Attention at Rest Stays at Rest: Breaking Visual Inertia for Cognitive Hallucination Mitigation

Anonymous ACL submission

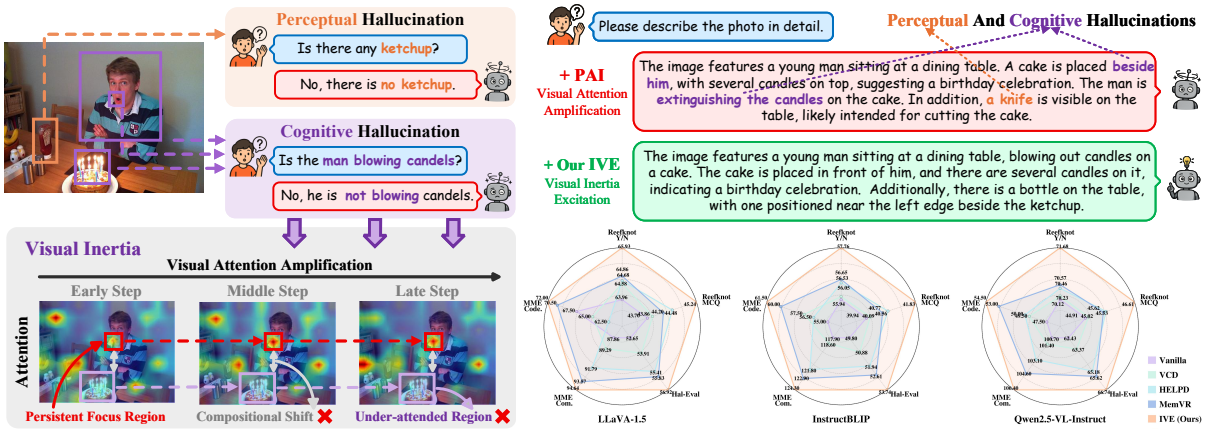


Figure 1: Overview of the cognitive hallucinations and their mitigation via our IVE. **Upper Left:** Perceptual and cognitive hallucinations in MLLMs. **Upper Right:** IVE effectively reduces cognitive hallucinations by exciting inertial visual attention. **Bottom Left:** Visual Inertia: visual attention remains persistently focused, hindering compositional reasoning. **Bottom Right:** Comparison with other methods across multiple models and benchmarks.

Abstract

Like a body at rest that stays at rest, we find that visual attention in multimodal large language models (MLLMs) exhibits pronounced inertia, remaining largely static once settled during early decoding steps and failing to support the compositional understanding required for cognitive inference. While existing hallucination mitigation methods mainly target perceptual hallucinations concerning object existence or attributes, they remain inadequate for such cognitive hallucinations that require inter-object relational deduction. Through token-wise analysis, we identify **visual inertia** as a contributing factor: attention to semantically critical regions remains persistently focused and fails to dynamically support relational inference. We thereby propose Inertia-aware Visual Excitation (IVE) that breaks this inertial pattern by modeling cognitive inference as the dynamic responsiveness of visual attention. Specifically, IVE selects visual tokens that are dynamically emerging relative to historical attention trends while distinguishing tokens exhibiting inertial behavior. To further facilitate compositional inference, IVE introduces an inertia-aware penalty that discourages over-concentration and limits the persistence of attention within localized regions. Extensive experiments show the effectiveness of IVE across various MLLMs and benchmarks without additional training.

1 Introduction

Multimodal Large Language Models (MLLMs) (Liu et al., 2024c,b; Li et al., 2023a; Zhu et al., 2024; Bai et al., 2025; Touvron et al., 2023a) have achieved remarkable progress in multimodal understanding and generation (Zhao et al., 2024a,b; Xie et al., 2024; Zhuang et al., 2024; Yin et al., 2025b). Despite these advances, recent studies (Dai et al., 2023b; Li et al., 2023b; Guan et al., 2024) show that even advanced MLLMs may produce descriptions inconsistent with the input image, a phenomenon known as MLLM hallucinations. Such hallucinations severely undermine model reliability and limit practical deployment (Wang et al., 2024; Hu et al., 2023; Mai et al., 2023; Liu et al., 2023).

Existing methods for mitigating MLLM hallucinations are either training-based (Yu et al., 2024a; Yue et al., 2024; Kim et al., 2024) or training-free (Huang et al., 2024; Wang et al., 2025; Leng et al., 2024; Yu et al., 2025; Zhuang et al., 2025). However, these methods predominantly address perceptual hallucinations concerning object existence or basic attributes. As recent evaluations (Fu et al., 2025; Liu et al., 2024e) show, reliable deployment also requires higher-order cognition built on perception, especially in scenarios involving rela-

tional inference (Chen et al., 2024; Mai et al., 2023; Wu et al., 2023). This exposes a harder failure mode, namely cognitive hallucination, where models fabricate inter-object relationships even after correctly recognizing individual objects (Figure 1 upper left). Such hallucinations are difficult to detect through object-level verification and remain largely unaddressed by existing methods.

To develop targeted solutions, it is necessary to understand what causes cognitive hallucinations internally. Prior work on perceptual hallucinations has traced their origins to the model's attention over visual tokens, finding that insufficient attention to task-relevant regions leads to fabricated objects (Yin et al., 2025a; Liu et al., 2024d; Yu et al., 2025). A natural question arises: **do cognitive hallucinations share the same root cause?** We argue that they do not, because cognitive inference requires not merely attending to individual objects, but dynamically integrating cues across multiple regions to capture inter-object relationships. This motivates our analysis of attention during cognitive tasks (cf. Section 2), where we uncover **Visual Inertia**: as the model generates output, attention to key visual regions remains persistently focused, failing to dynamically shift toward the compositional semantics required for relational inference (Figure 1 bottom left). This finding helps explain why naively amplifying visual attention (Liu et al., 2024d), though effective for perceptual hallucinations, yields diminishing returns for cognitive ones by reinforcing static rather than dynamic shifts.

To address this limitation, we propose Inertia-aware Visual Excitation (IVE) to mitigate cognitive hallucinations. Unlike previous methods that simply amplify attention, IVE explicitly breaks the inertial pattern of visual attention, reactivating it to emphasize overlooked visual regions that require compositional understanding of inter-object relationships. Specifically, we first identify visual tokens that dynamically emerge relative to historical attention trends while separating tokens exhibiting inertial behavior, enabling the model to focus on newly relevant regions. To further prevent persistent over-concentration on localized areas, IVE applies an Inertia-aware Attention Penalty that adaptively penalizes attention to tokens with sustained inertia, promoting the capture of compositional visual semantics necessary for relational inference. Extensive experiments demonstrate that IVE remains effective across various MLLMs and multiple benchmarks without additional training.

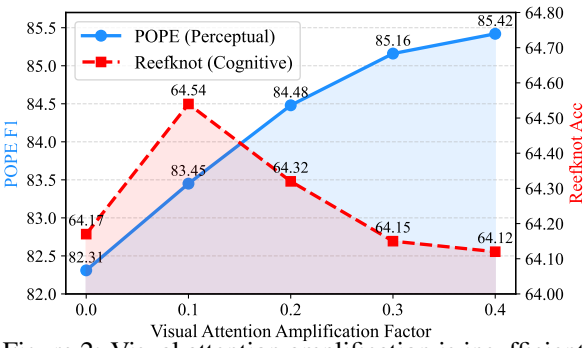


Figure 2: Visual attention amplification is insufficient for cognitive hallucination mitigation.

2 Preliminary and Motivation

2.1 Preliminary

A multimodal large language model (MLLM) consists of a visual encoder, a projector, and a large language model (LLM). Given a system prompt $\mathcal{S}$, a visual input $\mathcal{V}$, and a user instruction $\mathcal{I}$, the MLLM generates a response $\mathcal{Y}$:

$$\mathcal{F} : (\mathcal{S}, \mathcal{V}, \mathcal{I}) \mapsto \mathcal{Y} \tag{1}$$

The visual encoder and projector transform the image into text-aligned visual representations, which the LLM consumes in autoregressive decoding.

Tokens in MLLMs. MLLMs encode each modality as a token sequence, including system tokens $\mathbf{T}_s$, visual tokens $\mathbf{T}_v$, and instruction tokens $\mathbf{T}_i$, which are concatenated as

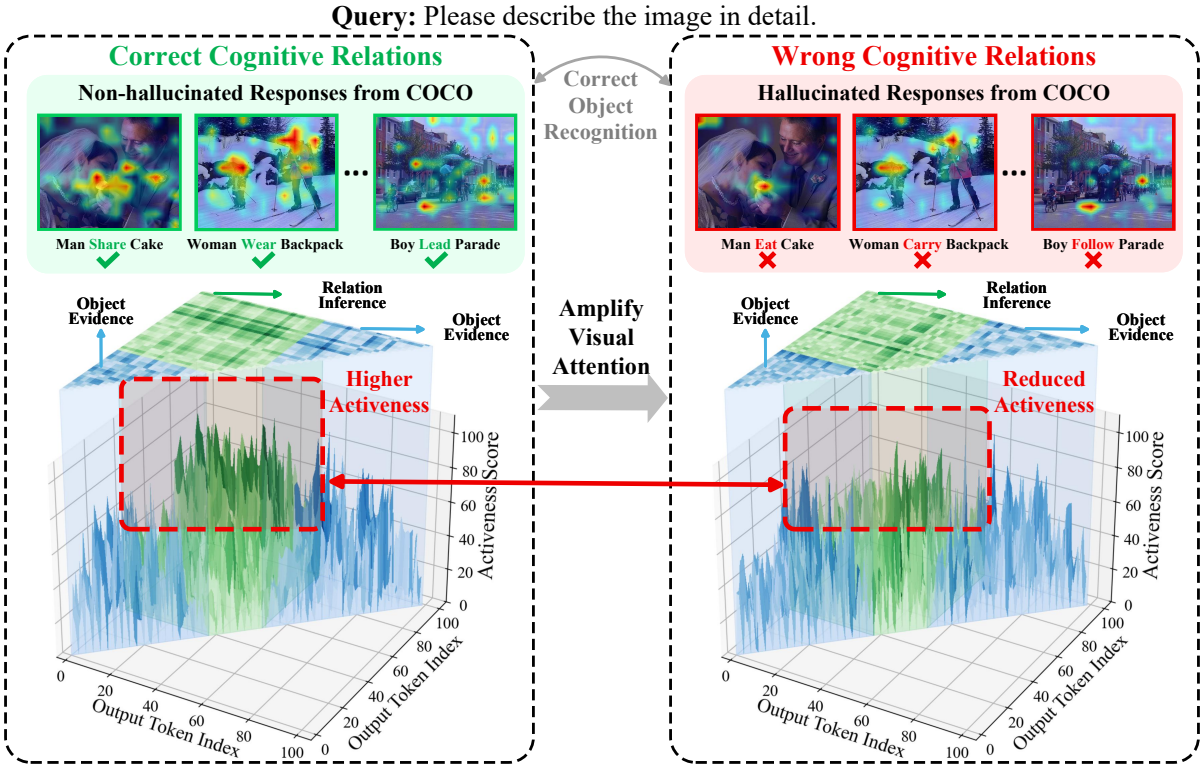
$$\mathbf{T} = [\mathbf{T}_s, \mathbf{T}_v, \mathbf{T}_i] \tag{2}$$

Their interactions are modeled by multi-head self-attention, which produce attention matrix $\mathbf{A}$.

2.2 Visual Inertia in Cognitive Hallucinations

Recent visual attention amplification studies (Yin et al., 2025a; Liu et al., 2024d; Yu et al., 2025) have identified imbalanced token attention as a primary cause of perceptual hallucinations. We investigate whether cognitive hallucinations arise from the same mechanism. Our analysis reveals an additional phenomenon we term **Visual Inertia**, where attention remains confined to limited regions and fails to support relational inference.

Necessary but Insufficient Visual Amplification. We revisit visual attention amplification as a diagnostic tool. Following PAI (Liu et al., 2024d), increasing attention to visual tokens consistently improves POPE (Li et al., 2023b) and yields initial gains on Reefknot (Zheng et al., 2024). However, these gains are limited and non-monotonic, indicating visual amplification is necessary but insufficient



(a) Baseline Model (b) Naive Attention Amplification Method

Figure 3: **The naive visual attention amplification method exacerbates visual inertia.** A visual activeness comparison between (a) the baseline model and (b) the naive attention amplification method PAI shows that the amplification strategy reduces visual attention activeness and is associated with cognitive hallucinations.

for cognitive hallucination mitigation. As shown in Figure 2, hallucinated responses remain confined to limited, relation-incomplete areas. We therefore shift focus from the amount of visual attention to its dynamics during generation.

Visual Inertia in Cognitive Hallucinations. We term this failure of dynamic visual attention **Visual Inertia**. During generation, attention becomes temporally persistent and spatially confined, preventing the model from integrating multiple regions required for relational inference. Unlike attention insufficiency, visual inertia concerns whether attention remains responsive to relational demands. To quantify it, we define visual activeness $\mathcal{A}$ as the mean Wasserstein distance between consecutive visual attention distributions over image patches:

$$\mathbf{p}_i = \frac{\sum_{h=1}^{H} \mathbf{A}_i^{(h)}}{\left\| \sum_{h=1}^{H} \mathbf{A}_i^{(h)} \right\|_1} \tag{3}$$

$$\mathcal{A} = \frac{1}{T-1} \sum_{i=1}^{T-1} W(\mathbf{p}_i, \mathbf{p}_{i+1}) \tag{4}$$

Here, $\mathbf{A}_i^{(h)}$ denotes the visual attention from the h-th head for the i-th output token, and $W(\cdot,\cdot)$ denotes the Wasserstein distance over image patches, with Manhattan ground cost on the 2D patch grid.

Intervention	Type	$\mathcal{A}$	$\mathbf{R}_{\text{score}} \uparrow$
Baseline	–	41.52	63.96
+ Random perturb.	Control	38.76	61.84
+ Shuffled prev-step	Control	34.21	58.93
+ Current-step focus	Control	30.47	56.18
+ Prev-step prop.	**Inertia**	**19.24**	**47.25**

Table 1: Intervention-based evidence on visual inertia.

We analyze pairwise distances of normalized visual attention over the first 100 tokens from 1,000 LLaVA-1.5 (Liu et al., 2024b) captions on a COCO subset (Lin et al., 2014). Figure 3 shows that naive attention amplification yields consistently lower pairwise distances than the base model during generation. This suggests that stronger visual attention does not necessarily improve attention dynamics, but may instead reinforce dominant regions and make attention trajectories more inertial.

Evidence on Visual Inertia via Counterfactual Intervention: While visual activeness is closely associated with cognitive hallucinations, correlation alone does not clarify whether temporally persistent attention is merely a byproduct or contributing mechanism. We therefore introduce counterfactual interventions:

$$\mathbf{A}_t' = \text{Norm}(\mathbf{A}_t + \Delta_t^A), \tag{5}$$

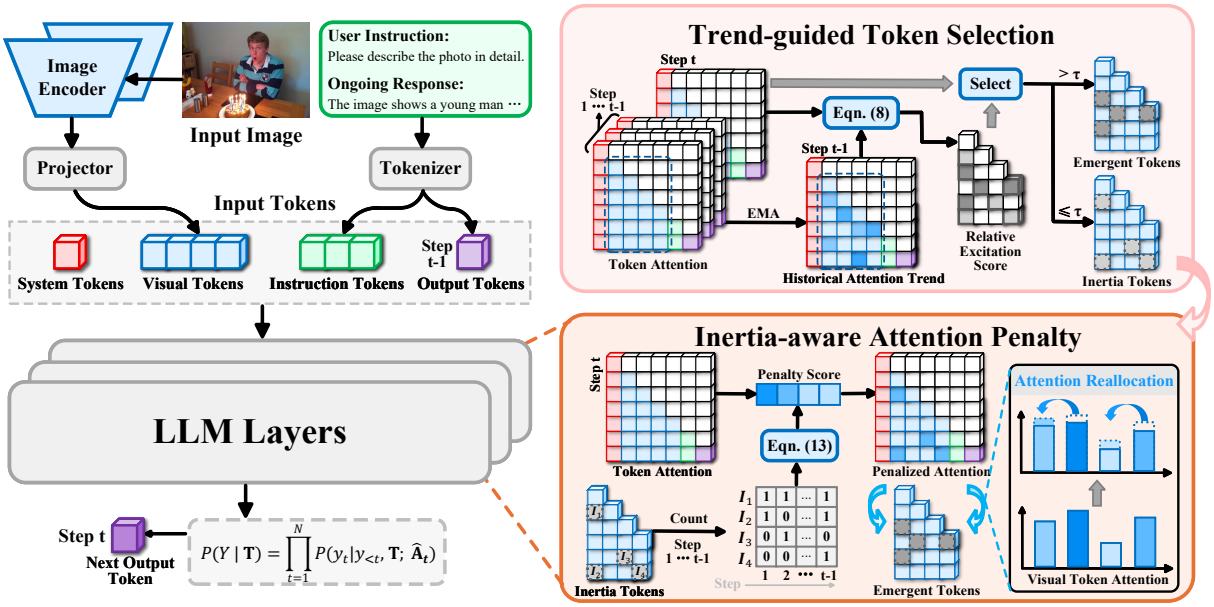


Figure 4: **Overview of IVE. Left:** IVE modulates token attention during autoregressive decoding. **Top Right:** Trend-guided Token Selection partitions visual tokens according to their deviation from historical attention trends, distinguishing emergent tokens from inertia tokens. **Bottom Right:** Inertia-aware Attention Penalty quantifies the persistence of inertia tokens, attenuates their influence, and reallocates the penalized attention to emergent tokens.

where Δ_t^A denotes the intervention applied to the attention map at step t. We compare three control interventions with previous-step propagation, which explicitly induces temporal persistence. As shown in Table 1, all interventions reduce both visual activeness and relation-level accuracy. Compared with the three control interventions, the inertia-inducing intervention causes the largest drop, decreasing $\mathcal{A}$ by 22.28 and $\mathbf{R}_{\text{score}}$ by 16.71. These interventions support visual inertia as a plausible contributing mechanism behind cognitive hallucinations. Intervention details are in the appendix.

3 Approach

Motivated by findings that cognitive hallucinations stem more closely from visual inertia than from insufficient attention alone, we propose Inertia-aware Visual Excitation (IVE) in Figure 4. Instead of amplifying visual attention, IVE modulates attention by its temporal deviation from historical trends, encouraging responses to newly relevant regions while suppressing persistently dominant ones.

3.1 Trend-guided Token Selection

Current attention cannot reveal whether visual token is newly relevant or simply remains historically dominant. To resolve this ambiguity, IVE compares current attention with its historical trend, partitioning visual tokens into emergent and inertia tokens. Emergent tokens become newly prominent relative to historical patterns, whereas inertia tokens remain highly attended without sufficient excitation.

Let $\mathbf{A}_t^{(h)} \in \mathbb{R}^N$ denote the attention vector from the current query token to all tokens at decoding step t in head h, where N is the number of tokens. We define the attention assigned to token j as the average over all heads:

$$a_{t,j} = \frac{1}{H} \sum_{h=1}^{H} \mathbf{A}_{t,j}^{(h)}, \tag{6}$$

where $\mathbf{A}_{t,j}^{(h)}$ denotes the attention weight from the current query token to token j in head h, and H is the number of attention heads.

We therefore maintain an Exponential Moving Average (EMA) as a historical reference:

$$\tilde{a}_{t-1,j} = \gamma \cdot \tilde{a}_{t-2,j} + (1-\gamma) \cdot a_{t-1,j}, \tag{7}$$

where $\gamma \in (0,1)$ controls temporal smoothing. We then define the relative excitation score:

$$S_{t,j} = \frac{a_{t,j} - \tilde{a}_{t-1,j}}{\sqrt{\tilde{a}_{t-1,j}(1-\tilde{a}_{t-1,j})} + \epsilon}, \tag{8}$$

where $\epsilon = 10^{-6}$ ensures numerical stability. This normalization highlights tokens whose attention rises relative to their historical baseline, while suppressing persistently dominant ones.

Token selection is performed using a threshold τ for emergent tokens, while inertia tokens are identified by current attention and relative excitation:

$$\mathcal{E}_t = \{ j \in \mathcal{V} \mid S_{t,j} > \tau \}, \tag{9}$$

$$\mathcal{I}_t = \{j \mid a_{t,j} > \mu_t^v,\ S_{t,j} \le \tau\}, \tag{10}$$

where

$$\mu_t^v = \frac{1}{N_v} \sum_{j \in \mathcal{V}} a_{t,j} \tag{11}$$

denotes the mean attention over visual tokens at step t. Here, $\mathcal{E}_t$ captures newly activated tokens, whereas $\mathcal{I}_t$ collects highly attended tokens without sufficient relative excitation.

3.2 Inertia-aware Attention Penalty

Selecting emergent tokens alone does not prevent attention from remaining concentrated on historically dominant regions. We therefore introduce Inertia-aware Attention Penalty to attenuate such persistent inertia.

We quantify the persistence of each inertia token by counting its previous selections:

$$C_{t,j} = \sum_{k=1}^{t-1} \mathbf{1}(j \in \mathcal{I}_k), \tag{12}$$

where no penalty is applied at $t = 1$.

To avoid overly suppressing useful visual evidence, we scale the penalty by the normalized persistence of each inertia token:

$$a'_{t,j} = 1 - \alpha \frac{C_{t,j}}{t-1}, \quad j \in \mathcal{I}_t, \tag{13}$$

where $\alpha \in (0, 1]$ controls the attenuation factor for inertia tokens. To further emphasize newly relevant regions, we reinforce emergent tokens according to their relative excitation:

$$\hat{a}_{t,j} = \begin{cases} a'_{t,j}, & j \in \mathcal{I}_t, \\ 1 + S_{t,j}, & j \in \mathcal{E}_t, \\ 1, & \text{otherwise}, \end{cases} \tag{14}$$

Here, $\hat{a}_{t,j}$ defines the token-level modulation factor for selected visual tokens at step t. We further implement this modulation by recalibrating cached value vectors rather than directly modifying attention probabilities after softmax. Specifically, for $j \in \mathcal{E}_t \cup \mathcal{I}_t$, we update the cached value vector of visual token j at layer l and head h as

$$V_j^{(l,h)} \leftarrow \hat{a}_{t,j}\, V_j^{(l,h)}. \tag{15}$$

The designated layer is used only to extract a stable token-level signal for emergent and inertia token selection. The resulting modulation is then applied across all heads and transformer layers at each decoding step, since visual inertia affects visual-token contributions cumulatively throughout decoding.

Algorithm 1: ***Inertia-aware Visual Excitation***

Input: Visual attention $\{a_{t,j}\}_{j\in\mathcal{V}}$, EMA $\{\tilde{a}_{t-1,j}\}_{j\in\mathcal{V}}$, inertia counts $\{C_{t,j}\}_{j\in\mathcal{V}}$, threshold τ, coefficients α, γ

 # Trend-guided Token Selection

for *each decoding step* t **do**

  Compute $\{S_{t,j}\}_{j\in\mathcal{V}}$

  $\mathcal{E}_t \leftarrow \{j \mid S_{t,j} > \tau\}$ ▷ *emergent*

  $\mathcal{I}_t \leftarrow \{j \mid a_{t,j} > \mu_t^v,\ S_{t,j} \le \tau\}$ ▷ *inertia*

  # Inertia-aware Attention Penalty

  for *each visual token* $j \in \mathcal{V}$ **do**

    if $j \in \mathcal{I}_t$ ***then*** ▷ *inertia attenuate*

      $C_{t,j} \leftarrow C_{t,j} + 1$

      $\hat{a}_{t,j} \leftarrow 1 - \alpha \frac{C_{t,j}}{t-1}$

    if $j \in \mathcal{E}_t$ ***then*** ▷ *emergent reinforce*

      $\hat{a}_{t,j} \leftarrow 1 + S_{t,j}$

    if $j \notin \mathcal{I}_t \cup \mathcal{E}_t$ ***then*** ▷ *attention unchanged*

      $\hat{a}_{t,j} \leftarrow 1$

  Recalibrate cached values by $\hat{a}_{t,j}$ ▷ *inject*

return generated response

3.3 Discussion

Visual Inertia Beyond Static Attention Imbalance. Attention aggregation in language and multimodal models has been shown to influence hallucination behaviors (Wang et al., 2023; Huang et al., 2024; Tang et al., 2025). Static attention imbalance describes how concentrated attention is at a single decoding step, whereas visual inertia concerns whether such concentration persists across decoding steps. Our analysis shows that stronger visual attention can still reduce visual activeness and degrade relational grounding, indicating that cognitive hallucinations are not captured by static attention magnitude alone. Moreover, the intervention that explicitly increases temporal persistence yields the largest degradation among several controls, suggesting that visual inertia reflects a temporal failure mode beyond static imbalance.

4 Experiments

In this section, we comprehensively evaluate our IVE method on multiple MLLMs and on hallucination and performance benchmarks. We also present ablation studies on key modules, hyperparameters and computational efficiency.

4.1 Experimental Settings

Benchmarks: We evaluate IVE on two main benchmark categories. Reefknot (Zheng et al., 2024) and Hal-Eval (Jiang et al., 2024a) target cognitive hallucinations involving complex inter-object relations. MME (Fu et al., 2025) and MM-

Category	Method	LLaVA-1.5 (Liu et al., 2024b)			InstructBLIP (Dai et al., 2023a)			Qwen2.5-VL-Instruct (Bai et al., 2025)		
		Perc.↓	Cogn.↓	R_{score} ↑	Perc.↓	Cogn.↓	R_{score} ↑	Perc.↓	Cogn.↓	R_{score} ↑
Y/N	Vanilla	37.67	33.99	63.96	45.31	42.47	55.94	37.22	20.58	70.12
	+VCD (Leng et al., 2024)	37.33	33.67	64.58	45.22	42.35	56.05	37.13	20.47	70.23
	+OPERA (Huang et al., 2024)	37.08	33.23	64.62	45.03	42.05	56.28	36.86	20.12	70.53
	+PAI (Liu et al., 2024d)	37.21	33.77	64.31	44.89	41.86	56.45	36.82	19.95	70.63
	+HELPD (Yuan et al., 2024)	37.05	33.31	64.68	44.58	41.94	56.53	36.79	20.04	70.46
	+Deco (Wang et al., 2025)	37.22	33.42	64.51	45.17	42.21	56.14	37.00	20.33	70.36
	+VASparse (Zhuang et al., 2025)	37.13	33.14	64.63	44.67	42.47	56.30	36.91	20.86	70.17
	+MemVR (Zou et al., 2025)	36.99	32.81	64.86	44.34	42.09	56.65	36.75	20.16	70.57
	+Ours	**35.71**	**31.94**	**65.93**	**43.62**	**40.41**	**57.76**	**35.83**	**18.76**	**71.68**
	Improvement	**-1.96**	**-2.05**	**+1.97**	**-1.69**	**-2.06**	**+1.82**	**-1.39**	**-1.82**	**+1.56**
MCQ	Vanilla	68.05	51.04	43.70	73.91	53.85	39.94	67.16	49.69	44.91
	+VCD (Leng et al., 2024)	67.86	50.90	43.86	73.67	53.75	40.09	67.02	49.58	45.02
	+OPERA (Huang et al., 2024)	67.12	50.21	44.56	72.98	53.23	40.66	66.51	49.13	45.50
	+PAI (Liu et al., 2024d)	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	44.48	73.18	53.27	40.77	66.44	49.04	45.62
	+Deco (Wang et al., 2025)	67.54	50.73	44.07	73.30	53.60	40.30	66.84	49.38	45.22
	+VASparse (Zhuang et al., 2025)	67.21	50.88	44.13	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	45.53
	+Ours	**66.02**	**49.84**	**45.24**	**71.88**	**51.96**	**41.83**	**65.51**	**48.12**	**46.61**
	Improvement	**-2.03**	**-1.20**	**+1.54**	**-2.03**	**-1.89**	**+1.89**	**-1.65**	**-1.57**	**+1.70**

Table 2: **Cognitive Hallucination Mitigation on Reefknot benchmark**. Results (%) of decoding strategies across three MLLMs. Perc./Cogn. denote perception/cognition hallucination rates, and R_{score} measures response accuracy.

Method	LLaVA-1.5 (Liu et al., 2024b)				InstructBLIP (Dai et al., 2023a)				Qwen2.5-VL-Instruct (Bai et al., 2025)			
	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
+VCD (Leng et al., 2024)	53.91	+1.26	37.08	+1.26	50.88	+1.08	67.41	+1.06	63.37	+0.94	58.26	+0.98
+OPERA (Huang et al., 2024)	55.47	+2.82	38.53	+2.71	51.73	+1.93	68.11	+1.76	64.95	+2.52	59.43	+2.15
+PAI (Liu et al., 2024d)	55.12	+2.47	38.02	+2.20	51.38	+1.58	67.96	+1.61	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	68.54	+2.19	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	67.88	+1.53	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	69.08	+2.73	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	69.47	+3.12	65.62	+3.19	60.83	+3.55
+Ours	**56.92**	**+4.27**	**40.11**	**+4.29**	**53.74**	**+3.94**	**70.58**	**+4.23**	**66.74**	**+4.31**	**61.92**	**+4.64**

Table 3: **Cognitive Hallucination Mitigation on Hal-Eval benchmark**. Results(%) on the relation hallucination subset of Hal-Eval benchmark across three MLLMs.

Bench (Liu et al., 2024e) assess broader MLLM performance on perception, cognition, and reasoning. Results on perceptual hallucinations, including POPE (Li et al., 2023b), are in the appendix.

Models: We evaluate IVE on three representative MLLMs: LLaVA-1.5 (Liu et al., 2024b), InstructBLIP (Dai et al., 2023a), and Qwen2.5-VL-Instruct (Bai et al., 2025). All models use 7B language backbones and pretrained vision encoders. Detailed prompt templates are in the appendix.

Implementation Details: Token selection uses layers 20, 20, and 18 for LLaVA-1.5 (Liu et al., 2024b), InstructBLIP (Dai et al., 2023a), and Qwen2.5-VL-Instruct (Bai et al., 2025), respectively, with cached value recalibration applied across all transformer layers. All experiments are conducted on NVIDIA H800 GPUs. Additional details and ablations are in the appendix.

4.2 Main Results

Cognitive Hallucination of MLLMs: We evaluate IVE on two cognitive hallucination benchmarks, Reefknot (Zheng et al., 2024) and Hal-Eval (Jiang et al., 2024a). As shown in Table 2, IVE consistently achieves the best results on both Y/N and MCQ settings across three backbones, surpassing the state-of-the-art method MemVR (Zou et al., 2025) by up to 1.27% in R_{score}. Table 3 further shows that IVE generalizes to the relation hallucination subset of Hal-Eval (Jiang et al., 2024a), where it also remains consistently better than competing methods. Taken together, these results indicate that IVE effectively mitigates cognitive hallucinations across backbones and benchmarks.

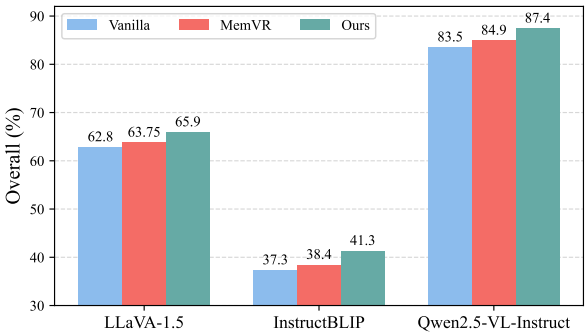


Figure 5: Results (%) on MMBench benchmark for evaluating multidimensional MLLM performance.

Method	LLaVA-1.5 (Liu et al., 2024b)				InstructBLIP (Dai et al., 2023a)				Qwen2.5-VL-Instruct (Bai et al., 2025)			
	Com.↑	Num.↑	Text.↑	Code.↑	Com.↑	Num.↑	Text.↑	Code.↑	Com.↑	Num.↑	Text.↑	Code.↑
Vanilla	87.86	105.0	97.50	67.50	117.9	40.00	92.50	55.00	100.7	52.50	92.50	47.50
+VCD (Leng et al., 2024)	89.29	101.5	98.00	65.00	118.6	45.00	94.00	57.50	101.4	55.00	94.00	49.50
+OPERA (Huang et al., 2024)	92.14	107.0	92.50	58.00	121.4	47.50	89.50	52.00	103.6	56.00	89.50	46.50
+PAI (Liu et al., 2024d)	91.43	106.5	91.50	56.50	120.7	47.00	89.00	51.50	102.9	55.50	89.00	46.00
+HELPD (Yuan et al., 2024)	91.79	106.0	95.50	62.50	121.8	46.50	94.00	56.50	103.1	55.00	93.50	50.00
+Deco (Wang et al., 2025)	90.71	102.0	100.0	70.00	120.4	46.00	96.00	59.50	102.3	54.50	96.00	52.50
+VASparse (Zhuang et al., 2025)	92.86	108.0	93.00	59.50	122.1	48.00	91.00	53.50	104.3	56.50	91.00	48.00
+MemVR (Zou et al., 2025)	93.57	109.0	101.5	70.50	122.9	47.50	97.00	60.00	104.6	57.00	97.00	53.00
+Ours	**94.64**	**110.0**	**103.0**	**72.00**	**124.3**	**49.00**	**98.00**	**61.50**	**106.4**	**58.00**	**98.00**	**54.50**
Improvement	**+6.79**	**+5.00**	**+5.50**	**+4.50**	**+6.43**	**+9.00**	**+5.50**	**+6.50**	**+5.71**	**+5.50**	**+5.50**	**+7.00**

Table 4: **Multidimensional MLLM performance on MME Benchmark.** Results(%) on the subset of MME benchmark across three MLLMs. Com., Num., Text., and Code denote commonsense reasoning, numerical calculation, text translation, and code reasoning respectively.

Selection	Penalty	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
✓	✓	**35.71**	**31.94**	**90.53**	**90.94**

Table 5: Ablation study on the key modules of IVE.

Model	Method	R_{score} ↑	Time↓	Memory↓
LLaVA-1.5	Vanilla	63.96	0:14:44	15.15G
	+VCD	64.58	0:38:23	16.06G
	+OPERA	64.62	0:17:51	15.72G
	+MemVR	64.86	0:16:07	15.54G
	+IVE	**65.93**	**0:15:47**	**15.38G**
InstructBLIP	Vanilla	55.94	0:47:02	31.13G
	+VCD	56.05	1:50:08	32.17G
	+OPERA	56.28	0:56:23	31.79G
	+MemVR	56.65	0:53:25	31.68G
	+IVE	**57.76**	**0:52:52**	**31.45G**
Qwen2.5-VL-Instruct	Vanilla	70.12	0:31:46	33.41G
	+VCD	70.23	1:32:21	34.65G
	+OPERA	70.53	0:36:52	33.95G
	+MemVR	70.57	0:34:02	33.72G
	+IVE	**71.68**	**0:33:45**	**33.61G**

Table 6: Comparison of inference latency and memory cost across decoding strategies on Reefknot benchmark.

Multidimensional Performance of MLLMs: Beyond hallucination-specific benchmarks, we further evaluate IVE on MME (Fu et al., 2025) and MMBench (Liu et al., 2024e) to assess broader multidimensional MLLM performance. As shown in Table 4, IVE consistently achieves the best results across multiple cognition-oriented MME tasks, including commonsense reasoning, numerical calculation, text translation, and code reasoning. Figure 5 further shows that IVE also outperforms the state-of-the-art method MemVR (Zou et al., 2025) on MMBench, indicating clear gains across perception, cognition, and reasoning dimensions.

4.3 Ablation Studies and Analysis

In this subsection, we analyze the key modules, coefficients, and computational efficiency of IVE. Additional results are provided in the appendix.

Ablation Study on Key Modules: Table 5 reports the ablation results of the two core modules in IVE. Either module alone improves over the vanilla model on Reefknot (Zheng et al., 2024) and POPE (Li et al., 2023b). Combining them yields the best results, showing that token selection and inertia suppression are complementary. This confirms that both modules contribute to the overall effectiveness of IVE from different aspects.

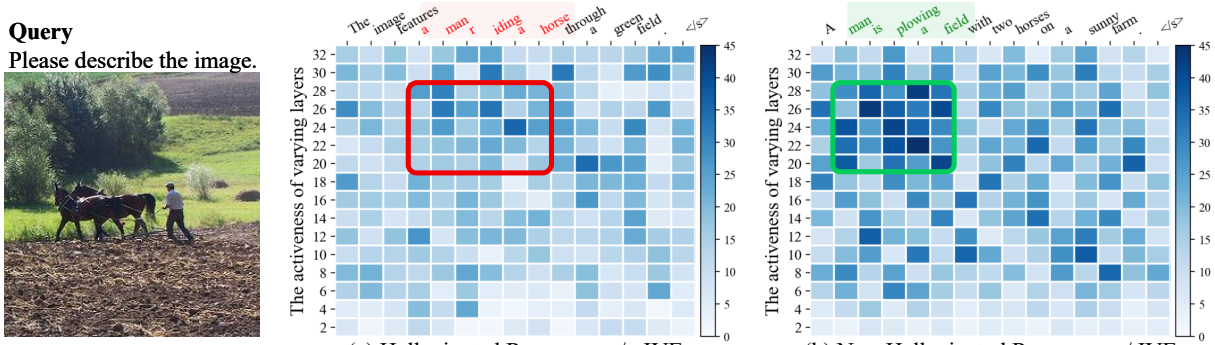
Ablation Study on Key Coefficients: Table 7 reports the effect of the three key coefficients, α, γ, and τ, on both perceptual and cognitive hallucinations. The best settings vary slightly across backbones, but all three models generally favor moderate coefficients. Across all tested settings, IVE consistently remains above MemVR (Zou et al., 2025), which further supports the practical applicability of the proposed method. This indicates that IVE remains stable within a reasonable coefficient range, without requiring highly sensitive tuning.

Analysis of Latency and Computational Cost: As shown in Table 6, we compare inference latency, GPU memory usage, and hallucination performance across decoding strategies on Reefknot (Zheng et al., 2024). Across three backbones, IVE achieves the best overall performance with marginal overhead, reaching R_{score} values of 65.93%, 57.76%, and 71.68%, respectively. Compared with other methods, it delivers higher accuracy with lower latency and memory usage.

Visualization of Visual Activeness: As shown in Figure 6, we visualize token-wise visual activeness across layers during generation. Without IVE, the hallucinated response shows weak activeness on relation-misleading tokens, such as *_r*, *iding*. In contrast, with IVE, the non-hallucinated response exhibits stronger activeness on relation-critical tokens, including *plowing*, *_a*, and *field*. This suggests that IVE encourages more active visual grounding for cognitive understanding.

Model	α					γ					τ				
	Val.	Reefknot		POPE		Val.	Reefknot		POPE		Val.	Reefknot		POPE	
		Perc.↓	Cogn.↓	Acc↑	F1↑		Perc.↓	Cogn.↓	Acc↑	F1↑		Perc.↓	Cogn.↓	Acc↑	F1↑
LLaVA-1.5	0.05	36.84	32.72	89.20	89.27	0.08	36.56	32.48	89.77	89.82	2.95	36.72	32.44	89.63	89.68
	0.10	35.71	31.94	90.53	90.94	0.09	35.71	31.94	90.53	90.94	3.00	35.71	31.94	90.53	90.94
	0.15	36.60	32.43	89.57	89.62	0.10	35.88	32.07	90.53	90.94	3.05	35.93	32.10	90.27	90.32
InstructBLIP	0.05	44.28	41.73	87.23	87.31	0.09	44.16	41.57	89.97	90.02	3.00	43.75	40.60	90.20	90.16
	0.10	43.62	40.41	90.20	90.16	0.10	43.62	40.41	90.20	90.16	3.05	43.62	40.41	90.20	90.16
	0.15	43.88	41.34	88.93	88.95	0.11	43.98	41.36	90.07	90.12	3.10	43.88	40.99	90.10	90.11
Qwen2.5-VL -Instruct	0.10	36.18	19.22	91.80	91.85	0.11	36.12	19.10	91.73	91.78	3.05	36.08	19.01	91.88	91.94
	0.15	35.83	18.76	92.10	92.15	0.12	35.83	18.76	92.10	92.15	3.10	35.83	18.76	92.10	92.15
	0.20	36.11	19.08	91.93	91.98	0.13	36.20	19.18	91.68	91.74	3.15	36.14	19.12	91.79	91.84

Table 7: Ablation study on key coefficients of IVE.



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

Figure 6: Visual activeness across tokens in cognitive hallucination tasks. Rows denote LLaVA-1.5 layers and columns denote decoded tokens. IVE strengthens activeness on relation-critical tokens in non-hallucinated responses.

5 Related Works

Multimodal Large Language Models: Multimodal large language models (MLLMs) extend large language models with visual encoders and cross-modal projectors for multimodal understanding and generation (Touvron et al., 2023a,b; Taori et al., 2023; Zhu et al., 2024). Representative systems such as LLaVA (Liu et al., 2024c,b), BLIP-2/InstructBLIP (Li et al., 2023a; Dai et al., 2023a), and Qwen2.5-VL-Instruct (Bai et al., 2025) have substantially improved multimodal perception, instruction following, and cross-modal alignment. Despite these advances, MLLMs remain vulnerable to hallucinations, especially in scenarios requiring compositional understanding.

Hallucination in MLLMs: Hallucinations in MLLMs arise when generated responses conflict with visual content, user instructions, or real-world facts (Zhang et al., 2025; Su et al., 2025; Ding et al., 2025). Benchmarks such as POPE (Li et al., 2023b) mainly evaluate perceptual hallucinations, while Reefknot (Zheng et al., 2024) and Hal-Eval (Jiang et al., 2024a) extend evaluation to cognitive hallucinations involving inter-object relations. Existing mitigation methods are either training-based (Jiang et al., 2024b; Liu et al., 2024a; Yu et al., 2024a; Yue et al., 2024; Gunjal et al., 2024; Kim et al., 2024; Sun et al., 2024; Yu et al., 2024b) or training-free (Wang et al., 2025; Huang et al., 2024; Lyu et al., 2025; Chuang et al., 2024; Leng et al., 2024; Lyu et al., 2026), but most primarily address perceptual hallucinations through additional supervision or inference-time visual recalibration.

6 Conclusion

In this paper, we propose Inertia-aware Visual Excitation (IVE) method to mitigate cognitive hallucinations. Unlike existing methods that mainly target perceptual hallucinations concerning object existence or attributes, IVE explicitly models cognitive inference as the dynamic responsiveness of visual attention. Inspired by the observation that attention to visual tokens often exhibits persistent inertia, IVE identifies visual tokens that emerge relative to historical attention trends while distinguishing tokens exhibiting inertial behavior, enabling adaptive attention reallocation. To further mitigate persistent attention on inertia tokens and promote compositional inference, IVE introduces an inertia-aware penalty that discourages over-concentration and constrains persistent attention within localized regions. Extensive experiments demonstrate that IVE is effective across various base MLLMs and multiple hallucination benchmarks.

Limitations

This work has several limitations. Our analysis of visual inertia is based on attention dynamics and intervention-based evidence, but it does not constitute a complete account of all mechanisms underlying cognitive hallucinations. Moreover, IVE operates only at decoding time, which limits it to inference-time regulation rather than parameter-level correction. Finally, although we evaluate across multiple MLLMs and benchmarks, the current study is centered on single-image cognitive hallucination settings, and its extension to larger models and more complex multimodal scenarios remains for future work.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *ICRA*, pages 14093–14100.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *ICLR*, pages 54158–54183.

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

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023b. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *EACL*, pages 2136–2148.

Yue Ding, Xiaofang Zhu, Tianze Xia, Junfei Wu, Xinlong Chen, Qiang Liu, and Liang Wang. 2025. D ^{2} hscore: Reasoning-aware hallucination detection via semantic breadth and depth analysis in llms. *arXiv preprint arXiv:2509.11569*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *AAAI*, volume 38, pages 18135–18143.

Mingzhe Hu, Shaoyan Pan, Yuheng Li, and Xiaofeng Yang. 2023. Advancing medical imaging with language models: A journey from n-grams to chatgpt. *arXiv preprint arXiv:2304.04920*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, pages 13418–13427.

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

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024b. Hallucination augmented contrastive learning for multimodal large language model. In *CVPR*, pages 27036–27046.

Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. 2025. Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models. In *ICML*.

Minchan Kim, Minyeong Kim, Junik Bae, Suhwan Choi, Sungkyung Kim, and Buru Chang. 2024. Exploiting semantic reconstruction to mitigate hallucinations in vision-language models. In *ECCV*, pages 236–252.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pages 13872–13882.

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

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

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Aligning large multi-modal model with robust instruction tuning. In *ICLR*, pages 57689–57733.

Haokun Liu, Yaonan Zhu, Kenji Kato, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. 2023. Llm-based human-robot collaboration framework for manipulation tasks. *arXiv preprint arXiv:2308.14972*.

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

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024e. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, pages 216–233.

Guangtao Lyu, Xinyi Cheng, Chenghao Xu, Qi Liu, Muli Yang, Fen Fang, Huilin Chen, Jiexi Yan, Xu Yang, and Cheng Deng. 2025. Revealing perception and generation dynamics in lvlms: Mitigating hallucinations via validated dominance correction. *arXiv preprint arXiv:2512.18813*.

Guangtao Lyu, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Xueting Li, Fen Fang, and Cheng Deng. 2026. Revealing and enhancing core visual regions: Harnessing internal attention dynamics for hallucination mitigation in lvlms. *arXiv preprint arXiv:2602.15556*.

Jinjie Mai, Jun Chen, Bing Li, Guocheng Qian, Mohamed Elhoseiny, and Bernard Ghanem. 2023. Llm as a robotic brain: Unifying egocentric memory and control. *arXiv preprint arXiv:2304.09349*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2025. Unsupervised real-time hallucination detection based on the internal states of large language models. In *ACL*, pages 14379–14391.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *ACL*, pages 13088–13110.

Barrett Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. 2025. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *ICLR*, volume 2025, pages 27745–27776.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *EMNLP*, pages 9840–9855.

Sheng Wang, Zihao Zhao, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. 2024. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 3(1):133.

Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. 2023. Embodied task planning with large language models. *arXiv preprint arXiv:2307.01848*.

Yuxin Xie, Zhihong Zhu, Xianwei Zhuang, Liming Liang, Zhichang Wang, and Yuexian Zou. 2024. Gpa: Global and prototype alignment for audio-text retrieval. In *Proc. Interspeech*, pages 5078–5082.

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

Yuguo Yin, Yuxin Xie, Wenyuan Yang, Dongchao Yang, Jinghan Ru, Xianwei Zhuang, Liming Liang, and Yuexian Zou. 2025b. Atri: Mitigating multilingual audio text retrieval inconsistencies by reducing data distribution errors. In *ACL*, pages 5491–5504.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *CVPR*, pages 12944–12953.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, pages 13807–13816.

Xinlei Yu, Chengming Xu, Guibin Zhang, Yongbo He, Zhangquan Chen, Zhucun Xue, Jiangning Zhang, Yue Liao, Xiaobin Hu, Yu-Gang Jiang, and 1 others. 2025. Visual multi-agent system: Mitigating hallucination snowballing via visual flow. *arXiv preprint arXiv:2509.21789*.

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

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. In *ACL*, pages 11766–11781.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Yian Zhao, Kehan Li, Zesen Cheng, Pengchong Qiao, Xiawu Zheng, Rongrong Ji, Chang Liu, Li Yuan, and Jie Chen. 2024a. Graco: Granularity-controllable interactive segmentation. In *CVPR*, pages 3501–3510.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. 2024b. Detrs beat yolos on real-time object detection. In *CVPR*, pages 16965–16974.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. 2024. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. In *ACL*, pages 6193–6212.

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

Xianwei Zhuang, Xuxin Cheng, Zhihong Zhu, Zhanpeng Chen, Hongxiang Li, and Yuexian Zou. 2024. Towards multimodal-augmented pre-trained language models via self-balanced expectation-maximization iteration. In *ACM MM*.

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

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and 1 others. 2025. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *ICML*.

Appendix Overview

This supplementary material provides prompt templates, extended experimental results, and qualitative case studies to support the main manuscript. The contents are organized as follows:

- **Prompt Templates for Datasets (§A):** Prompt formats used for Reefknot, POPE, MME, and MMBench.
- **Details for Motivation Analyses (§B):** Additional settings and analyses for the motivation experiments in Section 2.
- **More Details of Implementation (§C):** Additional implementation specifics for token selection and value-space recalibration.
- **More Experimental Results and Analysis (§D):** Additional radar-chart comparisons, heatmap and activeness analyses, perceptual hallucination results.
- **Case Studies Across Different Datasets (§E):** Qualitative examples from Reefknot, MME, POPE, and MMBench.

A Prompt Templates for Datasets

A.1 Prompt Templates for POPE Dataset

The POPE dataset (Li et al., 2023b) is specifically designed to evaluate object hallucinations and object–attribute consistency within perceptual hallucinations in MLLMs. In the POPE dataset (Li et al., 2023b), the input template for the model is presented below, where the prompts are highlighted in **green** and the image is highlighted in **red** for clearer visualization.

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
>
> **USER: IMAGE**
>
> **Existence:**
> **Is there a snowboard in the image? Please just answer yes or no.**
>
> **ASSISTANT:**

A.2 Prompt Templates for Reefknot Dataset

The Reefknot dataset (Zheng et al., 2024) is specifically designed to evaluate cognitive hallucinations that arise from complex inter-object relationships in visual scenes, which often require models to perform relational reasoning beyond simple visual recognition and object detection. In the Reefknot dataset (Zheng et al., 2024), the input template for the model is presented below, where the prompts are highlighted in **green** and the image is highlighted in **red** for clearer illustration of the input structure and the interaction format provided to the model during evaluation.

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
>
> **USER: IMAGE**
>
> **(i) Y/N:**
> **Is the dog smelling frisbee in this photo? Please answer yes or no.**
>
> **(ii) Multichoice:**
> **What is the relation with sign and building in this photo?**
> A. into B. over C. towards D. within
> **Please choose.**
>
> **ASSISTANT:**

A.3 Prompt Templates for MME Dataset

The MME dataset (Fu et al., 2025) provides a comprehensive benchmark for systematically evaluating various capabilities of MLLMs. Here, we specifically focus on the subsets related to cognitive hallucinations. We select the cognitive and reasoning-oriented tasks from MME (Fu et al., 2025), including commonsense reasoning, numerical calculation, text translation, and code reasoning, which are particularly prone to cognitive hallucinations. In the MME dataset (Fu et al., 2025), the input template for the model is presented below, where the prompts are highlighted in **green** and the image is highlighted in **red** for clearer illustration.

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

USER: **IMAGE**

The image shows a python code. Is the output of the code "36"?

ASSISTANT:

A.4 Prompt Templates for MMBench Dataset

The MMBench dataset (Liu et al., 2024e) provides a multidimensional and comprehensive assessment of MLLM performance across perception, cognition, and reasoning capabilities. In the MMBench dataset (Liu et al., 2024e), the input template for the model is presented below, where the prompts are highlighted in **green** and the image is highlighted in **red** for clearer presentation.

A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

USER: **IMAGE**

Which one describes capybaras?
A. Shy animals that hide in tall grass.
B. Wild guinea pigs that live in mountain.
C. Closest relatives of the hippopotamus.
D. Large rodents that are good swimmers.
Please choose.

ASSISTANT:

B Details for Motivation Analyses

B.1 Visual Attention Diversity Under Naive Amplification

To further analyze why naive visual attention amplification is insufficient for cognitive hallucination mitigation, we examine how visual attention evolves during decoding. Following the intuition that successful relational inference should involve attention shifts across multiple regions, we compare the diversity of visual attention patterns produced by the baseline model and by the naive amplification method PAI (Liu et al., 2024d).

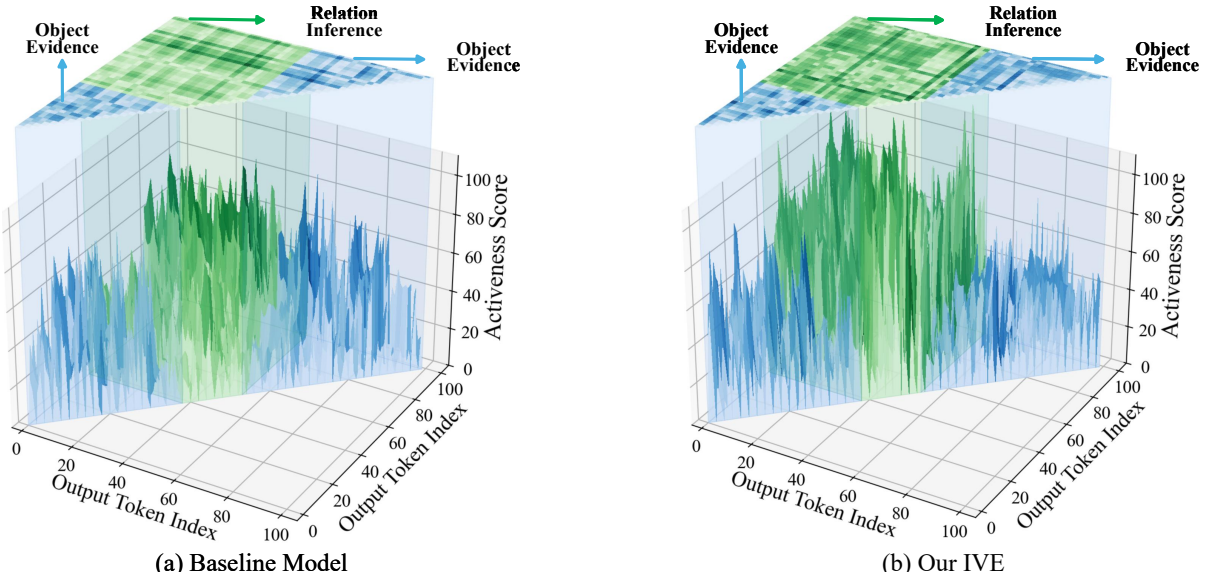
Specifically, we generate 1,000 captions using LLaVA-1.5 (Liu et al., 2024b) on a subset of COCO (Lin et al., 2014). During generation, we store the normalized visual attention weights of the first 100 generated tokens. For each sample, we then compute a pairwise distance matrix (100×100) over visual attention patterns, where the distance is measured by the Wasserstein distance with Manhattan ground cost on the 2D patch grid.

The resulting comparison is shown in Figure 3 of the main paper. If the model dynamically shifts attention across different image regions, the pairwise distances between token-level visual attention patterns should remain relatively large. In contrast, lower pairwise distances indicate that attention stays confined to similar regions over time, reflecting reduced visual activeness. Our results show that PAI consistently yields lower pairwise distances than the baseline model, suggesting that naive amplification makes visual attention more temporally persistent and less responsive to relational demands. For completeness, Figure 7 further compares the baseline with IVE, showing that IVE preserves higher pairwise distances than PAI and better maintains dynamic attention shifts during decoding.

B.2 Manual Intervention Settings

For the intervention analysis in Table 1, we apply several manual perturbations to the visual attention maps of LLaVA-1.5 (Liu et al., 2024b) on Reefknot (Zheng et al., 2024). The interventions are designed to separate temporal persistence from other simple attention perturbations:

- **Random perturbation.** We add random noise to the current-step visual attention map and then re-normalize it. This serves as a generic disturbance control, testing whether arbitrary attention corruption alone degrades relation reasoning.
- **Shuffled previous-step attention.** We inject the previous-step attention after shuffling its spatial positions. This preserves the rough attention magnitude from the previous step while breaking spatial continuity, allowing us to test whether historical attention helps only through its values or through aligned spatial persistence.



(a) Baseline Model (b) Our IVE

Figure 7: Visual activeness comparison between the baseline model and our IVE method.

- **Current-step focus.** We sharpen the current-step attention distribution toward its already dominant regions. This intervention increases over-concentration without introducing explicit temporal carryover, serving as a control for static attention imbalance.

- **Previous-step propagation.** We directly propagate previous-step attention to the current step. This intervention explicitly increases temporal persistence and therefore serves as the closest manual probe of the proposed visual inertia hypothesis.

Taken together, these settings test whether temporally persistent attention is more harmful to relation reasoning than generic disturbance, broken continuity, or stronger single-step concentration.

C More Details of Implementation

For all backbone models, token selection is computed from a designated attention layer, while cached value recalibration is applied across all transformer layers during decoding. In our experiments, the designated selection layer is set to layer 20 for LLaVA-1.5 (Liu et al., 2024b), layer 20 for InstructBLIP (Dai et al., 2023a), and layer 18 for Qwen2.5-VL-Instruct (Bai et al., 2025). These choices follow the same mid-to-late layer (Wang et al., 2025; Jung et al., 2025; Zhuang et al., 2025) preference observed in prior attention-based analyses, while the resulting token-level modulation is propagated throughout the full decoding stack.

D More Experimental Results and Analysis

This section provides additional qualitative and quantitative evidence for the effectiveness of IVE beyond the main paper. We further examine cross-backbone generalization, attention visualization, perceptual hallucination mitigation, visual activeness, and the sensitivity to token selection layers.

D.1 Radar Chart Results

Figure 8 presents a systematic comparison with hallucination mitigation methods across multiple benchmarks under two additional models, i.e., InstructBLIP (Dai et al., 2023a) and Qwen2.5-VL-Instruct (Bai et al., 2025). Notably, the results consistently indicate that IVE achieves superior performance compared with other methods under both base models, demonstrating strong effectiveness and robustness across diverse evaluation settings. This advantage is especially clear on benchmarks targeting cognitive hallucinations, such as Reefknot (Zheng et al., 2024) and MME (Fu et al., 2025), where IVE maintains stronger scores across different evaluation dimensions rather than improving only a single metric. The same trend also extends to perceptual hallucination benchmarks such as POPE (Li et al., 2023b), suggesting that inertia-aware modulation improves both relational reasoning and general visual grounding. Overall, Figure 8 complements the main-paper tables by showing that the gains of IVE are not isolated to one benchmark or one backbone.

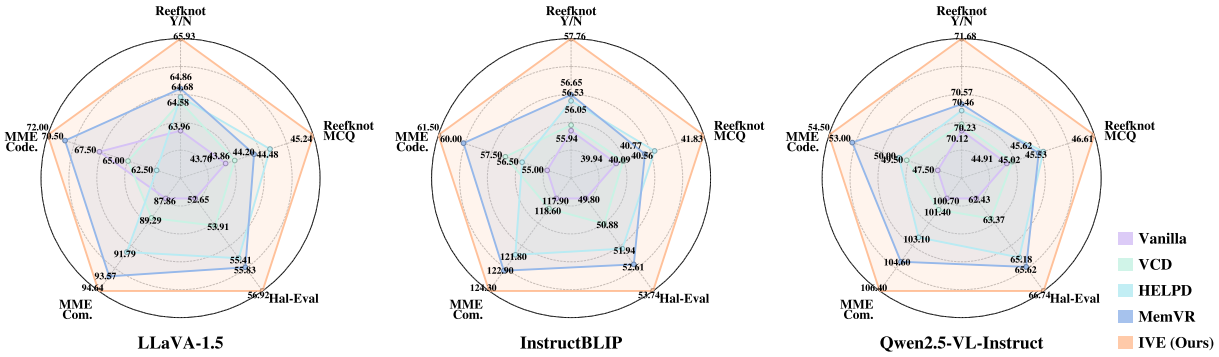


Figure 8: Radar chart comparison of hallucination mitigation methods across multiple benchmarks under LLaVA-1.5, InstructBLIP and Qwen2.5-VL-Instruct.

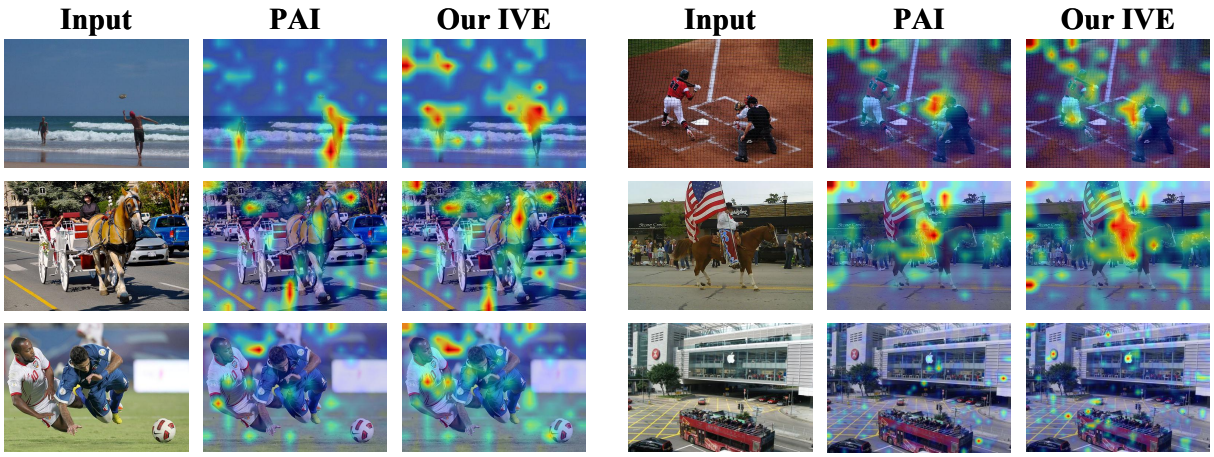


Figure 9: Attention heatmap comparison between our IVE and the naive attention amplification method PAI.

D.2 Heatmap Comparison

Figure 9 presents a qualitative comparison of attention heatmaps generated by different methods. The visualization shows that responses produced by IVE distribute attention over multiple regions that are relationally relevant to the queried objects, indicating a more comprehensive visual grounding process. In contrast, the naive attention amplification method PAI (Liu et al., 2024d) tends to concentrate attention within a limited spatial area, suggesting a narrower and less informative focus. This difference highlights that IVE does not simply strengthen attention magnitude, but reshapes attention dynamics toward broader cross-region evidence aggregation. As a result, the model is better able to integrate compositional cues from multiple visual regions, which is particularly important for relation-centric reasoning. The heatmaps therefore provide qualitative support for the central claim that mitigating visual inertia requires restoring dynamic visual responsiveness rather than merely amplifying already dominant regions.

D.3 Perceptual Hallucination Mitigation

Given its effectiveness in mitigating cognitive hallucinations, we further evaluated our method on perceptual hallucinations. As shown in Table 8, it achieves the best performance on the POPE benchmark (Li et al., 2023b), which targets object existence and attributes. Across different backbone models, our method consistently outperforms alternatives, and the largest gains in the POPE random setting are 9.61%, 9.75%, and 6.85%. The gains remain stable across the Random, Popular, and Adversarial splits, indicating that the effect is not restricted to a single evaluation setting. This suggests that although IVE is motivated by cognitive hallucinations, the proposed inertia-aware modulation also improves general visual grounding and benefits perceptual hallucination mitigation.

D.4 Visual Activeness Comparison

Figure 7 presents a comparison of visual attention dynamics between IVE and the naive visual attention amplification method PAI (Liu et al., 2024d) during image captioning. The results show that

Category	Method	LLaVA-1.5 (Liu et al., 2024b)				InstructBLIP (Dai et al., 2023a)				Qwen2.5-VL-Instruct (Bai et al., 2025)			
		Acc↑	Δ	F1↑	Δ	Acc↑	Δ	F1↑	Δ	Acc↑	Δ	F1↑	Δ
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 (Leng et al., 2024)	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 (Huang et al., 2024)	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 (Wang et al., 2025)	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 (Liu et al., 2024d)	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 (Zhuang et al., 2025)	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 (Zou et al., 2025)	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 (Leng et al., 2024)	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 (Huang et al., 2024)	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 (Wang et al., 2025)	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 (Liu et al., 2024d)	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 (Zhuang et al., 2025)	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 (Zou et al., 2025)	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 (Leng et al., 2024)	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 (Huang et al., 2024)	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 (Wang et al., 2025)	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 (Liu et al., 2024d)	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 (Zhuang et al., 2025)	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 (Zou et al., 2025)	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**

Table 8: **Hallucination Mitigation on POPE benchmark.** Accuracy and F1-score (%) comparison under different evaluation settings (i.e., Random, Popular and Adversarial). Δ denotes improvement over vanilla baseline.

Model	LLaVA-1.5					InstructBLIP					Qwen2.5-VL-Instruct				
	Layer	Reefknot		POPE		Layer	Reefknot		POPE		Layer	Reefknot		POPE	
		Perc.↓	Cogn.↓	Acc.↑	F1↑		Perc.↓	Cogn.↓	Acc.↑	F1↑		Perc.↓	Cogn.↓	Acc.↑	F1↑
Layer	16	36.14	32.36	90.18	90.24	16	44.02	40.97	89.94	90.01	14	36.09	19.14	91.72	91.76
	18	35.96	32.18	90.41	90.48	18	43.84	40.71	90.18	90.26	16	35.94	18.95	91.97	92.02
	20	**35.71**	**31.94**	**90.53**	**90.94**	20	**43.62**	**40.41**	**90.20**	**90.16**	18	**35.83**	**18.76**	**92.10**	**92.15**

Table 9: Ablation study on token selection layers of IVE.

PAI (Liu et al., 2024d) consistently exhibits lower pairwise distances, indicating reduced semantic activeness in visual attention. In contrast, IVE maintains higher pairwise distances, suggesting more active and dynamic attention transitions across visual regions. This difference implies that naive attention amplification tends to exacerbate visual inertia, restricting attention to limited regions and hindering the capture of compositional semantics required for relational inference. By comparison, IVE preserves more diverse attention trajectories across decoding steps, which better supports cross-region evidence integration.

D.5 Ablation on Token Selection Layers

Table 9 reports a small ablation on the designated token selection layer. Across all three backbones, mid-to-late layers (Wang et al., 2025; Jung et al., 2025; Zhuang et al., 2025) provide the most stable performance, while the selected layers used in the main experiments consistently yield the best trade-off across Reefknot (Zheng et al., 2024) and POPE (Li et al., 2023b). Earlier layers remain beneficial compared with the vanilla model, but their gains are generally smaller, suggesting that token selection becomes more reliable once visual-semantic alignment is more mature. This trend is consistent with our design choice to decouple stable token selection from all-layer value recalibration.

E Case Studies Across Different Datasets

To further illustrate the effectiveness of IVE, we present representative case studies across multiple widely used benchmark datasets. Figure 10 ∼ 17 present several representative examples where IVE successfully suppresses hallucinations. Compared with the baseline responses, the outputs generated by IVE demonstrate improved alignment with the visual content, capturing relevant objects, attributes, and inter-object relationships more reliably.

Across multiple benchmarks, including Reefknot (Zheng et al., 2024), MME (Fu et al., 2025), POPE (Li et al., 2023b), and MMBench (Liu et al., 2024e), IVE consistently demonstrates strong capability in mitigating hallucinations under diverse evaluation settings. To ensure a fair and controlled comparison, all experiments are conducted using the LLaVA-1.5 model (Liu et al., 2024b).

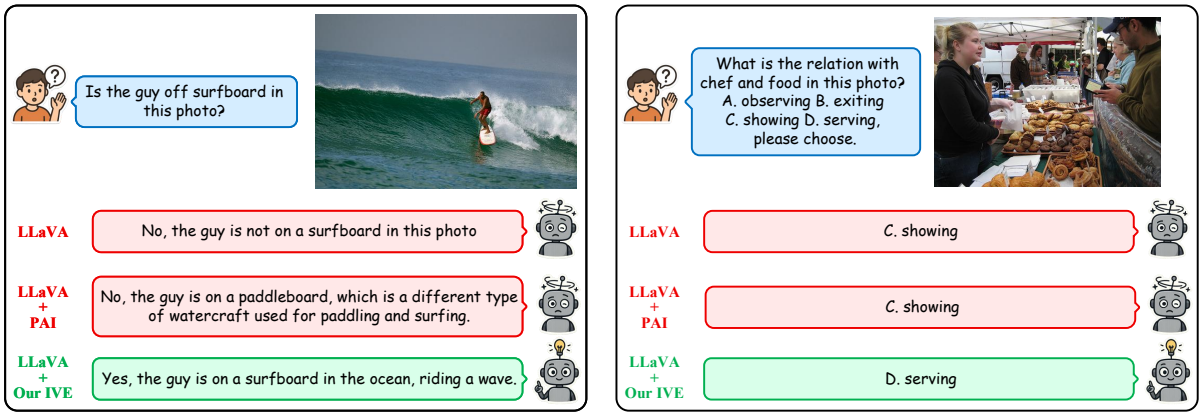


Figure 10: An illustration of IVE mitigating cognitive hallucinations on Reefknot.

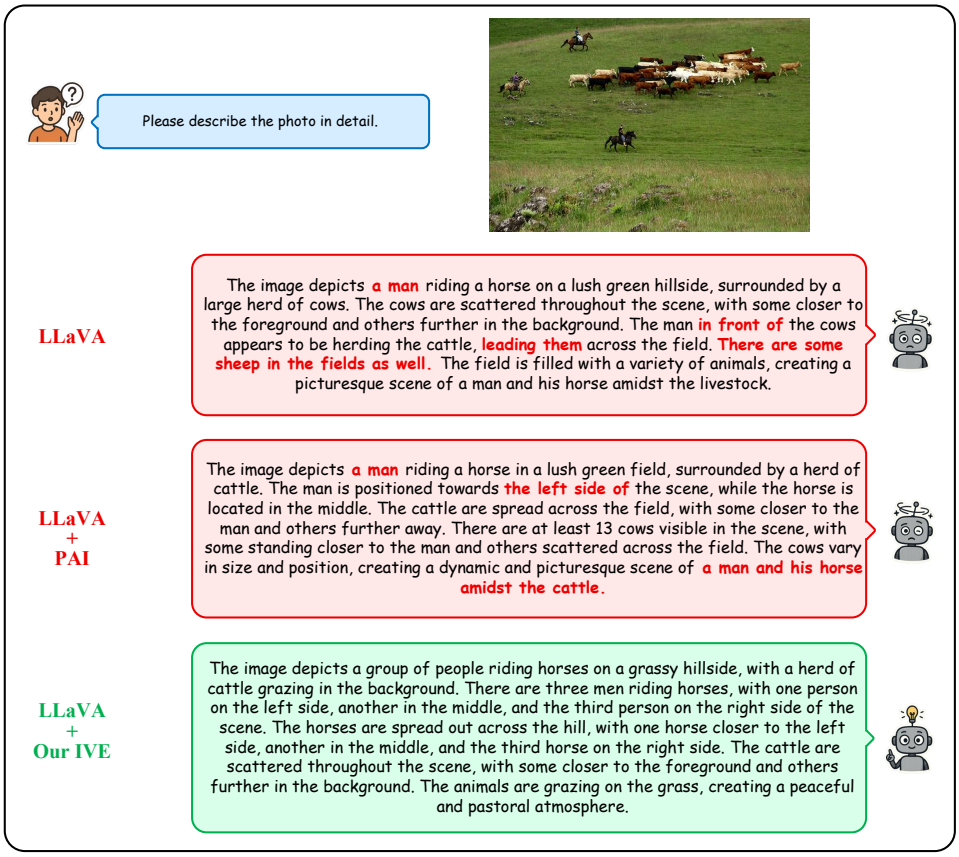


Figure 11: An illustration of IVE mitigating cognitive hallucinations on Reefknot.

Figure 12: An illustration of IVE mitigating cognitive hallucinations on Reefknot.

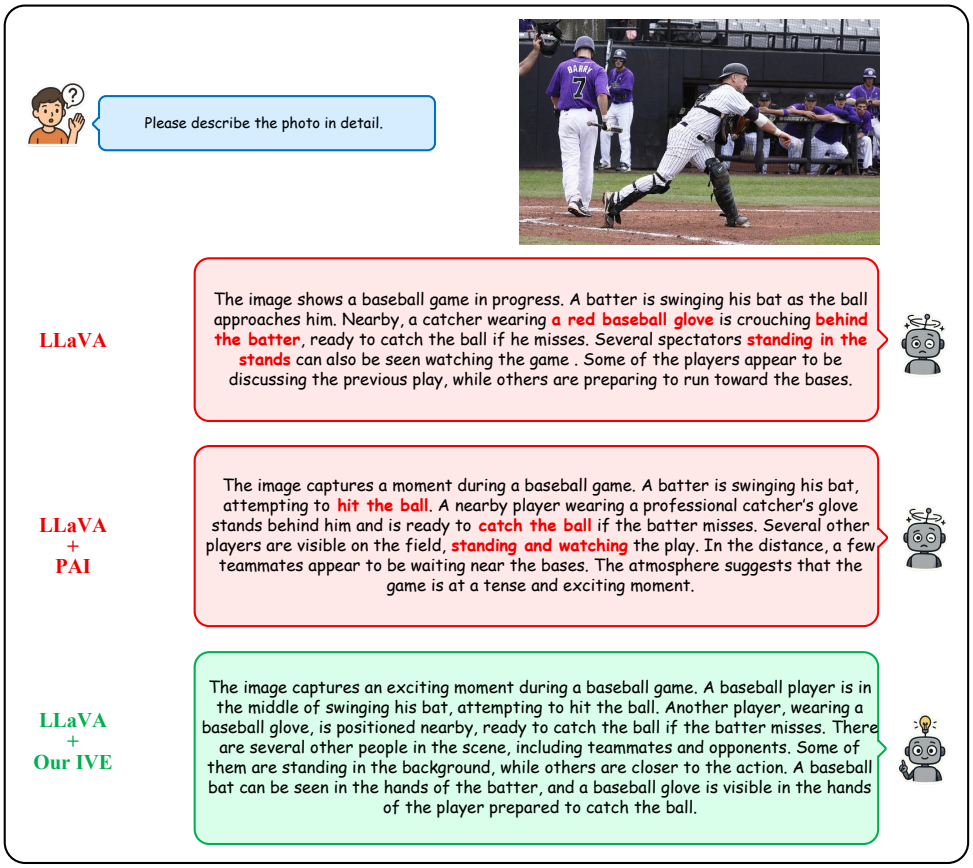


Figure 13: An illustration of IVE mitigating cognitive hallucinations on Reefknot.

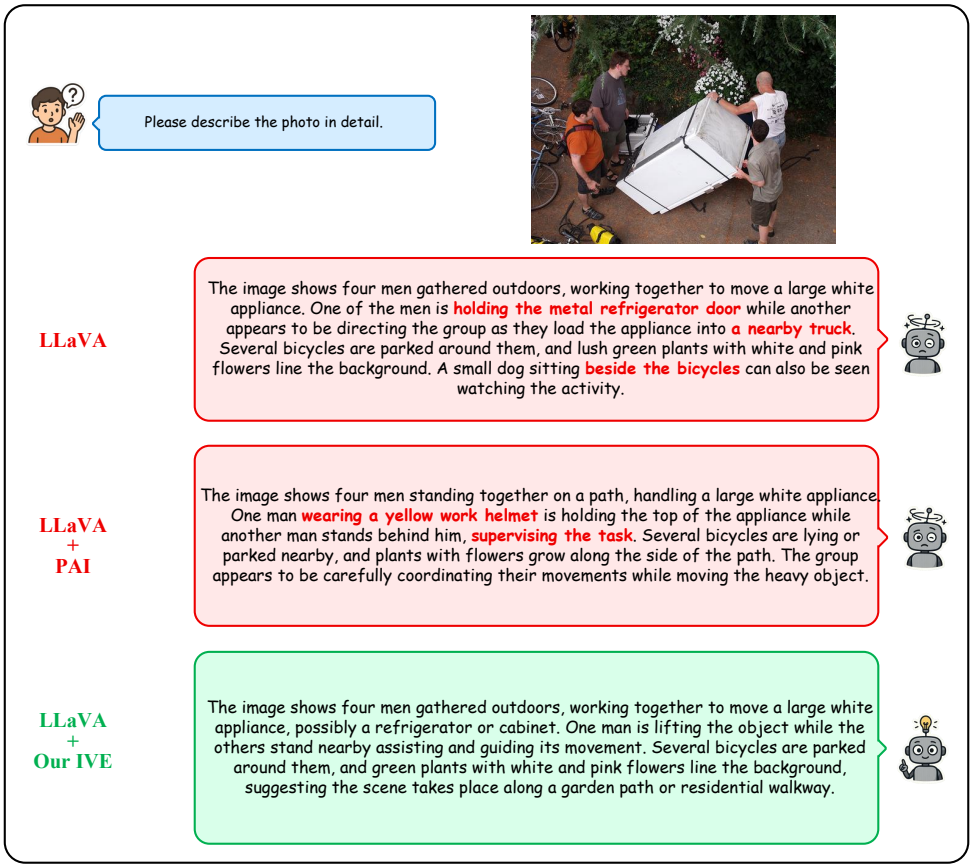


Figure 14: An illustration of IVE mitigating cognitive hallucinations on Reefknot.

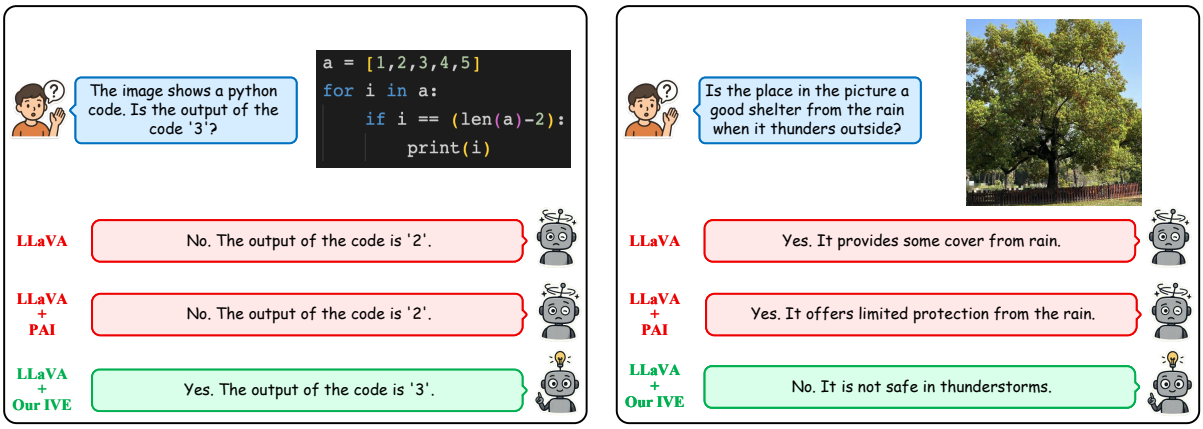


Figure 15: An illustration of IVE mitigating cognitive hallucinations on MME.

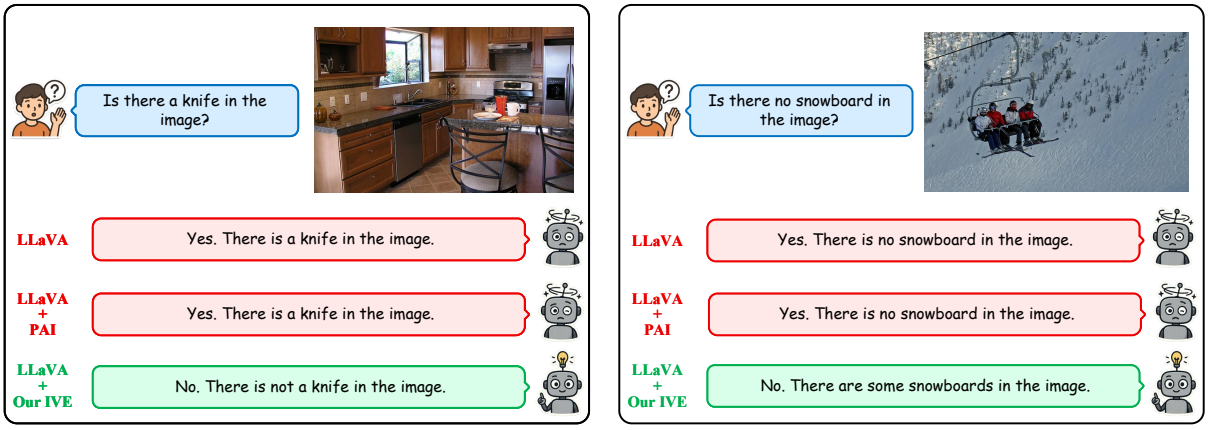


Figure 16: An illustration of IVE mitigating perceptual hallucinations on POPE.

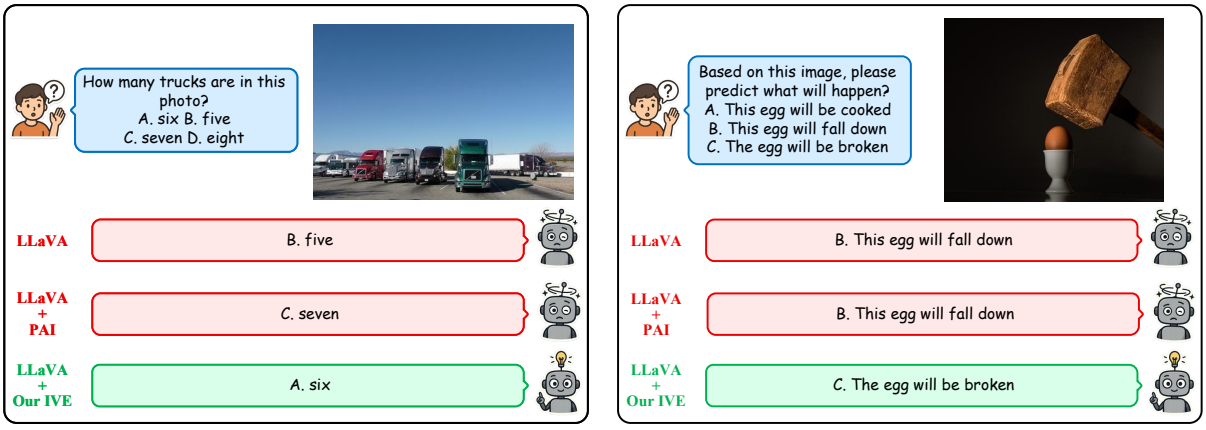


Figure 17: An illustration of IVE mitigating cognitive hallucinations on MMBench.