41.52	64.39
0.25	35.41	59.61	38.13	62.27
0.50	28.73	54.50	34.62	59.74
0.75	23.64	50.71	31.17	57.55
1.00	19.24	47.68	28.84	55.86

3 Inertia-aware Visual Excitation

The persistence observed in relation descriptions motivates calibrating visual contributions using attention history rather than current strength alone (Section 2). Inertia-aware Visual Excitation (IVE) combines token-specific historical references with emergent-token reinforcement and conditional inertia-token attenuation. Figure 4 outlines this training-free method.

3.1 History-Conditioned Visual Relevance

Consider two visual tokens that receive the same attention at the current decoding step. One may have remained at this level for several steps, while the other may have only recently risen from a weak historical level. A current-only signal treats them identically, although the latter reflects a change in visual focus. IVE preserves this distinction by evaluating each visual token relative to its own history.

To track recent changes in each token's attention, we maintain an exponential moving average as a short-term reference. At selection layer l_s, we read the attention vector $\mathbf{A}_t^{(h)} \in \mathbb{R}^N$ over all N tokens before value recalibration. Averaging the N_h heads gives $a_{t,j}$ for visual token $j \in \mathcal{V}$, with its

historical reference updated as

$$a_{t,j} = \frac{1}{N_h} \sum_{h=1}^{N_h} \mathbf{A}_{t,j}^{(h)}, \tag{4}$$

$$\widetilde{a}_{t-1,j} = \gamma \widetilde{a}_{t-2,j} + (1-\gamma) a_{t-1,j}. \tag{5}$$

The reference blends the previous observation and the older moving average with weights $1-\gamma$ and $\gamma \in (0,1)$, respectively. Here, t indexes output-predicting queries, with $t=1$ at the final prompt position. The recurrence applies for $t>1$, starting from $\widetilde{a}_{0,j}$.

Raw attention changes can differ in scale across visual tokens. We therefore compare the current signal with its historical reference using

$$S_{t,j} = \frac{a_{t,j} - \widetilde{a}_{t-1,j}}{\sqrt{\widetilde{a}_{t-1,j}(1-\widetilde{a}_{t-1,j})} + \epsilon}, \qquad \epsilon = 10^{-6}. \tag{6}$$

This normalized difference preserves the direction of change and accounts for historical attention scale. The signal reflects changes in both overall visual attention and its allocation across visual tokens. A large positive $S_{t,j}$ indicates increased prominence relative to the token's recent history.

Emergent tokens show a sufficient rise relative to history, while inertia tokens remain prominent with limited relative growth. To identify the latter, we combine the relevance score with current attention strength. Using the visual-token mean $\mu_t^v = N_v^{-1} \sum_{j \in \mathcal{V}} a_{t,j}$, where $N_v = |\mathcal{V}|$, we define

$$\begin{aligned} \mathcal{E}_t &= \{ j \in \mathcal{V} \mid S_{t,j} > \tau \}, \\ \mathcal{I}_t &= \{ j \in \mathcal{V} \mid a_{t,j} > \mu_t^v,\ S_{t,j} \le \tau \}, \end{aligned} \tag{7}$$

where τ is the relevance threshold. The above-mean condition focuses inertia selection on prominent contributions, distinguishing them from weakly attended tokens with similarly low scores.

3.2 Persistence-Aware Evidence Reweighting

The inertia set determines which visual contributions are eligible for attenuation. To distinguish repeated dominance from an occasional selection, IVE sets attenuation strength using prior inertia-selection frequency. With initial count C_j^{pre} and history length n_{pre}, we accumulate the history before output token t as

$$C_{t,j} = C_j^{\mathrm{pre}} + \sum_{k=1}^{t-1} \mathbf{1}[j \in \mathcal{I}_k], \qquad n_t = n_{\mathrm{pre}} + t - 1. \tag{8}$$

The ratio $C_{t,j}/n_t$ measures historical inertia-selection frequency, with higher values producing stronger attenuation for currently selected inertia tokens. The current query is excluded, and tokens without prior inertia selections receive no attenuation. The modulation factor before capping is

$$m_{t,j} = \begin{cases} 1 - \alpha \dfrac{C_{t,j}}{n_t}, & j \in \mathcal{I}_t,\ n_t > 0, \\ 1 + S_{t,j}, & j \in \mathcal{E}_t, \\ 1, & \text{otherwise}, \end{cases} \tag{9}$$

where $\alpha \in (0,1]$ controls attenuation, with inertia-token factors bounded by $[1-\alpha, 1]$. Emergent-token selection uses the original relevance score, while its multiplier is capped at 1.10, yielding fixed-strength reinforcement at the reported thresholds.

The token sets are reassessed at each step, so renewed attention can move an inertia token into the emergent set for reinforcement. These query-specific factors scale each token's contribution through its visual value. Let $\mathbf{V}_j^{(l,h),\mathrm{base}}$ denote the fixed original value at layer l and head h. Using the capped factor, we apply

$$\mathbf{V}_j^{(l,h)} \leftarrow m_{t,j} \mathbf{V}_j^{(l,h),\mathrm{base}}. \tag{10}$$

Tokens outside both sets regain their original values. Recalibration spans all layers and heads once per query, before attention–value aggregation at l_s. The adjusted values affect l_s and later layers immediately, and earlier layers on subsequent queries. Algorithm 1 details execution and history updates (Appendix B).

Table 2: Results (%) on Reefknot under Y/N and MCQ evaluation. Perc. and Cogn. denote the benchmark's perception- and cognition-hallucination rates.

		LLaVA-1.5			InstructBLIP			Qwen2.5-VL-Instruct		
Category	**Method**	**Perc.↓**	**Cogn.↓**	**R_{score} ↑**	**Perc.↓**	**Cogn.↓**	**R_{score} ↑**	**Perc.↓**	**Cogn.↓**	**R_{score} ↑**
Y/N	Vanilla	37.67	33.99	64.39	45.31	42.47	56.28	37.22	20.58	72.07
	Object-specific									
	+VCD Leng et al. (2024)	37.33	33.67	64.71	45.22	42.35	56.38	37.13	20.47	72.17
	General									
	+OPERA Huang et al. (2024)	37.08	33.23	65.07	45.03	42.05	56.63	36.86	20.12	72.49
	+PAI Liu et al. (2024c)	37.21	33.77	64.71	44.89	41.86	56.80	36.82	19.95	72.60
	+HELPD Yuan et al. (2024)	37.05	33.31	65.04	44.58	41.94	56.89	36.79	20.04	72.57
	+DeCo Wang et al. (2025)	37.22	33.42	64.90	45.17	42.21	56.48	37.00	20.33	72.31
	+VASparse Zhuang et al. (2025)	37.13	33.14	65.10	44.67	42.47	56.56	36.91	20.86	72.05
	+MemVR Zou et al. (2025)	36.99	32.81	65.34	44.34	42.09	56.92	36.75	20.16	72.52
	Relation-specific									
	+CAP Wu et al. (2025)	37.32	33.62	64.75	44.97	42.18	56.59	36.94	20.31	72.35
	+DtC Zheng et al. (2025)	36.94	32.87	65.33	44.63	41.82	56.94	36.71	20.08	72.58
	+RVE Qin et al. (2026)	37.06	33.04	65.19	44.79	42.01	56.76	36.97	20.15	72.42
	+IVE (Ours)	**34.92**	**31.11**	**67.21**	**42.94**	**39.31**	**59.09**	**34.91**	**17.60**	**74.76**
	Δ vs. Vanilla	**-2.75**	**-2.88**	**+2.82**	**-2.37**	**-3.16**	**+2.81**	**-2.31**	**-2.98**	**+2.69**
MCQ	Vanilla	68.05	51.04	43.70	73.91	53.85	39.94	67.16	49.69	44.91
	Object-specific									
	+VCD Leng et al. (2024)	67.86	50.90	43.85	73.67	53.75	40.09	67.02	49.58	45.02
	General									
	+OPERA Huang et al. (2024)	67.12	50.21	44.56	72.98	53.23	40.66	66.51	49.13	45.49
	+PAI Liu et al. (2024c)	66.93	51.10	44.00	74.05	53.83	39.91	67.58	49.25	45.08
	+HELPD Yuan et al. (2024)	66.84	50.58	44.39	73.18	53.27	40.57	66.44	49.04	45.58
	+DeCo Wang et al. (2025)	67.54	50.73	44.07	73.30	53.60	40.31	66.84	49.38	45.22
	+VASparse Zhuang et al. (2025)	67.21	50.88	44.07	73.91	53.48	40.20	66.51	49.21	45.44
	+MemVR Zou et al. (2025)	66.98	50.79	44.20	73.40	53.19	40.56	66.37	49.15	45.52
	Relation-specific									
	+CAP Wu et al. (2025)	67.40	50.81	44.06	73.49	53.21	40.52	66.79	49.33	45.27
	+DtC Zheng et al. (2025)	66.70	50.50	44.49	72.74	52.88	40.98	66.19	49.08	45.63
	+RVE Qin et al. (2026)	66.93	50.63	44.33	73.07	52.75	40.96	66.56	49.02	45.55
	+IVE (Ours)	**65.02**	**49.08**	**45.99**	**70.89**	**50.87**	**42.94**	**64.75**	**47.01**	**47.50**
	Δ vs. Vanilla	**-3.03**	**-1.96**	**+2.29**	**-3.02**	**-2.98**	**+3.00**	**-2.41**	**-2.68**	**+2.59**

4 EXPERIMENTS

We evaluate whether IVE reduces relation hallucination across three MLLMs while preserving performance on broader multimodal tasks. We also examine component contributions, robustness across configurations and decoding strategies, and computational cost.

4.1 EXPERIMENTAL SETTINGS

Benchmarks: Reefknot Zheng et al. (2025) and Hal-Eval Jiang et al. (2024) provide our primary evaluation of relation hallucination. Reefknot covers Y/N and multiple-choice relation prediction (scoring in Appendix B), while Hal-Eval distinguishes relation, object, attribute, and event hallucinations. MMRA Wu et al. (2026) evaluates multi-granularity, multi-image relational association. POPE Li et al. (2023b) provides an object-hallucination comparison. We use MMBench Liu et al. (2024d) and MME Fu et al. (2025) to assess whether IVE preserves performance on broader multimodal tasks.

Models: We evaluate IVE on the 7B variants of LLaVA-1.5 Liu et al. (2024a), InstructBLIP Dai et al. (2023), and Qwen2.5-VL-Instruct Bai et al. (2025).

Baselines: Besides Vanilla, we compare with object-specific VCD Leng et al. (2024). General methods include OPERA Huang et al. (2024), PAI Liu et al. (2024c), HELPD Yuan et al. (2024), DeCo Wang et al. (2025), VASparse Zhuang et al. (2025), and MemVR Zou et al. (2025). Relation-specific baselines include CAP Wu et al. (2025), DtC Zheng et al. (2025), and RVE Qin et al. (2026). CAP, DtC, and RVE are evaluated on both Reefknot settings and Hal-Eval Relation, with RVE reproduced from its paper description.

Implementation Details: We use selection layers 20, 20, and 18 for the three backbones and recalibrate visual values across all layers and heads within each decoding forward pass (Appendix B). Selection-layer ablations appear in Appendix Table 22. Main results use model-specific configurations selected on held-out Reefknot validation data and fixed across benchmarks before final evaluation. Experiments run on NVIDIA H800 GPUs.

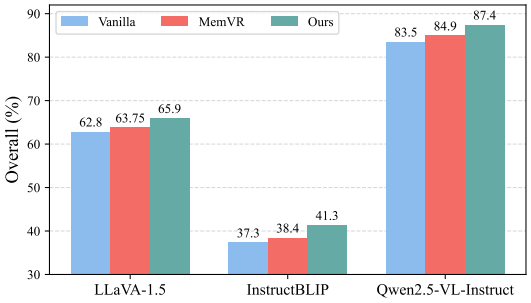
Table 3: Relation-type results (%) on Hal-Eval across three MLLMs.

Method	LLaVA-1.5				InstructBLIP				Qwen2.5-VL-Instruct			
	Acc↑	Δ	F1↑	Δ	Acc↑	Δ	F1↑	Δ	Acc↑	Δ	F1↑	Δ
Vanilla	52.65	+0.00	35.82	+0.00	49.80	+0.00	66.35	+0.00	62.43	+0.00	57.28	+0.00
Object-specific												
+VCD Leng et al. (2024)	53.91	+1.26	37.08	+1.26	50.88	+1.08	67.06	+0.71	63.37	+0.94	58.26	+0.98
General												
+OPERA Huang et al. (2024)	55.47	+2.82	38.53	+2.71	51.73	+1.93	67.41	+1.06	64.95	+2.52	59.43	+2.15
+PAI Liu et al. (2024c)	55.12	+2.47	38.02	+2.20	51.38	+1.58	67.29	+0.94	64.62	+2.19	59.08	+1.80
+HELPD Yuan et al. (2024)	55.41	+2.76	38.46	+2.64	51.94	+2.14	67.54	+1.19	65.18	+2.75	59.97	+2.69
+DeCo Wang et al. (2025)	54.76	+2.11	37.91	+2.09	51.21	+1.41	67.21	+0.86	64.18	+1.75	58.97	+1.69
+VASparse Zhuang et al. (2025)	55.34	+2.69	38.61	+2.79	52.27	+2.47	67.69	+1.34	65.41	+2.98	60.42	+3.14
+MemVR Zou et al. (2025)	55.83	+3.18	39.04	+3.22	52.61	+2.81	67.78	+1.43	65.62	+3.19	60.83	+3.55
Relation-specific												
+CAP Wu et al. (2025)	54.71	+2.06	37.96	+2.14	51.68	+1.88	67.42	+1.07	64.82	+2.39	59.74	+2.46
+DtC Zheng et al. (2025)	55.61	+2.96	38.83	+3.01	52.74	+2.94	67.77	+1.42	65.49	+3.06	60.91	+3.63
+RVE Qin et al. (2026)	55.38	+2.73	39.17	+3.35	52.49	+2.69	67.79	+1.44	65.73	+3.30	60.68	+3.40
+IVE (Ours)	**57.71**	**+5.06**	**41.29**	**+5.47**	**54.93**	**+5.13**	**68.93**	**+2.58**	**67.79**	**+5.36**	**62.84**	**+5.56**

Table 4: Accuracy (%) on MMRA with Qwen2.5-VL-Instruct.

Method	Accuracy↑
Vanilla	50.45
+VCD Leng et al. (2024)	51.12
+OPERA Huang et al. (2024)	51.38
+PAI Liu et al. (2024c)	51.27
+MemVR Zou et al. (2025)	52.16
+IVE (Ours)	**55.37**
Δ vs. Vanilla	**+4.92**

Figure 5: Results comparison (%) on MMBench.



4.2 Main Results

Relation Hallucination: On Reefknot Zheng et al. (2025), IVE attains the highest R_{score} and lowest $Cogn.$ hallucination rates across all three backbones in both Y/N and MCQ settings (Table 2). These gains are accompanied by lower $Perc.$ hallucination rates. The advantage over relation-specific baselines extends across these formats, covering both verification of a proposed relation and selection among alternatives. IVE also outperforms ADAPT's training-free variant and Attention Never Lie across all three backbones in both Reefknot settings (Appendix Table 12). On Hal-Eval Jiang et al. (2024), IVE achieves higher Relation accuracy and F1 than all compared methods across the three backbones (Table 3). For LLaVA-1.5, the 5.06-point accuracy gain over Vanilla exceeds the gains on Object, Attribute, and Event (Appendix Table 13). The larger benefit on Relation is consistent with the method's focus on relational grounding.

Relational Generalization on MMRA: To examine whether the gains extend to relations involving multiple images, we evaluate IVE on MMRA Wu et al. (2026) with Qwen2.5-VL-Instruct. IVE improves accuracy by 4.92 points over Vanilla and by 3.21 points over MemVR, the strongest baseline in this comparison (Table 4). This result supports the use of history-conditioned visual reweighting when relation prediction draws on evidence from multiple images.

Broader Multimodal Performance: IVE improves MMBench Liu et al. (2024d) scores over Vanilla and MemVR across the three backbones (Figure 5). Scores on the selected MME categories also improve over Vanilla across all three models (Appendix Table 18). These results indicate preserved performance on the broader multimodal tasks evaluated here. POPE and CHAIR further cover object-existence questions and generated captions, with CHAIR stratified by scene density and response length (Appendix Tables 19 and 20).

4.3 Ablation Studies and Efficiency

Component Contributions: We separately disable emergent-token reinforcement (*Reinforce*) and inertia-token attenuation (*Attenuate*) on LLaVA-1.5, retaining token-selection and history-update rules (Appendix C.3). Each branch improves over Vanilla on Reefknot Zheng et al. (2025) and POPE Li et al. (2023b), while their combination performs best across the reported metrics (Table 5(a)). To examine the role of token selection within this paired design, we compare relevance signals under

Table 5: Ablation, configuration robustness, and inference efficiency of IVE.

(a) Component contributions

Reinforce	Attenuate	Reefknot		POPE	
		Perc.↓	Cogn.↓	Acc.↑	F1↑
		37.67	33.99	83.29	81.33
✓		36.94	33.01	89.71	89.96
	✓	37.12	33.24	88.96	89.21
✓	✓	**34.92**	**31.11**	**90.53**	**90.94**

(b) Unified configuration

	Reefknot Y/N ($\mathbf{R_{score}}$ ↑)		
Backbone	**Vanilla**	**Unified**	**Tuned**
LLaVA-1.5	64.39	66.13	**67.21**
InstructBLIP	56.28	57.93	**59.09**
Qwen2.5-VL	72.07	73.38	**74.76**

(c) Hyperparameter sensitivity

Parameter	Value	Reefknot		POPE	
		Perc.↓	Cogn.↓	Acc.↑	F1↑
α	0.05	36.84	32.72	89.20	89.27
	0.10	**34.92**	**31.11**	**90.53**	**90.94**
	0.15	36.60	32.43	89.57	89.62
γ	0.08	36.56	32.48	89.77	89.82
	0.09	**34.92**	**31.11**	**90.53**	**90.94**
	0.10	35.88	32.07	90.53	90.94
τ	2.95	36.72	32.44	89.63	89.68
	3.00	**34.92**	**31.11**	**90.53**	**90.94**
	3.05	35.93	32.10	90.27	90.32

(d) Inference efficiency

Backbone	Method	$\mathbf{R_{score}}$ ↑	Time↓	Mem.↓
LLaVA-1.5	Vanilla	64.39	0:14:44	15.15G
	+VCD	64.71	0:38:23	16.06G
	+MemVR	65.34	0:16:07	15.54G
	+IVE	**67.21**	**0:15:47**	**15.38G**
InstructBLIP	Vanilla	56.28	0:47:02	31.13G
	+VCD	56.38	1:50:08	32.17G
	+MemVR	56.92	0:53:25	31.68G
	+IVE	**59.09**	**0:52:52**	**31.45G**
Qwen2.5-VL	Vanilla	72.07	0:31:46	33.41G
	+VCD	72.17	1:32:21	34.65G
	+MemVR	72.52	0:34:02	33.72G
	+IVE	**74.76**	**0:33:45**	**33.61G**

matched token budgets and fixed scaling factors. This matches token counts and scaling factors while varying which visual tokens receive the adjustments. Under these controls, the IVE score achieves the highest R_{score} among tested signals (Appendix Table 14). Matched random-token and random-value controls yield smaller gains than structured selection and reweighting (Appendix Table 17).

Unified Configuration: To assess performance with a single configuration, we apply shared coefficients and a common layer-selection rule across the three backbones (Table 5(b)). We use $\alpha = 0.10$, $\gamma = 0.10$, $\tau = 3.05$, and $l_{\text{sel}} = \text{round}(0.625L)$, where L denotes the number of transformer layers in each backbone. This unified configuration improves over Vanilla on all three backbones, while model-specific tuning yields further gains for each backbone.

Hyperparameter Sensitivity: We vary α, γ, and τ individually around the validation-selected LLaVA-1.5 setting while keeping the other coefficients fixed (Table 5(c)). This setting yields the lowest $Cogn.$ hallucination rate, but every tested setting improves over Vanilla on both Reefknot and POPE. Gains persist throughout the tested neighborhood, with their magnitude varying across configurations. Full three-backbone sensitivity results are reported in Appendix Table 21.

Inference Efficiency: On Reefknot, IVE improves R_{score} over Vanilla with roughly 6–12% longer runtime and less than 2% additional GPU memory (Table 5(d)). It also achieves higher R_{score} than VCD and MemVR across the three backbones, while requiring less runtime and memory than either.

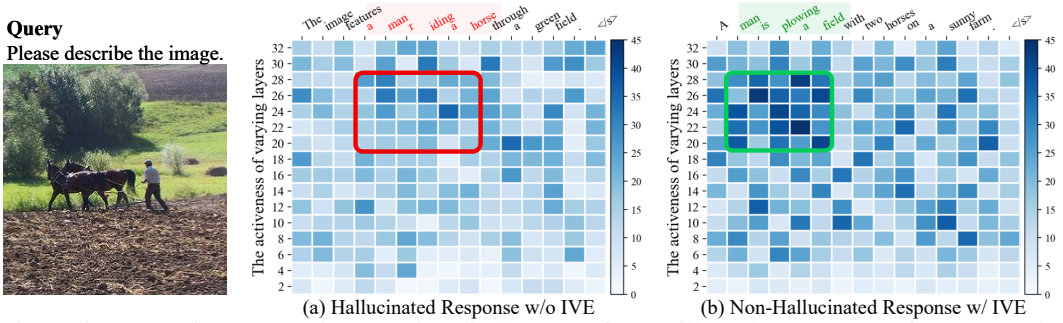
4.4 Analysis

Robustness Across Decoding Strategies: Because decoding strategies alter output sequences and visual-attention histories, we evaluate IVE under greedy decoding, beam search, and nucleus sampling on Reefknot with matched settings across methods. Across all three backbones and strategies, IVE outperforms Vanilla and MemVR, with larger R_{score} gains over Vanilla under greedy decoding than under beam search or sampling (Table 6). Gains persist even when beam search improves Vanilla.

Layer-wise Attention Displacement: Figure 6 maps consecutive visual-attention displacement across layers and output positions for a representative LLaVA-1.5 Reefknot example. This token-wise view localizes attention shifts across output positions. Vanilla exhibits limited displacement around the unsupported description *riding*, whereas IVE exhibits larger displacement around the supported description *plowing a field*. The contrast at relation-bearing positions reflects the persistent and shifting visual focus observed in Section 2. For LLaVA-1.5 and InstructBLIP, the method ordering is consistent under Wasserstein and Jensen–Shannon distances (Appendix Table 8).

Table 6: Y/N R_{score} on Reefknot under different decoding strategies.

Decoding	LLaVA-1.5			InstructBLIP			Qwen2.5-VL-Instruct		
	Vanilla	MemVR	IVE	Vanilla	MemVR	IVE	Vanilla	MemVR	IVE
Greedy	64.39	65.34	**67.21**	56.28	56.92	**59.09**	72.07	72.52	**74.76**
Beam ($b = 4$)	65.26	66.10	**67.10**	56.81	57.51	**58.57**	72.79	73.18	**74.20**
Sampling ($p = 0.9$)	63.94±0.42	64.85±0.45	**65.77±0.39**	55.96±0.51	56.66±0.48	**57.76±0.47**	71.82±0.38	72.24±0.41	**73.32±0.35**



(a) Hallucinated Response w/o IVE (b) Non-Hallucinated Response w/ IVE

Figure 6: Layer-wise consecutive attention displacement for Vanilla and IVE on a Reefknot example. Rows denote LLaVA-1.5 layers and columns denote decoded tokens.

5 Related Work

Relation Hallucination in MLLMs: MLLM hallucination research has often emphasized object existence, as in POPE Li et al. (2023b). Relation hallucination instead concerns image-unsupported relations among visible entities. Reefknot Zheng et al. (2025) and Hal-Eval Jiang et al. (2024) provide relation-focused and fine-grained evaluations, while MMRA Wu et al. (2026) extends relational association across granularities and images. Beyond evaluation, CAP Wu et al. (2025) addresses spatial-relation errors through prompting constraints, while RVE Qin et al. (2026) enhances action-relevant regions. DtC Zheng et al. (2025) combines confidence-based detection with intermediate-layer calibration to mitigate relation hallucination.

Visual Evidence Calibration: Training-free methods enhance visual grounding through attention and token-level calibration Liu et al. (2024c); Yin et al. (2025); Zhuang et al. (2025) or visual-evidence recovery Wang et al. (2025); Zou et al. (2025). OPERA Huang et al. (2024) links hallucination to attention aggregation onto generated summary tokens and penalizes over-trust in them. Visual-sink analyses Kang et al. (2025) identify highly attended tokens with limited semantic content, motivating attention redistribution to other visual tokens. We examine how dominant visual evidence persists across autoregressive decoding steps, complementing studies of within-step attention concentration. This persistence can involve informative tokens even when overall image attention remains strong. Temporal approaches also detect attention defocus in text windows or guide attention with an early visual anchor Zhao et al. (2026); Yao et al. (2026). IVE uses an evolving token-specific reference to couple reinforcement of newly relevant evidence with attenuation conditioned on repeated dominance. Appendix B.1 compares these signals and interventions.

6 Conclusion

Relation hallucinations can occur despite correct object recognition. In the relation descriptions analyzed here, hallucinated responses exhibit more persistent visual focus than correct ones, reflecting a form of visual inertia. Increasing previous-step attention carryover degrades relation prediction more than an entropy-matched current-step reference. Motivated by these observations, we introduced Inertia-aware Visual Excitation (IVE), a lightweight, training-free decoding-time method that uses each visual token's recent attention history to reweight its contributions. IVE favors newly relevant evidence while mildly attenuating persistently dominant contributions. Across three MLLMs, IVE improves relation prediction on Reefknot and Hal-Eval. These gains extend to the multi-image relational setting of MMRA, while performance on broader multimodal benchmarks is preserved.

AI USE STATEMENT

Generative AI tools assisted with literature review, experimental-design discussions, code development and experiment execution, result analysis, manuscript editing, and figure and table preparation. Research decisions and interpretation remained under author direction. The authors take responsibility for the final manuscript and all AI-assisted content, including citations, code, experimental results, and scientific claims.

ETHICS STATEMENT

IVE is intended to improve the reliability of multimodal predictions by reducing relation hallucinations. Our evaluation uses existing multimodal benchmarks. Although IVE reduces hallucination rates in the evaluated settings, the underlying models can still produce unsupported statements and retain biases inherited from pretraining data. Applications involving consequential decisions require task-specific validation and appropriate human oversight.

REPRODUCIBILITY STATEMENT

Section 3 specifies the history-conditioned selection and visual-value reweighting rules of IVE. Appendix B provides the decoding procedure and model-specific selection layers, while Section 4 describes experimental settings and evaluation metrics. The appendix additionally documents attention-displacement analyses, carryover controls, and configuration studies to support reproduction of the reported comparisons.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, volume 36, pp. 49250–49267, 2023.

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. In *NeurIPS*, 2025.

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 *CVPR*, pp. 13418–13427, 2024.

Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *ACM MM*, pp. 525–534, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *ICLR*, 2025.

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 *CVPR*, pp. 13872–13882, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, volume 36, pp. 34892–34916, 2024b.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *ECCV*, pp. 125–140, 2024c.

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 *ECCV*, pp. 216–233, 2024d.

Zhenxin Qin, Qiang Li, Qingzhuo Wang, Ruiyang Qin, Zhihua Wei, and Wen Shen. Mitigating action-relation hallucinations in lvlms via relation-aware visual enhancement. In *ACL*, pp. 24907–24923, 2026.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *ICLR*, pp. 13712–13736, 2025.

Jiarui Wu, Zhuo Liu, and Hangfeng He. Mitigating hallucinations in multimodal spatial relations through constraint-aware prompting. In *NAACL Findings*, pp. 3450–3468, 2025.

Siwei Wu, King Zhu, Yu Bai, Yiming Liang, Yizhi Li, Haoning Wu, Jiaheng Liu, Ruibo Liu, Xingwei Qu, Xuxin Cheng, et al. Mmra: A benchmark for evaluating multi-granularity and multi-image relational association capabilities in large visual language models. In *EACL Findings*, pp. 5405–5419, 2026.

Zhiyuan Yao, Zheren Fu, Zhixiao Zheng, Jiajun Li, Yi Tu, and Zhendong Mao. ADAPT: Attention dynamics alignment with preference tuning for faithful MLLMs. In *ECCV*, 2026.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *CVPR*, pp. 14625–14634, 2025.

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. *arXiv preprint arXiv:2509.21789*, 2025.

Fan Yuan, Chi Qin, Xiaogang Xu, and Piji Li. Helpd: Mitigating hallucination of lvlms by hierarchical feedback learning with vision-enhanced penalty decoding. In *EMNLP*, pp. 1768–1785, 2024.

Yufei Zhan, Ziheng Wu, Yousong Zhu, Rongkun Xue, Guanghao Zhou, Ruipu Luo, Zhenghao Chen, Can Zhang, Yifan Li, Zhentao He, Zheming Yang, Ming Tang, Minghui Qiu, and Jinqiao Wang. GThinker: Towards general multimodal reasoning via cue-guided rethinking. In *CVPR*, pp. 11954–11965, 2026.

Chenxi Zhao, Yan Zhou, and Jufeng Yang. Attention never lie: Visual attention defocus reveals and rectifies hallucinations in MLLMs. In *CVPR Findings*, pp. 8897–8906, 2026.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, Huiyu Zhou, and Xuming Hu. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. In *ACL Findings*, pp. 6193–6212, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, pp. 18378–18394, 2024.

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *CVPR*, pp. 4189–4199, 2025.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *ICML*, 2025.

APPENDIX OVERVIEW

This appendix provides supporting attention analyses, implementation details, extended evaluations, and qualitative examples. The contents are organized as follows:

- **Details for Motivation Analyses (§A):** Attention-displacement analyses, cross-backbone distance comparisons, and carryover controls supporting Section 2.
- **More Details of Implementation (§B):** Configuration selection, within-forward value recalibration, detailed pseudocode, and comparison with temporal attention methods.
- **More Experimental Results and Analysis (§C):** Type-wise Hal-Eval results, relevance-signal and random-control comparisons, POPE and CHAIR evaluations, broader capability checks, and configuration studies.
- **Case Studies (§D):** Qualitative comparisons between Vanilla and IVE.

A DETAILS FOR MOTIVATION ANALYSES

A.1 ATTENTION DISPLACEMENT UNDER NAIVE AMPLIFICATION

To further examine why naive visual attention amplification is insufficient for relation-hallucination mitigation, we compare the temporal attention patterns produced by the baseline model and the amplification method PAI Liu et al. (2024c). Overall attention strength describes the influence assigned to the image. Comparing visual-attention distributions across output positions further reveals whether the same visual-token contributions remain dominant throughout a response.

Specifically, we generate 1,000 captions using LLaVA-1.5 Liu et al. (2024a) on a subset of COCO Lin et al. (2014). During generation, we store the normalized visual attention weights of the first 100 generated tokens. For each sample, we then compute a pairwise distance matrix (100×100) over visual-attention patterns using the same one-dimensional Wasserstein distance over flattened patch indices as in Equation 2. This matrix covers both nearby and distant output positions, providing a trajectory-level view of whether similar visual-attention patterns persist over an extended portion of the response.

Normalizing the visual-token slice separates its relative allocation from the total attention assigned to the image. The index-based ground cost preserves the fixed visual-token ordering and assigns larger costs to more widely separated indices. In the pairwise matrix, entries near the diagonal compare nearby output positions, while entries farther from it compare more widely separated stages of generation. Extended low-distance regions thus indicate that similar indexed attention distributions recur across an interval of the response.

PAI exhibits lower displacement than Vanilla across many output-position pairs (Figure 7). Its normalized visual-attention distributions therefore remain more similar under the fixed token ordering over the course of generation. In this captioning comparison, amplification is accompanied by more persistent visual focus, extending the analysis beyond the overall attention strength measured in the main text.

Figure 8 compares Vanilla and IVE using the same normalization and flattened-index ground cost. Together, the two comparisons examine temporal attention patterns under both amplification and history-conditioned contribution reweighting. The pairwise view complements the consecutive-step average by retaining changes over short and long intervals. These COCO analyses compare decoding conditions during captioning. Their role differs from the main-text Reefknot surfaces, which compare prediction outcomes under one Vanilla configuration. The former describe how attention trajectories vary with the intervention, whereas the latter connect persistent visual focus to relation-prediction outcomes.

A.2 RELATION-DESCRIPTION PROTOCOL AND ATTENTION DISPLACEMENT

Task and generation. Each sample pairs an image with two entities specified in a Reefknot Y/N relation question. We generate a relation description with Vanilla LLaVA-1.5-7B using greedy decoding and a maximum of 128 new tokens. The task request is:

> Please describe the relationship between the [entity A] and the [entity B] in the image.

The entity placeholders are filled from the source question, and the request leaves response length unconstrained within the decoding limit. All candidate responses are generated under the same configuration before outcome grouping.

Outcome annotation. A target relation is *correct* when the response explicitly describes an image-supported relation between the specified entities. It is *unsupported* when the response makes an image-unsupported claim about that relation, including responses that mix supported and unsupported target-relation statements. Responses that omit the target relation or are too ambiguous to judge are labeled *unanswered/ambiguous*. A partially informative response counts as correct only if it still expresses a supported target relation without a conflicting target-relation claim. Additional unsupported entity, attribute, or relation claims are recorded separately from target correctness.

Outcome groups. For the main-text surfaces, the correct group requires a correct target relation and no additional hallucinated claims. The hallucinated group contains responses with an unsupported target relation. Unanswered or ambiguous responses, and target-correct responses containing additional hallucinations, are excluded from this two-group comparison. After annotation, we retain responses with at least 50 valid output tokens and randomly select 500 eligible responses per group using seed 42. Selection is independent of the measured attention displacement. We compare separately sampled outcome groups, aligning attention trajectories by output-token index. The contrast characterizes a group-level association between attention persistence and relation errors.

Attention positions and aggregation. We record attention at the transformer block denoted by $l_s = 20$, averaging over its heads and normalizing the visual-token slice as in Equation 1. Position t refers to the query that predicts the corresponding output token. We use the first 50 valid output-token positions, excluding BOS, EOS, and padding tokens, and retain only responses that provide all 50 positions. Shorter responses are excluded without padding, repetition, or interpolation. The 24×24 patches are flattened in row-major order to indices $0, \ldots, 575$. We compute the weighted one-dimensional Wasserstein distance with SciPy's `wasserstein_distance`, passing the indices as support values and the normalized attention distributions as weights. The ground cost is $|a - b|$, giving distances in $[0, 575]$. Each response contributes one 50×50 matrix, and we average these matrices elementwise within each outcome group, giving every selected response equal weight.

Independent prediction-level analysis. The consecutive-displacement analysis uses a separate Reefknot sample cohort with the same relation-description request, generation settings, and target-relation annotation rules, retaining its natural correctness distribution. Relation accuracy counts target-correct responses over all evaluated responses in each displacement group, including unanswered or ambiguous responses in the denominator. Target-relation correctness is assessed independently of additional hallucinations elsewhere in the response. A pairwise matrix describes the temporal structure of one trajectory, but a single summary is useful when comparing many predictions with their outcomes. We summarize each trajectory by averaging Wasserstein distance over consecutive output positions:

$$D_{\text{attn}} = \frac{1}{T-1} \sum_{t=1}^{T-1} W_{\text{idx}}(\mathbf{p}_t, \mathbf{p}_{t+1}) \,, \tag{11}$$

where T is the number of analyzed output positions and $\mathbf{p}_t$ is the normalized visual-attention distribution defined in Equation 1. Each term measures the change from one output position to the next, and averaging prevents longer analyzed sequences from receiving a larger score simply because they contain more transitions. A low value indicates relatively similar attention distributions at successive positions. A higher value indicates greater mean change under the fixed flattened-token ordering.

Relation accuracy increases with consecutive attention displacement (Table 7). The same statistic has a positive Spearman correlation with relation accuracy ($\rho = +0.31$) and a negative correlation with the corresponding hallucination rate ($\rho = -0.28$), with $p < 0.001$ in both cases. These results associate changes in visual focus with prediction outcome on this separate sample.

Table 7: Relation accuracy across consecutive attention-displacement groups on Reefknot.

Group	D_{attn} range	Samples	Relation accuracy ↑
Low	$[0, 32)$	∼330	58.3%
Medium	$[32, 48)$	∼340	64.7%
High	$[48, \infty)$	∼330	71.2%

Table 8: Attention distances and Y/N R_{score} for LLaVA-1.5 and InstructBLIP. D_{attn} uses Wasserstein distance over flattened patch indices, and D_{JS} uses Jensen–Shannon distance.

(a) LLaVA-1.5

Method	D_{attn}	D_{JS}	R_{score} ↑
Vanilla	41.52	0.193	64.39
+PAI	38.64	0.181	64.71
+IVE	**47.21**	**0.224**	**67.21**

(b) InstructBLIP

Method	D_{attn}	D_{JS}	R_{score} ↑
Vanilla	39.83	0.187	56.28
+PAI	37.12	0.174	56.80
+IVE	**45.67**	**0.216**	**59.09**

The two summaries organize their samples differently. The main-text surfaces first separate predictions by correctness and then compare their trajectories. Here, predictions are grouped by displacement, and relation accuracy is measured within each group. Each displacement group contains both correct and incorrect predictions. The increasing accuracy across groups provides complementary evidence for the association between attention change and relation accuracy.

A.3 Attention Displacement Across Backbones

We compare Vanilla, PAI, and IVE on Reefknot to examine whether their attention-displacement patterns are consistent across LLaVA-1.5 and InstructBLIP. For InstructBLIP, we project its LLM-side Q-Former query attention through Q-Former cross-attention to obtain a patch-level distribution, flatten the resulting patch order, and then compute the same index-based Wasserstein distance. Each visual encoder defines its own support and patch indexing, so we compare method rankings within each backbone.

We also report Jensen–Shannon distance, denoted D_{JS}, which is invariant to applying the same permutation of patch indices to both distributions and measures allocation changes independently of index separation. Together, the Wasserstein and Jensen–Shannon comparisons test whether the within-backbone method ordering persists across distance measures.

Both distance measures give the same ordering for both backbones, with lower attention change under PAI than Vanilla and higher change under IVE (Table 8). The agreement across backbones and distance choices supports the consistency of this temporal pattern.

PAI modestly improves R_{score} despite reducing the two attention-distance measures, while IVE combines larger attention changes with the strongest performance in each backbone.

A.4 Progressive Carryover Control

The carryover intervention in Equation 3 mixes the previous-step visual-attention distribution into the current one. Increasing the mixing weight makes the previous spatial pattern more influential, while the current distribution continues to contribute. To compare temporal carryover at a similar level of attention concentration, the matched control uses the same entropy but reshapes only the current-step distribution:

$$\mathbf{p}_t^{\text{ctrl}}(\lambda) = \text{Norm}\left(\mathbf{p}_t^{\beta_t}\right), \qquad \mathcal{H}\left(\mathbf{p}_t^{\text{ctrl}}\right) = \mathcal{H}(\mathbf{p}_t^{\text{carry}}). \tag{12}$$

Table 9: IVE under induced previous-step carryover at $\lambda = 0.5$ on Reefknot Y/N.

Condition	D_{attn}	Cogn.↓	R_{score} ↑
Vanilla	41.52	33.99	64.39
Carryover	28.73	39.84	54.50
Carryover + IVE	36.81	35.12	59.77

Table 10: Supporting attention interventions on Reefknot Y/N. Previous-step propagation corresponds to the $\lambda = 1$ carryover condition in Table 1. D_{attn} is mean consecutive attention displacement.

Intervention	D_{attn}	R_{score} ↑
Vanilla	41.52	64.39
Random perturbation	38.76	62.27
Shuffled previous-step	34.21	59.36
Current-step focus	30.47	56.61
Previous-step propagation	19.24	47.68

At each step, $\beta_t > 0$ is chosen so that the control and carryover distributions have the same entropy. Here, Norm denotes ℓ_1 normalization and $\mathcal{H}$ denotes entropy. The resulting relative distribution is mapped back to the visual-token attention slice while preserving its total mass.

We initialize the previous-attention state from the last eligible input-text query before the final prompt position. Carryover and its matched reference are applied before attention–value aggregation, with outputs selected from the resulting logits. The previous-attention state advances once per query.

The exponent adjusts concentration while preserving the ordering of the current attention weights. Values above one sharpen the distribution toward its dominant entries, whereas values below one flatten it. This current-step transformation contains no previous-step spatial allocation, and preserving total visual-attention mass isolates redistribution within the visual slice.

Under progressive carryover, both R_{score} and consecutive displacement decline more strongly than under the matched reference. The widening performance gap in Table 1 supports a contribution from temporal persistence under this intervention. Entropy specifies the concentration being matched, while displacement records the accompanying change in the trajectory.

Recovery under induced carryover. We additionally apply IVE under carryover at $\lambda = 0.5$ to examine whether history-conditioned reweighting remains useful when persistence is explicitly strengthened. The relevant comparison is between carryover alone and carryover with IVE, with Vanilla providing the no-carryover reference. The combined condition follows the initialization and value-update procedure in Appendix B. IVE partially recovers the lost R_{score} and moves both attention displacement and the $Cogn.$ hallucination rate toward their Vanilla levels (Table 9). The joint improvement connects the method's behavioral benefit with a change in the induced attention pattern.

A.5 Manual Intervention Settings

For the intervention analysis in Table 10, we apply several manual perturbations to the visual attention maps of LLaVA-1.5 Liu et al. (2024a) on Reefknot Zheng et al. (2025). These settings provide a broader comparison of attention perturbations alongside the progressive carryover analysis. All perturbations precede attention–value aggregation, with input-initialized attention history for the temporal conditions (Appendix A.4).

- **Random perturbation.** We add random noise to the current-step visual attention map and then re-normalize it, introducing a generic disturbance without carrying forward the previous spatial pattern.
- **Shuffled previous-step attention.** We inject previous-step attention after shuffling its spatial positions. This retains the previous attention values while changing which regions receive them.

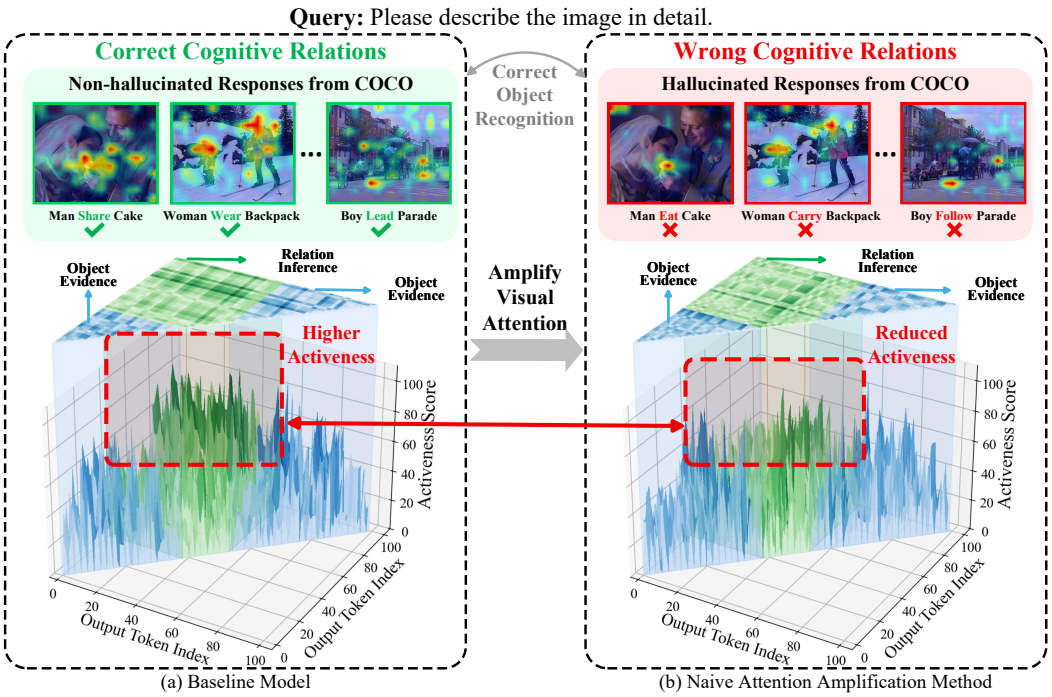


Figure 7: **Pairwise visual-attention displacement under naive amplification.** On the COCO captioning analysis, PAI exhibits lower pairwise displacement than the baseline across many output-token pairs. This supporting result concerns the effect of amplification and is distinct from the outcome-linked Reefknot comparison in the main text.

- **Current-step focus.** We sharpen the current-step distribution toward its already dominant regions, increasing concentration without introducing previous-step attention.
- **Previous-step propagation.** We mix the current- and previous-step distributions with equal weight and normalize the result, corresponding to $\lambda = 1$ in Equation 3.

The shuffled and unshuffled conditions test whether previous attention values retain their spatial assignment. Current-step focus strengthens regions that already dominate the present distribution, and random perturbation supplies a generic disturbance. These conditions compare distinct perturbation schemes at their respective intervention strengths.

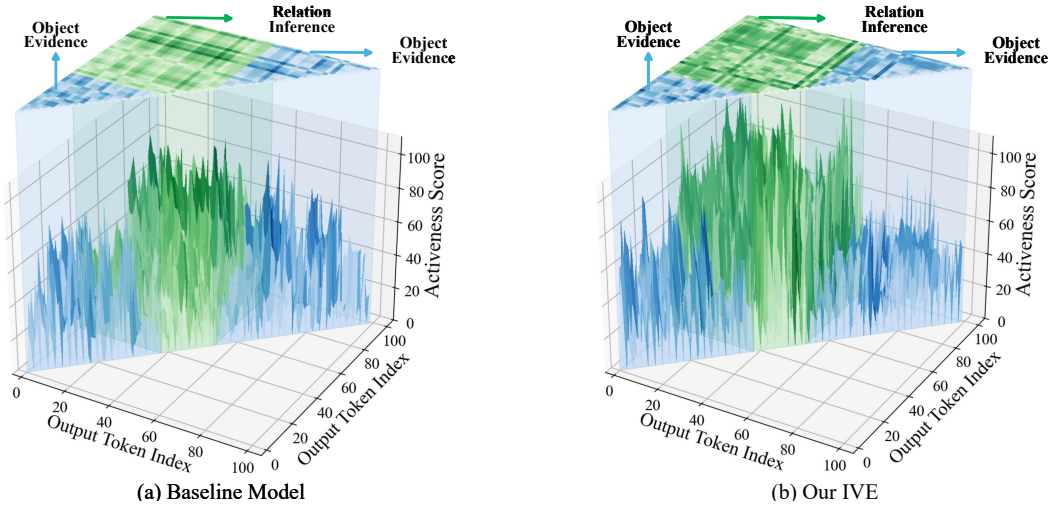
Previous-step propagation produces the largest R_{score} reduction among these settings, accompanied by the lowest consecutive displacement. Its $\lambda = 1$ setting combines the current and previous distributions with equal weight. The entropy-matched comparison in Section 2 separately examines carryover at comparable attention concentration.

B More Details of Implementation

Token selection uses head-averaged attention over all tokens at layer 20 for LLaVA-1.5 Liu et al. (2024a), layer 20 for InstructBLIP Dai et al. (2023), and layer 18 for Qwen2.5-VL-Instruct Bai et al. (2025). Value recalibration spans all attention heads and transformer layers. Displacement analysis uses the unit-normalized visual slice at output positions.

Reefknot evaluation score. For each Reefknot task t, we report its task-conditioned component of the benchmark score:

$$R_{score}(t) = 100 - \frac{N_{P,t}H_{P,t} + N_{C,t}H_{C,t}}{N_{P,t} + N_{C,t}}.$$



(a) Baseline Model (b) Our IVE

Figure 8: Pairwise visual-attention displacement for Vanilla and IVE on COCO captioning samples.

Here, $H_{P,t}$ and $H_{C,t}$ are the perception- and cognition-hallucination rates in percent, and $N_{P,t}$ and $N_{C,t}$ are the corresponding sample counts from the released splits. We report this score separately for Y/N and MCQ.

Hal-Eval metrics. For the discriminative evaluations in Tables 3 and 13, the positive class is an image-consistent description (*Yes*). Accuracy and binary F1 are computed from the same predictions. Within each reported evaluation condition, we pool the confusion counts over examples and compute $F_1 = 2TP/(2TP + FP + FN)$. Both metrics are reported as percentages.

Decoding settings. Table 6 compares greedy decoding, beam search with four beams, and nucleus sampling with top-$p = 0.9$, temperature 1.0, and no top-k truncation. For each backbone and method, sampling uses five runs with seeds 0, 1, 2, 3, and 4. We report the mean and sample standard deviation of the five full-evaluation R_{score} values, using the same decoding settings across methods.

Configuration selection. We select α, γ, τ, and the token-selection layer separately for each backbone using only held-out Reefknot validation data, disjoint from the final evaluation samples. Each configuration is fixed before testing and reused across that backbone's benchmark evaluations, including Hal-Eval, POPE, and MMRA. The unified configuration is also fixed before evaluation and shared across the three backbones. The coefficient and layer sweeps in Tables 21 and 22 examine sensitivity around the selected configurations.

Within-forward value recalibration. We retain fixed copies of the original visual values separately from the working cache. Before generation, the EMA reference, inertia counts, and history length are accumulated in causal order from image-attending input-text queries preceding the final prompt position, excluding padding and visual-token queries, while visual values remain unchanged. For each output-producing query, the hook at l_s reads attention before attention–value aggregation and computes token sets and factors from the preceding history. Emergent tokens are selected using the original relevance scores, and their multipliers are capped at 1.10 before cache updates. At the reported thresholds, $1 + S_{t,j} > 1.10$ for every selected emergent token, so its applied multiplier is 1.10.

At each query, every visual-token entry in the working cache is assigned its capped factor times the corresponding original value, across all layers and heads. Tokens outside both selected sets have factor one and are restored to their original values. The original copies remain unchanged, and the active tensor at l_s uses the reconstructed entries without reapplying the factors.

Algorithm 1: Within-forward recalibration from fixed original values in IVE.

Input: MLLM f with L layers, final input token x_P, KV cache for $\mathbf{x}_{<P}$, fixed original visual values $\{\mathbf{V}_j^{(l,h),\text{base}}\}$, history state $(\mathbf{b}, \mathbf{C}, n)$, visual-token set $\mathcal{V}$, selection layer l_s, maximum output length $T_{\max}$, threshold τ, coefficients α, γ
Output: Generated response
$\epsilon \leftarrow 10^{-6}$, $\mathbf{y} \leftarrow \varnothing$
$q \leftarrow x_P$ with its original position and causal mask
for $t \leftarrow 1$ **to** $T_{\max}$ **do**
  Process q through layers $1, \ldots, l_s - 1$, extending the cache
  Compute $\{\mathbf{A}_t^{(h)}\}_{h=1}^{N_h}$ at layer l_s before value aggregation
  $a_{t,j} \leftarrow N_h^{-1} \sum_{h=1}^{N_h} \mathbf{A}_{t,j}^{(h)}$ for every $j \in \mathcal{V}$
  if $n > 0$ **then**
    Compute $\mathbf{S}_t, \mathcal{E}_t, \mathcal{I}_t$ from $\mathbf{a}_t, \mathbf{b}$ using Equations 6 and 7
    Compute $m_{t,j}$ using Equation 9 with $C_{t,j} = C_j$ and $n_t = n$
    Cap $m_{t,j}$ at 1.10 for $j \in \mathcal{E}_t$
  else
    $\mathcal{E}_t \leftarrow \varnothing$, $\mathcal{I}_t \leftarrow \varnothing$, and $m_{t,j} \leftarrow 1$ for $j \in \mathcal{V}$
  foreach *transformer layer l and attention head h* **do**
    $\mathbf{V}_j^{(l,h)} \leftarrow m_{t,j} \mathbf{V}_j^{(l,h),\text{base}}$ for every visual token $j \in \mathcal{V}$
  Use the updated visual entries in layer l_s's active value tensor
  Complete layer l_s and layers $l_s + 1, \ldots, L$, without reapplying the factors
  Compute $\mathbf{z}_t$ from the resulting final hidden state
  Select y_t from $\mathbf{z}_t$ with the base decoding rule and append it to $\mathbf{y}$
  if $n = 0$ **then**
    $\mathbf{b} \leftarrow \mathbf{a}_t$
  else
    $\mathbf{b} \leftarrow \gamma \mathbf{b} + (1 - \gamma)\mathbf{a}_t$
  $C_j \leftarrow C_j + \mathbf{1}[j \in \mathcal{I}_t]$ for $j \in \mathcal{V}$, and $n \leftarrow n + 1$
  if y_t *is the end-of-sequence token* **then**
    break
  $q \leftarrow y_t$ at the next sequence position
return $\mathbf{y}$

Layer l_s and later layers use the adjusted values in the current forward pass. Earlier layers have already executed and use the newly assigned cache entries on the next query. Each output is produced by one forward pass, with its resulting logits passed to the original decoding rule.

History updates and persistence counts. After value recalibration, the pre-intervention attention observation updates the EMA with weights γ and $1 - \gamma$ for the old reference and new observation. The reported coefficients give greater weight to the latest observation. Inertia counts and history length advance once, excluding the current query from its own factor. A query without history uses identity modulation and initializes the reference.

The frequency $C_{t,j}/n_t$ determines the inertia factor within $[1 - \alpha, 1]$. With the reinforcement cap, the applied value multiplier lies in $[1 - \alpha, 1.10]$ relative to the fixed original value at every query. The EMA, inertia counts, and history length evolve across queries, while each value update is computed from the original value.

State is local to each sample and beam hypothesis, following cache branching and reordering. Algorithm 1 uses $(\mathbf{b}, \mathbf{C}, n)$ for the preceding EMA, inertia counts, and history length. Reported runtimes include input processing and cache updates.

B.1 Comparison with Temporal Attention Methods

Temporal attention provides information about both the changing relevance of visual tokens and the deterioration of visual grounding during generation. The methods below use this information at different levels, from individual visual contributions to candidate text continuations. Table 11

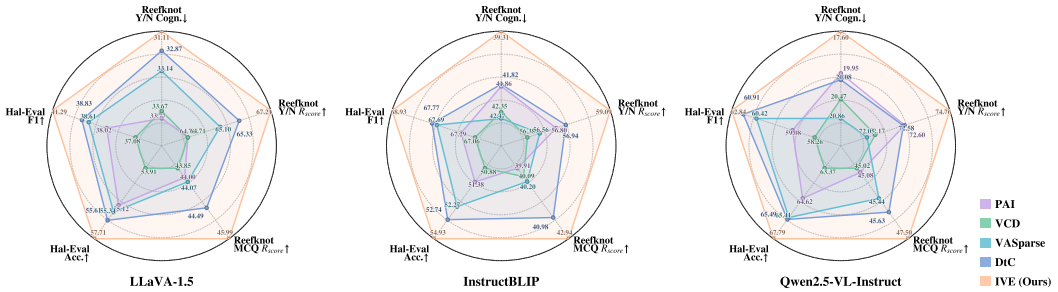


Figure 9: Radar comparison on Reefknot and Hal-Eval across three MLLMs. Each axis uses a padded raw-score range whose favorable endpoint is the best reported value. The Y/N Cogn. axis is reversed so that farther outward consistently indicates better performance.

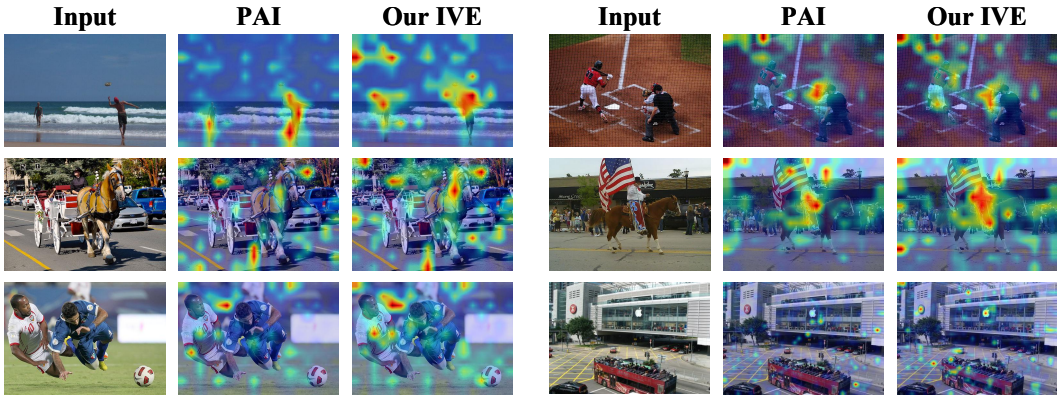


Figure 10: Visual-attention heatmaps for IVE and the amplification method PAI.

summarizes their reference signals and interventions. Our analysis focuses on persistence associated with relation errors and examines the effect of increased previous-step carryover at comparable attention concentration.

Attention Never Lie. Attention Never Lie Zhao et al. (2026) links hallucination to visual-attention defocus and examines attention differences between nearby and more distant output positions. Its correction mechanism trains a detector on attention features from text windows and uses detector scores to select candidate continuations. Attention history guides continuation selection, and the paper also evaluates combining its detector with visual-attention amplification. IVE applies token-specific attention history directly to visual-value reinforcement and attenuation within the base decoding procedure.

ADAPT. ADAPT Yao et al. (2026) characterizes attention degradation through fading visual focus, temporal rigidity, and spatial bias. It constructs a query-dependent visual anchor from early attention across multiple layers and uses degradation scores to trigger sparse, anchor-guided attention steering. This correction supports a training-free variant, while the full method additionally incorporates preference tuning. The early anchor guides attention toward stable visual evidence. IVE maintains an evolving reference for each token and conditions visual-value reinforcement and attenuation on its recent trajectory and inertia-selection history.

Visual Inertia terminology. GThinker Zhan et al. (2026) uses Visual Inertia to describe a model's tendency to retain its initial visual interpretation. Its cue-guided rethinking framework learns to revisit visual cues through supervised fine-tuning and reinforcement learning. In this paper, visual inertia denotes persistent reuse of previously dominant visual evidence across decoding, characterized through attention trajectories and carryover interventions.

Table 11: Reference signals and interventions in temporal attention methods.

Method	Reference or signal	Intervention
Attention Never Lie	Learned detection from windowed attention features	Detector-guided selection of candidate continuations
ADAPT	Attention degradation relative to an early visual anchor	Sparse anchor-guided attention steering, with preference tuning in the full method
IVE	EMA-relative attention rise and prior inertia-selection frequency	Paired reinforcement and conditional attenuation of visual values

Table 12: Comparison with temporal-attention methods on Reefknot across three backbones.

Backbone	Method	Y/N		MCQ	
		Cogn.↓	R_{score} ↑	**Cogn.**↓	R_{score} ↑
LLaVA-1.5	Vanilla	33.99	64.39	51.04	43.70
	ADAPT (training-free)	32.87	65.34	50.31	44.72
	Attention Never Lie†	32.36	65.90	49.86	45.35
	IVE	**31.11**	**67.21**	**49.08**	**45.99**
InstructBLIP	Vanilla	42.47	56.28	53.85	39.94
	ADAPT (training-free)	40.94	57.65	52.18	41.68
	Attention Never Lie†	41.12	57.52	51.91	41.92
	IVE	**39.31**	**59.09**	**50.87**	**42.94**
Qwen2.5-VL	Vanilla	20.58	72.07	49.69	44.91
	ADAPT (training-free)	19.22	73.31	48.03	46.72
	Attention Never Lie†	18.96	73.53	48.27	46.51
	IVE	**17.60**	**74.76**	**47.01**	**47.50**

†Uses a trained detector.

The relevance-signal comparison in Table 14 evaluates token-scoring statistics under matched selection budgets and fixed reinforcement and attenuation factors. This diagnostic examines the selection signal separately from the full method's thresholding and frequency-dependent attenuation.

C More Experimental Results and Analysis

We extend the main evaluations with type-wise hallucination results, controlled comparisons of visual reweighting, and object-hallucination tests on both binary questions and generated captions. Additional analyses examine broader multimodal performance and the effects of coefficient and layer choices.

C.1 Temporal-Attention Baselines on Reefknot

Table 12 lists Vanilla, the training-free variant of ADAPT Yao et al. (2026), Attention Never Lie Zhao et al. (2026), and IVE for comparison on Reefknot across the three evaluated backbones. We reproduce ADAPT-TF and Attention Never Lie following their official repositories and default inference protocols, without tuning on the Reefknot test set. ADAPT excludes preference-tuned components, while Attention Never Lie follows the repository-provided detector setup. All methods use matched backbones, prompts, decoding settings, and evaluation splits.

IVE achieves higher R_{score} and lower $Cogn.$ than both temporal baselines on all three backbones in both Y/N and MCQ. The R_{score} margin over the better-performing temporal baseline in each setting ranges from 0.64 to 1.44 points.

Table 13: Accuracy and binary F1 (%) across Hal-Eval hallucination types with LLaVA-1.5.

Type	Vanilla		+MemVR		+IVE (Ours)	
	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
Object	62.20	48.50	63.70	50.10	**64.10**	**50.80**
Attribute	62.20	48.50	63.50	49.80	**63.90**	**50.30**
Relation	52.65	35.82	55.83	39.04	**57.71**	**41.29**
Event	62.70	58.90	64.10	60.40	**64.80**	**61.20**

C.2 Hal-Eval Results Across Hallucination Types

Hal-Eval Jiang et al. (2024) separates hallucinations into Object, Attribute, Relation, and Event categories. Table 13 complements the three-backbone Relation results in the main text by reporting all four categories for LLaVA-1.5. This comparison asks whether the gains are larger for relations than for other hallucination types. IVE improves accuracy and F1 over both Vanilla and MemVR in all four categories, with its largest gains over Vanilla occurring on Relation. The type-wise pattern is consistent with the relation-focused motivation while retaining improvements on Object, Attribute, and Event.

Relation has the lowest Vanilla accuracy and F1 among the four categories. IVE narrows its performance gap to the other categories, although Relation remains the most difficult category on both metrics.

C.3 Component and Selection-Signal Ablations

Reinforcement and attenuation. Table 5(a) examines the two reweighting branches on LLaVA-1.5 using $l_s = 20$, $\alpha = 0.10$, $\gamma = 0.09$, and $\tau = 3.00$. Both single-branch conditions retain the full token-selection and history-update rules. Reinforce only sets the inertia-token multiplier to one, while Attenuate only sets the emergent-token multiplier to one. The enabled branch uses Equation 9, including the 1.10 reinforcement cap. Inertia counts continue to update when attenuation is disabled, and each condition maintains history along its own generated trajectory. Enabling both branches gives full IVE, while disabling both gives identity modulation. All conditions share evaluation samples, prompts, scoring, and deterministic decoding, with visual values reconstructed from fixed originals at each query.

Matched selection signals. To compare which tokens different signals prioritize, Table 14 uses a fixed-budget diagnostic on Reefknot Y/N with LLaVA-1.5. We average the emergent- and inertia-set sizes of full IVE over valid intervention queries on held-out validation data and round each mean to the nearest integer. This gives reinforcement and attenuation budgets of $K_+ = 8$ and $K_- = 32$, fixed before testing. At each query, score-based conditions reinforce the highest-scoring K_+ visual tokens and attenuate the lowest-scoring K_- among the remainder. Their multipliers are fixed at 1.10 and 0.90, respectively, relative to the original visual values, with all other entries restored to their originals. Evaluation inputs, decoding, and selection layer remain matched across signals. The IVE-score row uses this ranking and scaling protocol in place of the full method's threshold-based sets and frequency-dependent attenuation.

We rank unrounded scores and break exact ties by ascending original visual-token index in both selections. Random refresh uses an independent generator for each sample, initialized with seed $42+i$, where i is the fixed sample index. Each query draws a fresh visual-token permutation, assigning the first K_+ entries to reinforcement and the next K_- to attenuation. The generator advances across queries without reseeding, and the two selected sets remain disjoint.

Token-level scores. Cumulative attention measures past prominence, while previous-step differences retain the direction of the latest change. The IVE score in Equation 6 combines a short-term EMA reference with attention-scale normalization. For entropy and concentration, we use each token's contribution to the corresponding statistic. With $p_{t,j} = a_{t,j} / \sum_{k \in \mathcal{V}} a_{t,k}$, these scores are

$$s_{t,j}^{\text{ent}} = -p_{t,j} \log(p_{t,j} + \epsilon), \qquad s_{t,j}^{\text{conc}} = p_{t,j}^2, \qquad \epsilon = 10^{-6}.$$

Table 14: Selection-signal comparison on Reefknot Y/N with LLaVA-1.5. All non-Vanilla rows use matched token budgets and fixed $1.10/0.90$ reinforcement/attenuation factors. D_{attn} measures consecutive attention displacement.

Signal	Property	Perc.↓	Cogn.↓	$\mathbf{R_{score}}$ ↑	D_{attn}
Vanilla	–	37.67	33.99	64.39	41.52
Random refresh	No structure	37.44	33.71	64.64	42.18
Cumulative attention	Token-level, nondirectional	37.12	33.28	65.02	43.51
Previous-step difference	Token-level, directional	36.89	32.97	65.30	44.23
Entropy contribution	Token-level, current-only	37.31	33.44	64.85	42.87
HHI contribution	Token-level, current-only	37.38	33.51	64.78	42.63
Attention-sinking score	Token-level, nondirectional	37.08	33.19	65.09	43.74
IVE score (matched budget)	**Token-level, directional, smoothed**	**35.35**	**31.73**	**66.67**	**44.72**

Table 15: Comparison of attenuation rules under the full IVE selection protocol.

Attenuation rule	Perc.↓	Cogn.↓	$\mathbf{R_{score}}$ ↑	Mean attenuation
Fixed, strength-matched	35.21	31.60	66.81	0.061
History-frequency (IVE)	**34.92**	**31.11**	**67.21**	0.063

Both provide token-level rankings from the current visual-attention distribution. The concentration score preserves the ordering of $p_{t,j}$ and therefore implements current-saliency selection under the matched budget.

The attention-sinking control uses the sink-dimension activation metric of Kang et al. (2025). Let $\mathbf{h}_j \in \mathbb{R}^D$ be visual token j's prefill hidden state at the input to layer $l_s = 20$. For LLaVA-1.5-7B, the sink dimensions are $\mathcal{D}_{\text{sink}} = \{1415, 2533\}$, giving

$$\phi_j = \frac{\max_{d \in \mathcal{D}_{\text{sink}}} |\mathbf{h}_j[d]|}{\sqrt{D^{-1} \sum_{d=1}^{D} \mathbf{h}_j[d]^2}}, \qquad s_{t,j}^{\text{sink}} = -\phi_j.$$

The negative sign assigns higher attenuation priority to tokens with larger sink activations. These scores remain fixed across output queries and use the same budgets and value multipliers as the other ranking controls.

All tested signals improve R_{score} over Vanilla, with previous-step difference performing best among the alternatives. The IVE score achieves the highest R_{score} and the lowest hallucination rates in both reported categories under matched selection budgets and scaling factors. This comparison evaluates the ranking signal under fixed selection budgets and modulation strengths.

The history-conditioned score also produces the largest consecutive attention displacement in this comparison. This statistic describes the attention changes accompanying its prediction gains, complementing the benchmark metrics that evaluate relational grounding.

Attenuation rule. To examine whether attenuation benefits from prior inertia-selection frequency, we retain the full IVE token-selection rules, emergent-token reinforcement, and history updates, replacing only the inertia-token attenuation rule with a uniform coefficient. This coefficient is fixed before testing to the mean attenuation measured on held-out validation data. Table 15 also reports the mean applied attenuation, $1 - m_{t,j}$, over selected inertia tokens on the test set.

History-frequency attenuation lowers both hallucination rates and improves R_{score} over the fixed alternative (Table 15). Test-set mean attenuation is similar across conditions, with a slightly higher value for IVE. The comparison favors assigning attenuation according to prior inertia selections over applying one coefficient to every selected token under the evaluated protocol.

Raw and visual-normalized attention. We compare raw attention $a_{t,j}$ with visual-normalized attention $p_{t,j}$ to examine whether retaining total visual-attention strength benefits token selection. Each condition maintains an EMA in the corresponding input space and uses the fixed $K_+ = 8$,

Table 16: Comparison of attention signals under matched selection budgets and scaling factors.

Score input	Perc.↓	Cogn.↓	$\mathbf{R_{score}}$ ↑
Vanilla (no intervention)	37.67	33.99	64.39
Visual-normalized attention ($p_{t,j}$)	35.81	31.99	66.32
Raw attention ($a_{t,j}$; IVE score)	**35.35**	**31.73**	**66.67**

Table 17: Matched random perturbation controls and IVE on Reefknot Y/N. The token control matches the token budget, and the value control matches the modulation magnitude.

Method	D_{attn}	Cogn.↓	R_{score} ↑
Vanilla	41.52	33.99	64.39
+Random token perturbation (same budget)	42.18	33.84	64.52
+Random value modulation (same magnitude)	42.63	33.71	64.60
+IVE	**47.21**	**31.11**	**67.21**

$K_{-} = 32$ selection budgets and $1.10/0.90$ scaling factors from Table 14. The raw-attention row in Table 16 reports the same matched-budget IVE-score condition.

Both inputs reduce hallucination rates relative to Vanilla, with raw attention providing a further improvement (Table 16). The normalized-input result shows that the paired intervention remains effective when token rankings use only within-image allocation history. Under the tested scoring rule, retaining overall visual-attention strength alongside relative allocation yields an additional benefit with the same selection budgets and scaling factors.

C.4 Matched Random Controls

The preceding comparison varies token rankings under fixed budgets and scaling factors. We separately compare full IVE with random token perturbation at the same token budget and random visual-value modulation at the same magnitude. These controls examine whether unstructured perturbations produce gains comparable to history-conditioned selection and reweighting on Reefknot.

The two matching conditions address different aspects of the intervention. Matching the token budget keeps the number of affected tokens comparable, while matching the modulation magnitude controls the size of the value perturbation. They provide reference points for unstructured changes under comparable intervention budgets.

Both random controls modestly improve R_{score} over Vanilla, whereas IVE gives a larger improvement and a lower $Cogn.$ hallucination rate (Table 17). The comparison favors structured selection and reweighting over the matched random perturbations. Random perturbations also slightly increase consecutive attention displacement, but their prediction gains remain smaller than those of IVE.

C.5 Additional Results on MME

We report selected MME Fu et al. (2025) categories to examine whether the relation-oriented intervention preserves performance on other multimodal tasks. The comparison covers the same three backbones and reports category-specific scores. The categories involve different demands on the visual input, including reading textual or numerical content and evaluating statements about that content. A category-level view shows how performance changes across these tasks, with each column expressed in its benchmark-specific score.

IVE improves every selected category over Vanilla across all three backbones and remains competitive with the strongest baseline results. The gains extend across the reported tasks, supporting performance preservation beyond the primary relation-hallucination evaluations.

The baseline methods exhibit different profiles across the categories: some strengthen commonsense or numerical performance while falling below Vanilla on text or code, whereas IVE remains above

Table 18: Scores on selected MME categories across three MLLMs. Com., Num., Text., and Code denote commonsense reasoning, numerical calculation, text translation, and code reasoning, respectively.

Model	Method	Com.↑	Num.↑	Text.↑	Code.↑
LLaVA-1.5	Vanilla	87.86	105.0	97.50	67.50
	+VCD	89.29	101.5	98.00	65.00
	+OPERA	92.14	107.0	92.50	58.00
	+PAI	91.43	106.5	91.50	56.50
	+HELPD	91.79	106.0	95.50	62.50
	+DeCo	90.71	102.0	100.0	70.00
	+VASparse	92.86	108.0	93.00	59.50
	+MemVR	93.57	109.0	101.5	70.50
	+IVE (Ours)	**94.64**	**110.0**	**103.0**	**72.00**
	Δ vs. Vanilla	**+6.79**	**+5.00**	**+5.50**	**+4.50**
InstructBLIP	Vanilla	117.9	40.00	92.50	55.00
	+VCD	118.6	45.00	94.00	57.50
	+OPERA	121.4	47.50	89.50	52.00
	+PAI	120.7	47.00	89.00	51.50
	+HELPD	121.8	46.50	94.00	56.50
	+DeCo	120.4	46.00	96.00	59.50
	+VASparse	122.1	48.00	91.00	53.50
	+MemVR	122.9	47.50	97.00	60.00
	+IVE (Ours)	**124.3**	**49.00**	**98.00**	**61.50**
	Δ vs. Vanilla	**+6.43**	**+9.00**	**+5.50**	**+6.50**
Qwen2.5-VL-Instruct	Vanilla	100.7	52.50	92.50	47.50
	+VCD	101.4	55.00	94.00	49.50
	+OPERA	103.6	56.00	89.50	46.50
	+PAI	102.9	55.50	89.00	46.00
	+HELPD	103.1	55.00	93.50	50.00
	+DeCo	102.3	54.50	96.00	52.50
	+VASparse	104.3	56.50	91.00	48.00
	+MemVR	104.6	57.00	97.00	53.00
	+IVE (Ours)	**106.4**	**58.00**	**98.00**	**54.50**
	Δ vs. Vanilla	**+5.71**	**+5.50**	**+5.50**	**+7.00**

Vanilla in all reported columns. The relation benchmarks provide the direct evaluation of the paper's primary objective, and these category-level results assess compatibility with other multimodal tasks.

C.6 Radar Chart Results

Figure 9 brings together the relation-focused results from Reefknot Zheng et al. (2025) and HalEval Jiang et al. (2024) for LLaVA-1.5, InstructBLIP Dai et al. (2023), and Qwen2.5-VL-Instruct Bai et al. (2025). The five axes cover Reefknot Y/N Cogn. and R_{score}, Reefknot MCQ R_{score}, and HalEval Relation accuracy and F1. IVE exhibits the strongest overall profile across the three backbones, complementing the detailed results in the main tables.

C.7 Heatmap Comparison

Figure 10 compares IVE with the amplification method PAI Liu et al. (2024c). In the illustrated response, IVE attends across several regions involved in the queried relation, whereas PAI remains concentrated within a smaller area. The contrast appears across successive output positions, providing a spatial view of the different attention patterns accompanying history-conditioned reweighting and amplification.

Reading the maps in temporal order reveals whether strongly attended regions continue to dominate across output positions. The displayed contrast captures the evolution of spatial emphasis and complements the distance summaries by retaining the image locations associated with each attention pattern.

Table 19: **Object-hallucination results on POPE.** Accuracy and F1 (%) across the Random, Popular, and Adversarial settings. Parenthetical values denote improvements over Vanilla.

Category	Method	LLaVA-1.5		InstructBLIP		Qwen2.5-VL-Instruct	
		Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
Random	Vanilla	83.29 (+0.00)	81.33 (+0.00)	80.71 (+0.00)	80.41 (+0.00)	85.27 (+0.00)	85.30 (+0.00)
	+VCD	86.27 (+2.98)	86.34 (+5.01)	84.53 (+3.82)	83.68 (+3.27)	88.60 (+3.33)	88.67 (+3.37)
	+OPERA	88.30 (+5.01)	88.36 (+7.03)	85.80 (+5.09)	85.89 (+5.48)	89.87 (+4.60)	89.93 (+4.63)
	+DeCo	88.13 (+4.84)	88.20 (+6.87)	84.87 (+4.16)	84.79 (+4.38)	89.63 (+4.36)	89.68 (+4.38)
	+PAI	88.42 (+5.13)	88.51 (+7.18)	85.20 (+4.49)	85.29 (+4.88)	89.77 (+4.50)	89.82 (+4.52)
	+VASparse	88.57 (+5.28)	88.63 (+7.30)	86.20 (+5.49)	86.28 (+5.87)	89.97 (+4.70)	90.02 (+4.72)
	+MemVR	88.83 (+5.54)	88.89 (+7.56)	86.57 (+5.86)	86.64 (+6.23)	90.40 (+5.13)	90.45 (+5.15)
	+Ours	**90.53** (+7.24)	**90.94** (+9.61)	**90.20** (+9.49)	**90.16** (+9.75)	**92.10** (+6.83)	**92.15** (+6.85)
Popular	Vanilla	81.88 (+0.00)	80.06 (+0.00)	78.22 (+0.00)	78.36 (+0.00)	83.73 (+0.00)	83.83 (+0.00)
	+VCD	83.67 (+1.79)	83.77 (+3.71)	80.70 (+2.48)	80.80 (+2.44)	85.03 (+1.30)	85.12 (+1.29)
	+OPERA	84.60 (+2.72)	84.69 (+4.63)	81.97 (+3.75)	82.07 (+3.71)	86.33 (+2.60)	86.41 (+2.58)
	+DeCo	83.90 (+2.02)	83.99 (+3.93)	81.33 (+3.11)	81.44 (+3.08)	85.47 (+1.74)	85.54 (+1.71)
	+PAI	84.13 (+2.25)	84.23 (+4.17)	81.60 (+3.38)	81.71 (+3.35)	85.91 (+2.18)	85.98 (+2.15)
	+VASparse	84.77 (+2.89)	84.84 (+4.78)	82.17 (+3.95)	82.27 (+3.91)	86.57 (+2.84)	86.64 (+2.81)
	+MemVR	85.00 (+3.12)	85.09 (+5.03)	82.50 (+4.28)	82.60 (+4.24)	86.83 (+3.10)	86.91 (+3.08)
	+Ours	**86.90** (+5.02)	**87.01** (+6.95)	**84.73** (+6.51)	**85.45** (+7.09)	**89.70** (+5.97)	**89.75** (+5.92)
Adversarial	Vanilla	76.47 (+0.00)	76.61 (+0.00)	74.37 (+0.00)	74.51 (+0.00)	81.87 (+0.00)	81.98 (+0.00)
	+VCD	77.20 (+0.73)	77.34 (+0.73)	76.80 (+2.43)	76.94 (+2.43)	83.47 (+1.60)	83.57 (+1.59)
	+OPERA	78.00 (+1.53)	78.13 (+1.52)	77.50 (+3.13)	77.63 (+3.12)	84.67 (+2.80)	84.76 (+2.78)
	+DeCo	77.57 (+1.10)	77.68 (+1.07)	77.07 (+2.70)	77.21 (+2.70)	83.87 (+2.00)	83.97 (+1.99)
	+PAI	77.83 (+1.36)	77.96 (+1.35)	77.37 (+3.00)	77.49 (+2.98)	84.30 (+2.43)	84.39 (+2.41)
	+VASparse	78.30 (+1.83)	78.42 (+1.81)	78.13 (+3.76)	78.53 (+4.02)	84.87 (+3.00)	84.93 (+2.95)
	+MemVR	78.60 (+2.13)	78.73 (+2.12)	78.40 (+4.03)	78.42 (+3.91)	85.47 (+3.60)	85.56 (+3.58)
	+Ours	**80.53** (+4.06)	**82.17** (+5.56)	**80.50** (+6.13)	**81.70** (+7.19)	**88.17** (+6.30)	**88.23** (+6.25)

C.8 Object Hallucination on POPE

POPE Li et al. (2023b) tests whether history-conditioned contribution reweighting also benefits object-existence judgments. Table 19 reports accuracy and F1 across the Random, Popular, and Adversarial settings for all three backbones. These settings vary the object queries, allowing the comparison to extend beyond one collection of candidate objects. Accuracy summarizes the binary decisions, while F1 reflects the balance between precision and recall for positive predictions. Parenthetical values report changes relative to Vanilla within the same backbone and setting.

IVE improves both metrics over Vanilla and the competing methods in each setting. Scores generally decrease from Random to Popular and Adversarial queries across the compared methods, while the advantage of IVE persists. This evaluation complements Reefknot by assessing temporal-relevance-based visual contribution updates on object-existence decisions.

C.9 Object Hallucination in Generated Captions

Object-existence questions constrain the model to a short response, while captioning allows it to introduce objects throughout generation. We therefore report CHAIR Rohrbach et al. (2018) results on a COCO subset. CHAIR_i measures the fraction of mentioned objects that are hallucinated, and CHAIR_s measures the fraction of captions containing at least one hallucinated object. We report both overall measures and stratify CHAIR_i by scene density and response length.

The density and length groups provide two complementary views of the captioning results. Density groups distinguish scenes by object count, and length groups distinguish generated responses by token count. These marginal stratifications show how each method behaves in more crowded scenes and longer responses.

Hallucination rates rise with scene density and response length for all compared methods (Table 20). IVE yields the lowest overall rates and remains below the baselines in each density and length group, including dense scenes and long responses. These results extend the object-hallucination comparison to the unsupported object mentions introduced during caption generation.

Table 20: CHAIR results (%) overall and by scene density and response length. Density counts objects, and length counts output tokens. Lower values are better.

Method	Overall		Scene density (CHAIR_i)			Response length (CHAIR_i)		
	$\mathrm{CHAIR}_i \downarrow$	$\mathrm{CHAIR}_s \downarrow$	Sparse ≤ 5	Medium 6–10	Dense ≥ 11	Short ≤ 50	Medium 51–100	Long > 100
Vanilla	52.4	68.1	38.2	54.7	71.3	44.3	53.8	61.9
+PAI	51.8	67.5	37.9	54.1	71.6	44.1	53.4	62.7
+MemVR	51.2	66.9	37.5	53.8	70.9	43.7	52.9	61.4
+IVE	**50.3**	**65.8**	**37.1**	**53.2**	**70.8**	**43.2**	**52.4**	**61.1**

IVE maintains lower hallucination rates in the densest scenes and longest responses, with smaller gains over Vanilla than in the other groups. The benefit concerns the reliability of the object content that is generated, complementing the object-existence evaluation on POPE.

C.10 Full Configuration Sensitivity

Table 21 reports coefficient sensitivity across the three backbones on Reefknot and POPE. Each sweep varies one coefficient around the validation-selected configuration while keeping the remaining settings fixed. The attenuation coefficient α bounds the reduction of persistent contributions at each update, γ controls smoothing of the recent attention history, and τ sets the threshold for identifying emergent tokens. The main text retains the LLaVA-1.5 subset, while the full table shows the corresponding local variations on the other backbones.

These coefficients influence different parts of the update. Changing α adjusts attenuation after an inertia token has been selected. Changing γ modifies the historical reference against which current attention is compared, with larger values assigning more weight to the older reference. Changing τ alters the threshold for the emergent-token branch. Examining them separately helps identify whether performance is sensitive to the magnitude of reweighting, the construction of attention history, or the boundary used for token selection.

Every tested setting improves over Vanilla on both Reefknot and POPE, while the selected settings yield the lowest $Cogn.$ rates within each backbone's sweeps. Gains persist throughout the tested neighborhoods, with their magnitude varying across configurations. This analysis examines neighborhoods around the tuned settings, complementing the shared-configuration comparison in the main text.

The relation-oriented metric varies non-monotonically across each coefficient sweep. The selected settings balance the effects of the relevance reference and contribution adjustment within the tested neighborhoods. The POPE columns show how the same local changes affect object-level evaluation. The unified-configuration experiment applies one shared choice across backbones, complementing these model-specific local sweeps.

C.11 Ablation on Token Selection Layers

The selection layer determines where IVE reads the attention signal, while value recalibration continues to apply across transformer layers. Table 22 varies the selection layer while keeping the remaining method fixed. The chosen mid-to-late layers yield the lowest Reefknot Zheng et al. (2025) hallucination rates for each backbone and retain competitive accuracy and F1 on POPE Li et al. (2023b). POPE F1 can peak at a neighboring layer. Earlier layers also improve over Vanilla, with smaller gains on Reefknot.

Changing the selection layer changes both the attention history and the point at which recalibration enters the current forward pass. Cache updates still span all layers. The results favor the chosen layers for relation evaluation, with gains also retained at neighboring layers.

Table 21: Full coefficient sensitivity of IVE across three backbones on Reefknot and POPE.

Model	Coeff.	Val.	Reefknot		POPE	
			Perc.↓	Cogn.↓	Acc.↑	F1↑
LLaVA-1.5	α	0.05	36.84	32.72	89.20	89.27
		0.10	34.92	31.11	90.53	90.94
		0.15	36.60	32.43	89.57	89.62
	γ	0.08	36.56	32.48	89.77	89.82
		0.09	34.92	31.11	90.53	90.94
		0.10	35.88	32.07	90.53	90.94
	τ	2.95	36.72	32.44	89.63	89.68
		3.00	34.92	31.11	90.53	90.94
		3.05	35.93	32.10	90.27	90.32
InstructBLIP	α	0.05	44.28	41.73	87.23	87.31
		0.10	42.94	39.31	90.20	90.16
		0.15	43.88	41.34	88.93	88.95
	γ	0.09	44.16	41.57	89.97	90.02
		0.10	43.70	40.79	90.07	90.12
		0.11	42.94	39.31	90.20	90.16
	τ	3.00	43.75	40.60	90.20	90.16
		3.05	42.94	39.31	90.20	90.16
		3.10	43.88	40.99	90.10	90.11
Qwen2.5-VL-Instruct	α	0.10	36.18	19.22	91.80	91.85
		0.15	34.91	17.60	92.10	92.15
		0.20	36.11	19.08	91.93	91.98
	γ	0.11	36.12	19.10	91.73	91.78
		0.12	34.91	17.60	92.10	92.15
		0.13	36.20	19.18	91.68	91.74
	τ	3.05	36.08	19.01	91.88	91.94
		3.10	34.91	17.60	92.10	92.15
		3.15	36.14	19.12	91.79	91.84

Table 22: Ablation study on token selection layers of IVE.

Model	Layer	Reefknot		POPE	
		Perc.↓	Cogn.↓	Acc.↑	F1↑
LLaVA-1.5	16	36.14	32.36	90.18	90.24
	18	35.96	32.18	90.41	90.48
	20	**34.92**	**31.11**	**90.53**	**90.94**
InstructBLIP	16	44.02	40.97	89.94	90.01
	18	43.84	40.71	90.18	90.26
	20	**42.94**	**39.31**	**90.20**	**90.16**
Qwen2.5-VL-Instruct	14	36.09	19.14	91.72	91.76
	16	35.94	18.95	91.97	92.02
	18	**34.91**	**17.60**	**92.10**	**92.15**

D CASE STUDIES

Figures 11–18 present Cases 1–8 comparing Vanilla and IVE. Cases 1–5 are drawn from Reefknot Zheng et al. (2025), Case 6 from MME Fu et al. (2025), Case 7 from POPE Li et al. (2023b), and Case 8 from MMBench Liu et al. (2024d). Each comparison uses LLaVA-1.5 Liu et al. (2024a) with the same visual input and prompt.

Cases 1–5 illustrate unsupported relations predicted by the baseline, showing how responses can mention the correct entities yet misrepresent their relationship. The paired responses highlight

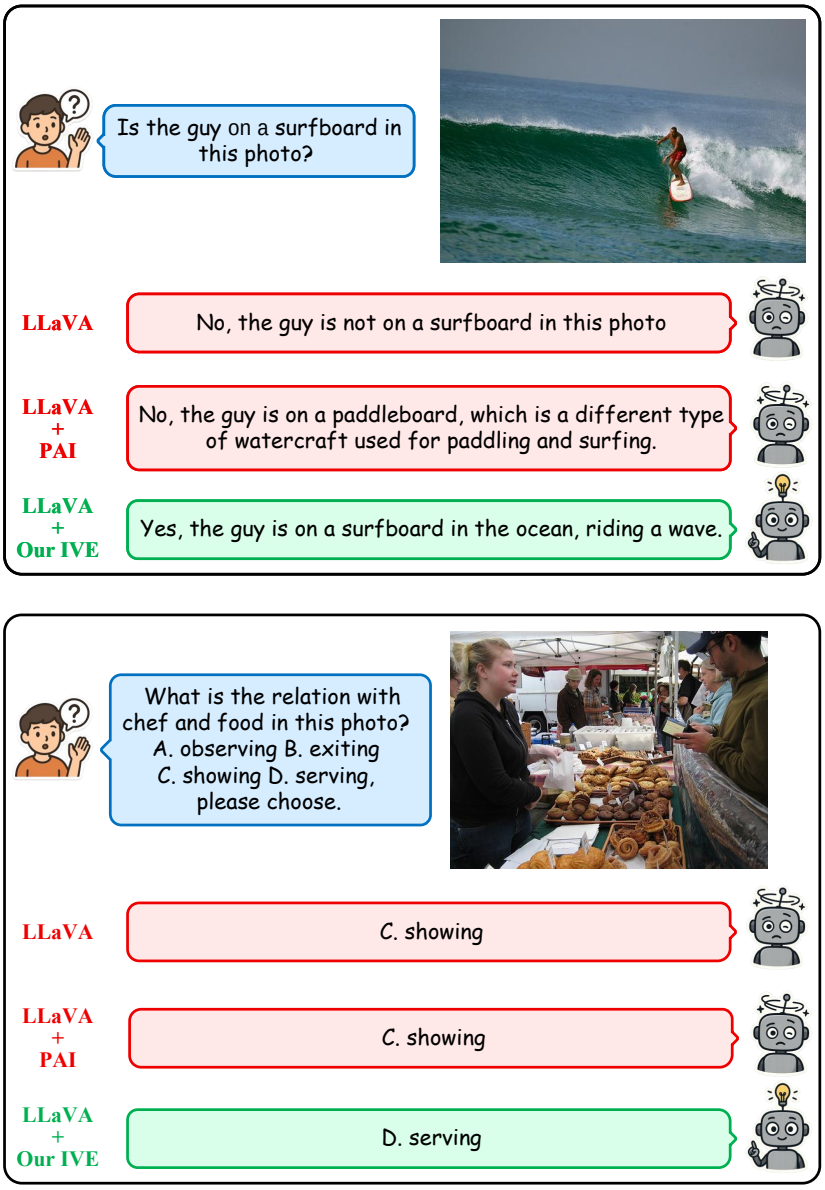


Figure 11: Case 1.

differences in the relation-bearing assertions, complementing the aggregate relation-hallucination results.

Case 7 concerns whether a named object is present, contrasting the methods' object-existence decisions. Cases 6 and 8 illustrate differences in responses to the task-specific questions of the broader capability evaluations.

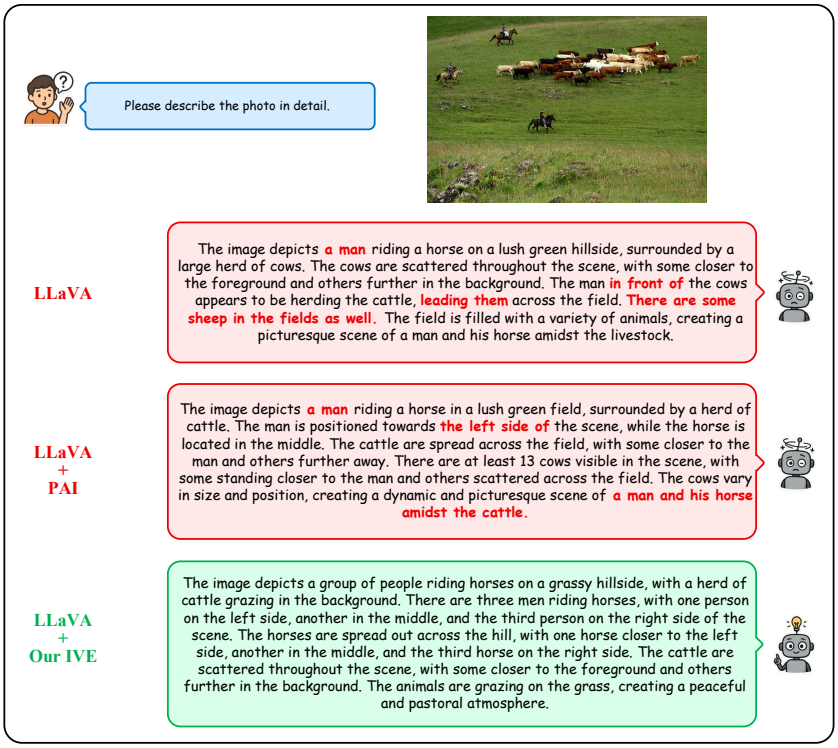


Figure 12: Case 2.

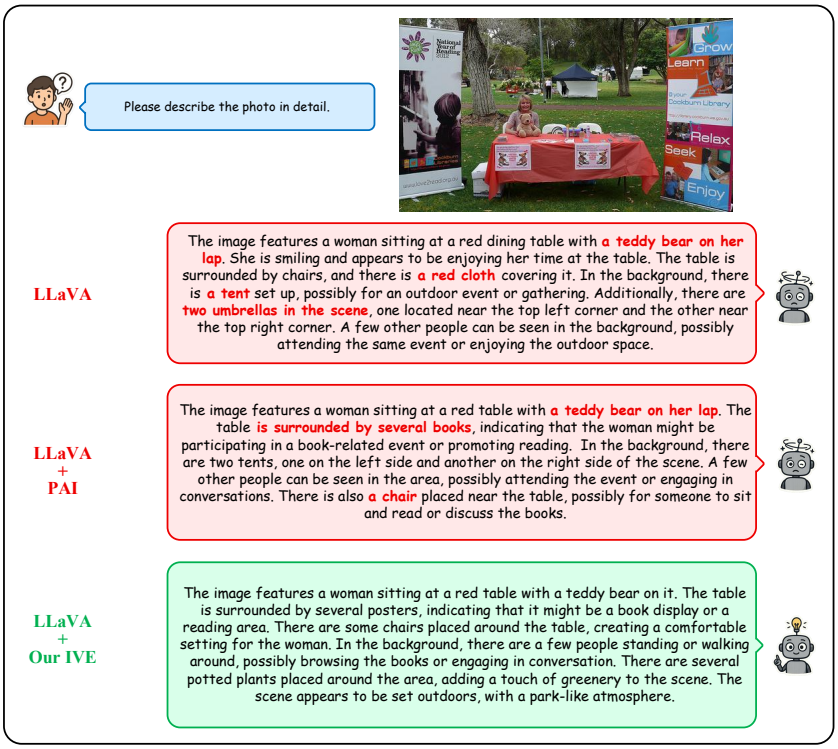


Figure 13: Case 3.

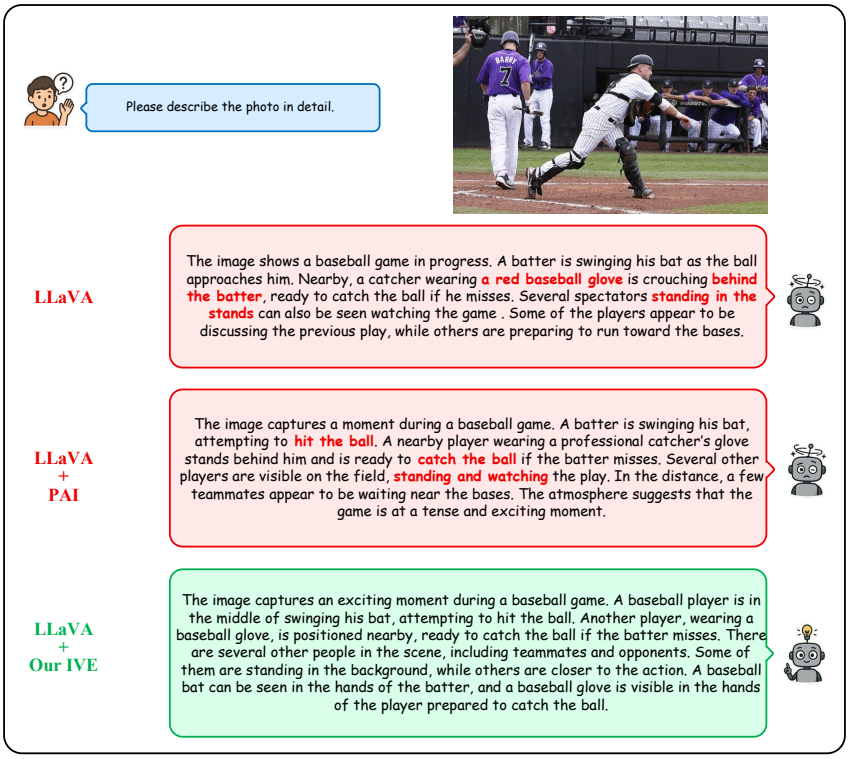


Figure 14: Case 4.

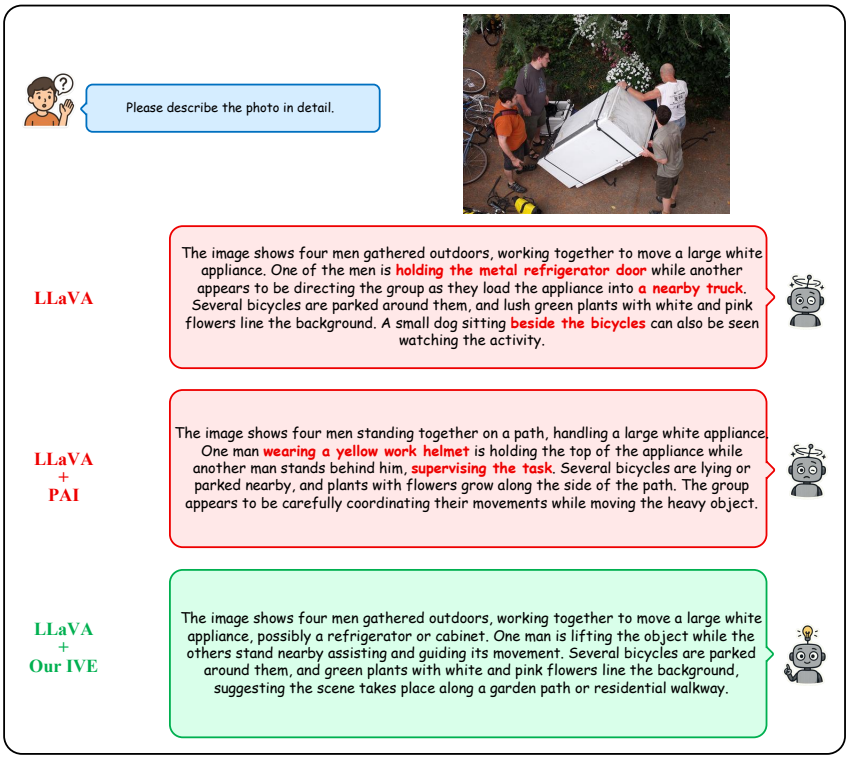
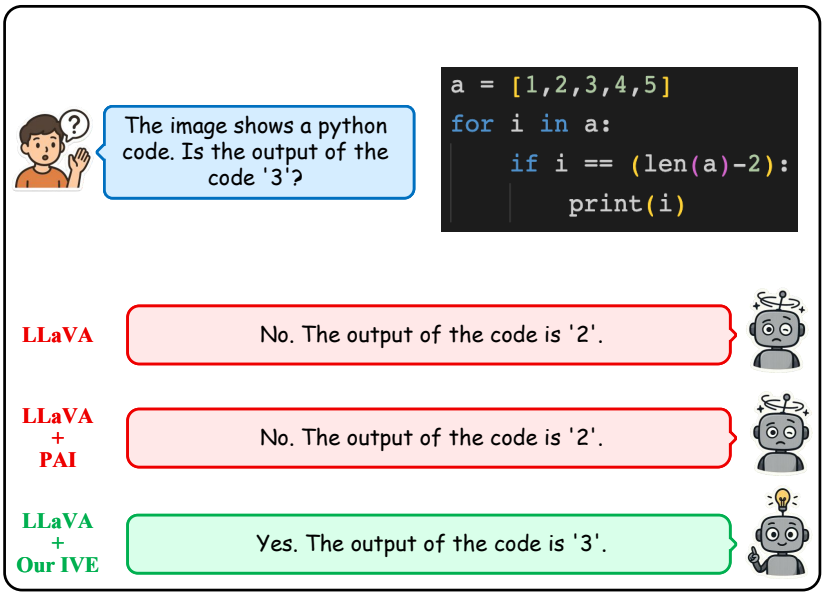


Figure 15: Case 5.



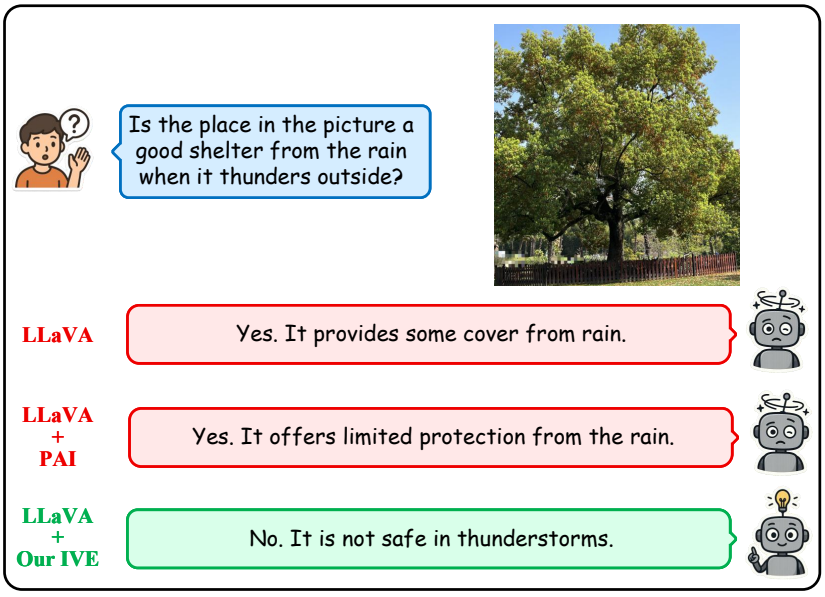


Figure 16: Case 6.

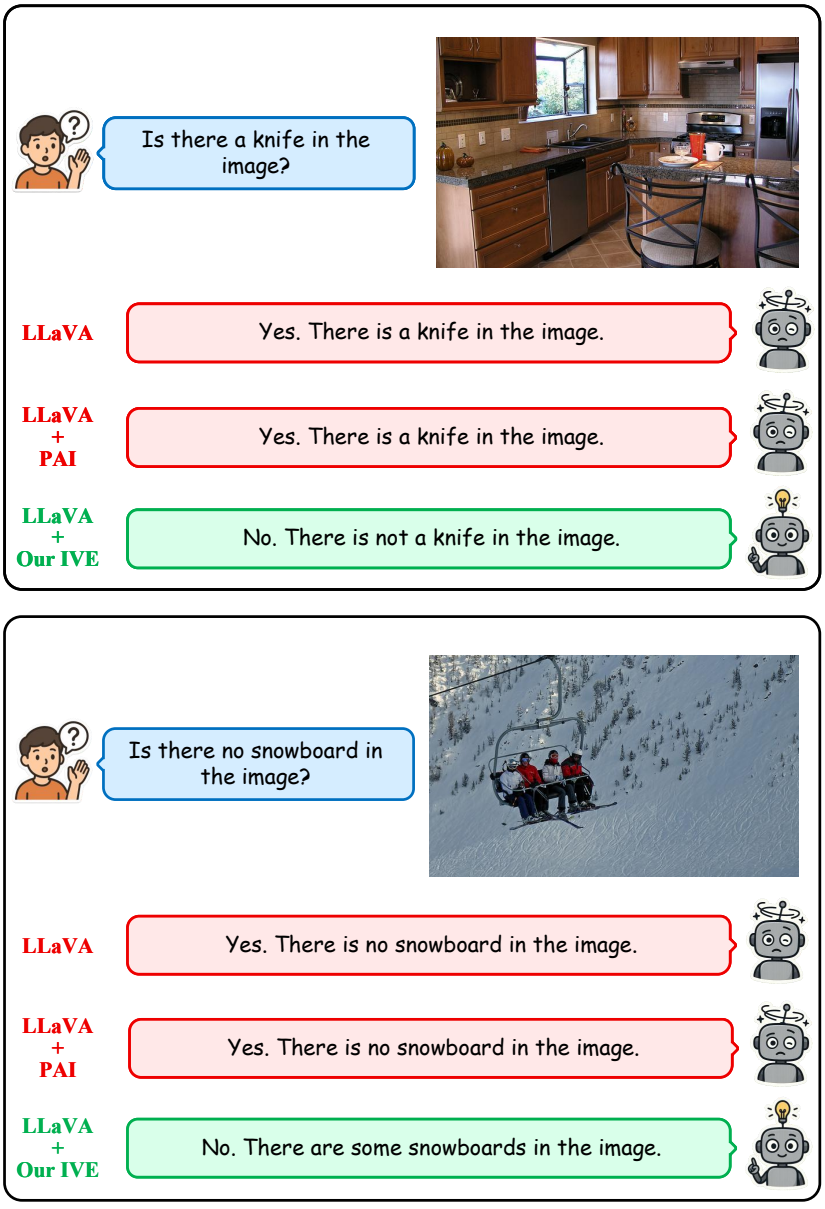


Figure 17: Case 7.

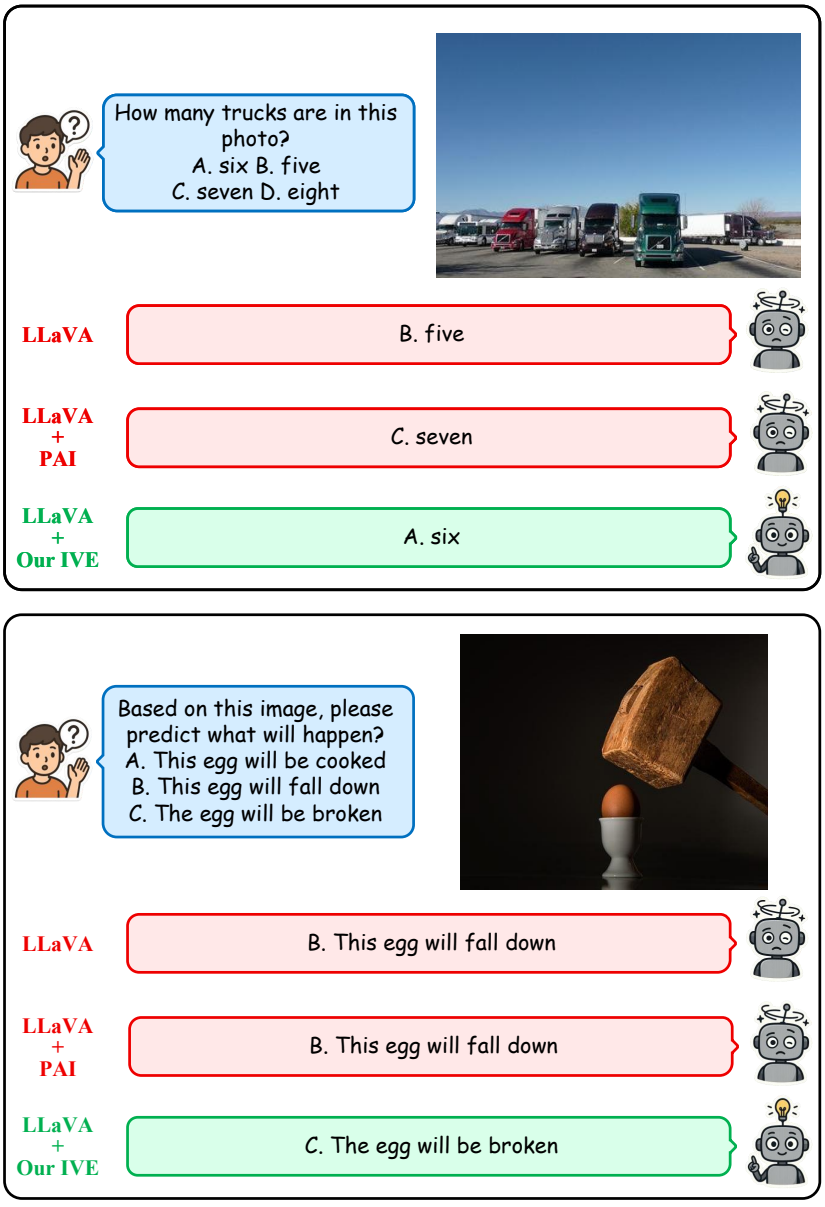


Figure 18: Case 8.