Focus Matters: Attention-Value Dynamics for Efficient Hallucination Mitigation in Vision-Language Models

Sohyeon Kim[1] **Sang Yeon Yoon**[2] **Kyeongbo Kong**[1,3*]

[1] Dept. of Electrical & Electronics Eng., Pusan National University
[2] Dept. of Computer Eng. & AI, Pukyong National University
[3] AI Collaboration Hub Research Center, Pusan National University
Republic of Korea

{shkim0503,kbkong}@pusan.ac.kr palsc@pukyong.ac.kr

Abstract

Large Vision-Language Models (LVLMs) have achieved impressive progress in multimodal reasoning, yet they remain prone to object hallucinations, generating descriptions of objects that are not present in the input image. In this work, we investigate hallucination from the perspective of attention-value dynamics inside LVLM vision encoders. We identify a consistent three-phase structure of visual processing—diffusion, focus, and rediffusion—and show that the focus phase is where attention most clearly separates strongly and weakly supported visual tokens. However, low attention does not necessarily imply negligible downstream influence: low-attention tokens in the focus phase can still exert non-negligible value-side influence on the attention output relative to their small attention mass. Through controlled phase-wise value interventions, we find that hallucination behavior is particularly sensitive to the value content of low-attention tokens during this phase. Replacing or neutralizing these values reduces hallucination metrics while largely preserving grounded object evidence. A token-level teacher-forcing analysis further shows that the intervention reduces the probability of hallucinated object tokens with much smaller effects on ground-truth object tokens. In addition, Visual Attention Ratio (VAR) analysis shows that focus-phase intervention is accompanied by increased attention to visual tokens during decoding. Based on these observations, we instantiate a simple training-free inference-time intervention that replaces focus-phase low-attention values with an image-level mean value vector using statistics from a single forward pass. Experiments across multiple LVLM backbones demonstrate that this analysis-derived intervention reduces object hallucination with negligible additional runtime and remains compatible with existing inference-time mitigation methods. Our project page is available at: https://paper.pnu-cvsp.com/FocusMatters/

1 Introduction

Large Vision-Language Models (LVLMs) have achieved remarkable progress in multimodal understanding and image-grounded text generation [1, 2, 3, 4, 5]. Despite these advances, they remain vulnerable to *object hallucination* [6, 7], generating descriptions of objects that are not present in the input image. Such hallucinations undermine the reliability of LVLMs and limit their use in applications requiring faithful visual grounding. This issue becomes increasingly important as LVLMs are incorporated into multimodal agents. In such settings, visual outputs may influence subsequent reasoning, planning, tool use, or inter-agent communication, and unreliable visual information can propagate across multiple stages of an agentic pipeline, amplifying the impact of an

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

initial hallucination. At the same time, agentic systems often invoke multimodal models repeatedly across multiple reasoning steps or agents, making computational efficiency increasingly important. Therefore, as LVLMs are deployed in more complex multimodal and agentic systems, both reliable visual grounding and efficient inference become essential.

Existing hallucination mitigation methods largely intervene during decoding or within the language model. Decoder-stage methods adjust token probabilities, contrast visual conditions, or modify generation dynamics [8, 9, 10], while language-side attention methods reweight the model's reliance on visual tokens during generation [11, 12]. Although effective, these methods operate after visual features have already been extracted, without directly addressing how individual visual tokens contribute to hallucination. Recent studies on visual token pruning [13] have also shown that selectively removing certain visual tokens can mitigate hallucination, suggesting a potential connection between unreliable visual information and hallucination behavior. Vision-encoder-side methods intervene earlier by suppressing unreliable visual signals before they reach the language model. A representative approach, *Adversarial Uncertainty Estimation* (AUE) [14], identifies uncertain visual tokens through iterative adversarial optimization using Projected Gradient Descent (PGD) [15]. However, this per-image optimization introduces substantial latency; on Qwen-2.5-VL, for example, AUE increases total inference time from 2.60s to 33.46s, making it approximately $12.9\times$ slower than standard inference. This motivates a closer analysis of where hallucination-sensitive visual signals arise inside the vision encoder and whether they can be controlled without iterative optimization.

In this work, we study object hallucination through the lens of *attention-value dynamics* inside LVLM vision encoders. Self-attention propagates visual information by aggregating value vectors according to attention weights, so attention scores alone do not fully characterize a token's downstream influence. As summarized in Fig. 1, we observe a consistent three-phase structure across vision encoder layers: *diffusion*, *focus*, and *rediffusion*. Attention is broadly distributed in the diffusion phase, becomes concentrated on a small subset of tokens in the focus phase, and becomes diffuse again in the rediffusion phase.

The focus phase is important because it is where attention most clearly separates strongly supported tokens from weakly supported ones. However, low attention does not necessarily mean that a token's value content is removed from the computation. We find that focus-phase low-attention tokens receive small attention mass but still exert non-negligible value-side influence on the attention output. These tokens do not dominate the visual representation, but their value contribution is disproportionately large relative to their attention mass. This attention-value mismatch raises the question of whether weakly supported value-side signals are related to hallucinated object generation.

To examine this question, we conduct controlled phase-wise value interventions. We use value modification as a diagnostic tool to test how different value groups affect hallucination behavior. Modifying low-attention values during the focus phase reduces hallucination metrics while largely preserving caption fidelity, whereas analogous interventions in the diffusion or rediffusion phases have weaker effects. In contrast, suppressing high-attention values substantially harms caption fidelity, indicating that strongly attended values contain important grounded visual evidence. This contrast suggests that hallucination reduction is not simply caused by removing visual information, but is closely related to neutralizing weakly supported value-side signals in the focus phase.

We further validate this observation using token-level teacher forcing. For hallucinated and ground-truth object tokens identified from generated captions, we recompute token log-probabilities under the same token sequence while applying focus-phase value interventions. The intervention reduces hallucinated object token probabilities while having much smaller effects on ground-truth object probabilities. Visual Attention Ratio (VAR) analysis [12] further shows that this intervention is accompanied by increased attention to visual tokens during decoding. Together, these results provide diagnostic evidence that focus-phase low-attention value signals are associated with hallucination-sensitive visual information.

Based on these findings, we instantiate a simple analysis-derived intervention: focus-phase low-attention mean value replacement. The method identifies low-attention tokens using attention statistics from a single forward pass and replaces their value vectors with the image-level mean value vector. This neutralizes token-specific value deviations without removing tokens, requiring additional training, or performing adversarial optimization. Experiments across multiple LVLM backbones show that this intervention reduces object hallucination with negligible runtime overhead and remains compatible with existing inference-time mitigation methods.

Our contributions are summarized as follows:

- **Attention-value dynamics analysis.** We identify a diffusion–focus–rediffusion structure in LVLM vision encoders and show that the focus phase exhibits an attention-value mismatch: low-attention tokens receive small attention mass while retaining non-negligible value-side influence.
- **Diagnostic evidence for hallucination-sensitive value signals.** Through phase-wise value interventions and token-level teacher-forcing analysis, we provide diagnostic evidence that focus-phase low-attention value signals affect hallucinated object likelihood under controlled interventions.
- **Analysis-derived lightweight intervention.** We instantiate a training-free intervention that replaces focus-phase low-attention values with an image-level mean value vector using statistics from a single forward pass.
- **Empirical validation across LVLM backbones.** We evaluate the intervention across multiple LVLMs and show a favorable trade-off between hallucination reduction, caption fidelity, and inference efficiency.

2 Related Work

Large Vision-Language Models. Large Vision-Language Models (LVLMs) [1, 2, 3, 4, 5] extend large language models (LLMs) [16, 17, 18] with visual inputs. A typical LVLM consists of a vision encoder [19, 20, 21], a modality connector, and an LLM. Despite strong multimodal performance, LVLMs remain prone to object hallucination, generating objects or relationships that are inconsistent with the input image.

Hallucination Mitigation in LVLMs. Hallucination mitigation methods can be broadly divided into training-based and training-free approaches. Training-based methods use hallucination-aware supervision or preference optimization [22, 23, 24, 25, 26], but require additional data and training. Training-free methods instead intervene during inference. Decoder-side methods modify token probabilities or generation dynamics [8, 9, 10], while language-side attention methods adjust the decoder's reliance on visual tokens [11, 12]. These methods are complementary to ours but operate after visual features have already been extracted. Vision-encoder-side methods such as AUE [14] intervene earlier by estimating unreliable visual tokens through iterative PGD-based optimization [15], which introduces substantial latency. In contrast, our work analyzes attention-value dynamics inside the vision encoder and applies a single-pass, training-free intervention during the focus phase.

3 Attention-Value Dynamics in Vision Encoders

To identify where hallucination-sensitive visual signals arise inside LVLM vision encoders, we analyze how visual information is routed and transformed across layers. In self-attention, attention weights determine which tokens are referenced, while value vectors determine what information is propagated. Therefore, attention scores alone do not fully characterize a token's downstream influence. In particular, a token receiving low attention may still affect the attention output if its value content contributes non-negligibly to the weighted aggregation. Motivated by this distinction, we jointly analyze layer-wise attention concentration and low-attention value influence across vision encoder layers.

3.1 Three-Phase Structure from Attention Concentration

We first quantify how attention distributions evolve across layers. For a given layer ℓ and attention head h, let $\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{N\times N}$ denote the self-attention matrix over visual tokens. For architectures with a CLS token, we define the reference attention distribution from the CLS token to the N spatial tokens as

$$a^{(\ell,h)} = \{A^{(\ell,h)}_{\mathrm{cls},i}\}_{i=1}^{N}. \tag{1}$$

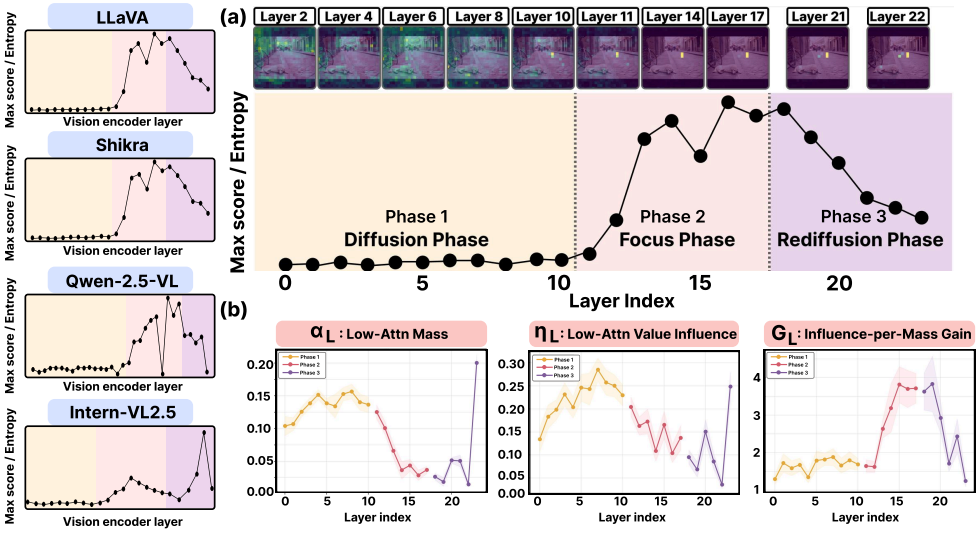

Figure 1: **Attention-value dynamics across vision encoder layers.** (a) Layer-wise attention concentration ratio $R^{(\ell)}$ reveals a consistent diffusion–focus–rediffusion structure. (b) Low-attention value influence across layers. We measure the attention mass $\alpha_{\mathcal{L}}$ assigned to low-attention tokens, their normalized value contribution $\eta_{\mathcal{L}}$, and the influence per attention mass $G_{\mathcal{L}} = \eta_{\mathcal{L}}/(\alpha_{\mathcal{L}} + \epsilon)$. During the focus phase, $\alpha_{\mathcal{L}}$ decreases sharply while $\eta_{\mathcal{L}}$ remains non-negligible, yielding a large $G_{\mathcal{L}}$. This indicates that low-attention tokens can exert disproportionate value-side influence relative to their small attention mass.

For architectures without an explicit CLS token, such as Qwen-2.5-VL, we instead use the average attention distribution over all visual query tokens:

$$a_i^{(\ell,h)} = \frac{1}{N} \sum_{j=1}^{N} A_{j,i}^{(\ell,h)}. \tag{2}$$

To measure attention concentration, we use two statistics: attention entropy and maximum attention score. The attention entropy is defined as

$$H^{(\ell,h)} = -\sum_{i=1}^{N} a_i^{(\ell,h)} \log a_i^{(\ell,h)}, \tag{3}$$

which measures how broadly attention mass is distributed across visual tokens. The maximum attention score is defined as

$$M^{(\ell,h)} = \max_i a_i^{(\ell,h)}, \tag{4}$$

which captures the dominance of the most attended token. We summarize layer-wise attention concentration as

$$R^{(\ell)} = \frac{\mathbb{E}_h[M^{(\ell,h)}]}{\mathbb{E}_h[H^{(\ell,h)}]}. \tag{5}$$

A larger $R^{(\ell)}$ indicates that attention is concentrated on a small subset of tokens, whereas a smaller value indicates a more diffuse attention distribution.

As shown in Fig. 1(a), the layer-wise evolution of $R^{(\ell)}$ reveals a consistent three-phase structure. Although the absolute magnitude differs across models, the depth-wise pattern remains similar: attention is broadly distributed in early layers, becomes sharply concentrated in intermediate layers, and becomes more diffuse again in later layers. Based on this pattern, we divide the vision encoder into three phases:

- **Diffusion phase.** In the early layers, $R^{(\ell)}$ remains low, indicating that attention is broadly distributed across visual tokens.

- **Focus phase.** In the intermediate layers, $R^{(\ell)}$ sharply increases and reaches its peak. We refer to this interval as the focus phase because attention progressively concentrates on a small subset of tokens, thereby most clearly separating strongly supported tokens from weakly supported ones.
- **Rediffusion phase.** In the later layers, $R^{(\ell)}$ decreases again, indicating that the previously concentrated attention pattern becomes more distributed.

The detailed boundary determination procedure is provided in Appendix B.1. This diffusion–focus–rediffusion pattern provides a compact view of how visual token interactions evolve across layers. The focus phase is especially important because it is where the encoder most clearly separates strongly supported tokens from weakly supported ones. We next examine whether low-attention tokens in this phase can still influence the attention output through their value content.

3.2 Low-Attention Value Influence in the Focus Phase

Attention concentration describes how attention mass is distributed, but it does not directly measure how much value content from a token is reflected in the attention output. For each layer ℓ and head h, self-attention computes

$$\mathbf{O}^{(\ell,h)} = \mathbf{A}^{(\ell,h)}\mathbf{V}^{(\ell,h)}, \tag{6}$$

where $\mathbf{A}^{(\ell,h)}$ is the attention matrix and $\mathbf{V}^{(\ell,h)}$ is the value matrix. Thus, even tokens receiving small attention mass can still affect the output if their value content remains non-negligible after weighted aggregation.

We define the low-attention token set $\mathcal{L}^{(\ell)}$ as the bottom 25% of visual tokens ranked by received attention mass. To quantify whether this group still influences the attention output, we measure three quantities: $\alpha_{\mathcal{L}}^{(\ell)}$, the total attention mass assigned to low-attention tokens; $\eta_{\mathcal{L}}^{(\ell)}$, their normalized value contribution to the attention output; and

$$G_{\mathcal{L}}^{(\ell)} = \frac{\eta_{\mathcal{L}}^{(\ell)}}{\alpha_{\mathcal{L}}^{(\ell)} + \epsilon}, \tag{7}$$

which measures value influence per unit attention mass. Detailed definitions of these quantities, including head-wise aggregation, are provided in Appendix B.2.

Fig. 1(b) shows the layer-wise behavior of these quantities. In the diffusion phase, $\alpha_{\mathcal{L}}$ and $\eta_{\mathcal{L}}$ vary together, yielding a relatively small $G_{\mathcal{L}}$. In the focus phase, however, $\alpha_{\mathcal{L}}$ drops sharply as attention concentrates on a small subset of tokens, while $\eta_{\mathcal{L}}$ remains non-negligible. As a result, $G_{\mathcal{L}}$ increases substantially. This indicates that low-attention tokens do not dominate the attention output, but their value content exerts a disproportionate influence relative to their small attention mass. Additional low-attention value influence plots for Qwen-2.5-VL and Shikra-7B are provided in Appendix B.3.

This provides a more precise interpretation of the focus phase. It is not important simply because token values are large. Rather, it is the stage where attention most clearly separates strongly and weakly supported tokens, while the value content of weakly supported tokens can still be reflected in the attention output. This attention-value mismatch motivates the phase-wise value intervention analysis in the next section.

4 Phase-Wise Value Interventions and Hallucination Behavior

We now test whether the focus-phase low-attention value signals identified in Sec. 3 are related to object hallucination. To this end, we conduct controlled value interventions across the diffusion, focus, and rediffusion phases. We use these interventions as diagnostic tools and measure their effects on caption-level hallucination, token-level object likelihood, and decoder-side visual attention. Unless otherwise specified, the diagnostic experiments in this section are conducted on LLaVA-1.5-7B.

4.1 Caption-Level Effects of Phase-Wise Value Intervention

We partition visual tokens into high- and low-attention groups using the received attention mass defined in Sec. 3.2. As shown in Fig. 2, for each phase, we evaluate four diagnostic value interventions:

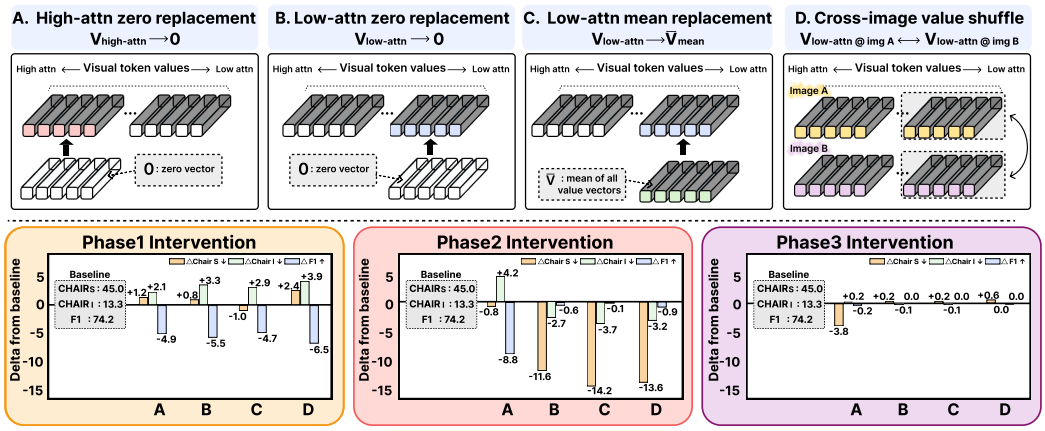

Figure 2: **Phase-wise value interventions and hallucination behavior.** We compare four diagnostic interventions across diffusion, focus, and rediffusion phases: high-attention zero replacement, low-attention zero replacement, low-attention mean replacement, and cross-image value shuffle. Low-attention value interventions in the focus phase produce the strongest reduction in CHAIR_S and CHAIR_I while largely preserving F1. Metric definitions are provided in Appendix A.

(A) replacing high-attention values with zero vectors, (B) replacing low-attention values with zero vectors, (C) replacing low-attention values with the image-level mean value vector, and (D) replacing low-attention values with low-attention values from another image. The high-attention intervention tests the effect of removing strongly supported visual evidence, while the three low-attention interventions remove or disrupt token-specific value content from weakly supported tokens.

We apply each intervention independently to the diffusion, focus, and rediffusion phases, and evaluate generated captions using CHAIR_S, CHAIR_I, and F1. CHAIR_S and CHAIR_I measure sentence-level and instance-level object hallucination, respectively, while F1 measures caption fidelity; detailed benchmark definitions are provided in Appendix A. As shown in Fig. 2, interventions in the diffusion and rediffusion phases produce only limited changes in hallucination metrics. In contrast, modifying low-attention values during the focus phase consistently reduces hallucination. Low-attention zero replacement, mean replacement, and cross-image value shuffle all reduce CHAIR_S and CHAIR_I while largely preserving F1.

This shared trend is important. The three low-attention interventions use different replacement values, but they share one common effect: they remove or disrupt the original token-specific value content of focus-phase low-attention tokens. Thus, the reduction is unlikely to be caused by a particular replacement value. Rather, it suggests that the original value content of these weakly supported tokens contains hallucination-sensitive signals.

The contrast with high-attention intervention further supports this interpretation. In the focus phase, high-attention zero replacement substantially degrades F1, indicating that strongly attended values contain important grounded visual evidence. By contrast, modifying low-attention values reduces hallucination without a comparable loss in caption fidelity. To further examine whether this effect depends on the selection of low-attention tokens, we perform matched controls that keep the intervention phase, 25% token budget, and mean-replacement operation fixed while varying only the selected token group. As shown in Table 1, low-attention selection in the focus phase yields substantially lower CHAIR_S and CHAIR_I than either random or high-attention selection ($p < 0.001$ for both metrics). Full phase-wise results and 95% confidence intervals are provided in Appendix C.1.1.

Table 1: **Matched controls in the focus phase.** The phase, 25% token budget, and mean-replacement operation are fixed, while only the selected token group is varied.

Metric	Baseline	Random	High	Ours
$\mathrm{CHAIR}_S \downarrow$	45.0	39.6 ± 2.3	42.4	**30.8**
$\mathrm{CHAIR}_I \downarrow$	13.3	12.6 ± 0.5	18.1	**9.6**

Ours vs. Random/High: $p < .001$ for both CHAIR_S and CHAIR_I. Full phase-wise results and 95% CIs are provided in Appendix C.1.1.

Table 2: **Focus-phase low-attention value interventions.** Hallu LogP Δ denotes the average log-probability change of hallucinated object tokens relative to the original model. Drop Rate denotes the percentage of samples where hallucinated-token log-probability decreases. GT LogP Δ denotes the corresponding change for ground-truth object tokens. Metric definitions are provided in Appendix A.

Intervention	**CHAIR$_S\downarrow$**	**CHAIR$_I\downarrow$**	**F1$\uparrow$**	**Hallu LogP $\Delta\downarrow$**	**Drop Rate$\uparrow$**	**GT LogP $\Delta \approx 0$**
Baseline	45.0	13.3	74.2	0.000	0.0	0.000
Zero replacement	33.4	10.6	73.6	-0.136	63.5	-0.010
Mean replacement	30.8	9.6	74.1	-0.156	67.3	-0.020
Cross-image shuffle	31.4	10.1	73.3	-0.190	68.2	-0.052

As a complementary diagnostic, we also test the opposite direction by amplifying the contribution of focus-phase low-attention values while keeping the original attention weights and value vectors unchanged. When the amplification factor increases from 1 to 5, CHAIR$_S$ increases from 45.0 to 57.6 and CHAIR$_I$ from 13.3 to 46.5, while F1 decreases from 74.2 to 34.7. The full modulation sweep, statistical trend analysis, and matched amplification controls are provided in Appendix C.1.2.

Together, these results indicate that the observed effect is not simply a consequence of applying value intervention to visual tokens or of using a particular replacement rule, but is closely associated with focus-phase low-attention value signals.

4.2 Token-Level Verification with Teacher Forcing

Caption-level metrics show that hallucination decreases after focus-phase low-attention value intervention, but they do not directly reveal whether hallucinated object tokens themselves become less likely. We therefore perform token-level teacher-forcing analysis.

For captions generated by the original model, we identify hallucinated object tokens and ground-truth object tokens. We then recompute their log-probabilities under the same token sequence while applying focus-phase value interventions. For an object token span $y_s, \ldots, y_e$, we compute the average log-probability

$$\mathrm{LP} = \frac{1}{e-s+1} \sum_{t=s}^{e} \log p(y_t \mid \mathbf{x}, y_{<t}), \tag{8}$$

where $\mathbf{x}$ is the input image and $y_{<t}$ is the fixed textual prefix. We report the log-probability change relative to the original model.

Table 2 summarizes the results. All three low-attention value interventions reduce the log-probability of hallucinated object tokens. Mean replacement decreases hallucinated-token log-probability by -0.156 on average, while cross-image value shuffle yields a decrease of -0.190. The drop rate exceeds 67%, indicating that hallucinated-token probability decreases in a majority of samples.

The effect on ground-truth object tokens is much smaller. For mean replacement, the ground-truth log-probability change is only -0.020, and F1 remains close to the baseline. This suggests that the intervention does not globally suppress object generation. Rather, it preferentially reduces hallucinated object likelihood while largely preserving visually grounded object evidence.

Among the tested interventions, mean replacement provides the best trade-off between hallucination reduction and caption fidelity. It achieves the lowest CHAIR_S and CHAIR_I while maintaining F1 nearly identical to the baseline. We therefore use focus-phase low-attention mean replacement as the default intervention in the full experiments.

4.3 Decoder-Side Visual Attention Analysis

We further analyze how focus-phase value intervention affects the language model's use of visual tokens during decoding. Following prior work [12], we use the Visual Attention Ratio (VAR), which measures the total attention weight assigned by a generated token to image tokens. For a generated token y_k at language-model layer l and head h, VAR is defined as

$$\mathrm{VAR}^{(l,h)}(y_k) = \sum_{i=1}^{N} A_k^{(l,h)}(a_{k,i}), \tag{9}$$

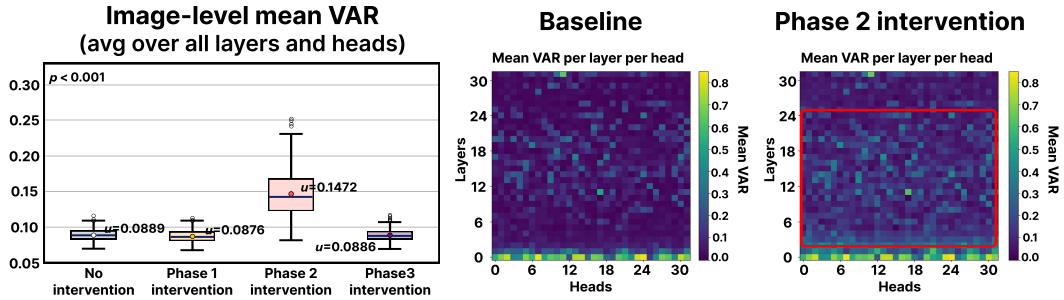

Figure 3: **Visual Attention Ratio analysis under phase-wise value interventions. Left:** Distribution of image-level mean VAR across intervention settings. Focus-phase value intervention yields a significantly higher VAR compared to the baseline ($p < 0.001$). **Right:** Layer-head VAR heatmaps showing increased visual attention in intermediate language-model layers after focus-phase value intervention.

where $A_k^{(l,h)}(a_{k,i})$ denotes the attention score from the generated token to the i-th visual token.

As shown in Fig. 3, focus-phase value intervention is accompanied by a significant increase in image-level mean VAR, from 0.0889 to 0.1472 ($p < 0.001$). In contrast, diffusion- and rediffusion-phase interventions produce negligible changes, with mean VAR values of 0.0876 and 0.0886, respectively. The layer-head heatmaps further show increased attention to visual tokens across multiple language-model layers and heads after focus-phase intervention.

Together with the CHAIR and teacher-forcing results, this suggests that neutralizing focus-phase low-attention values reduces hallucinated object likelihood and is accompanied by stronger decoder-side visual attention, while largely preserving ground-truth object information.

5 Experiments

Based on the attention-value analysis in Sec. 3 and the phase-wise value intervention study in Sec. 4, we evaluate the proposed focus-phase low-attention value replacement strategy. In the main paper, we focus on three aspects: hallucination mitigation and compatibility with existing methods, inference efficiency, and broader evaluation and capability preservation.

5.1 Experimental Setup

Models. We evaluate our method on four LVLM backbones with different architectures and scales: LLaVA-1.5-7B, LLaVA-1.5-13B [2], Qwen-2.5-VL-7B [5], and InternVL-2.5-8B [4]. LLaVA-1.5-7B and LLaVA-1.5-13B are widely used LVLMs based on CLIP [20] visual features, while Qwen-2.5-VL and InternVL-2.5 represent more recent LVLM architectures. Unless otherwise specified, all models are evaluated under standard greedy decoding.

Baselines. We compare our method with the original model output (Orig.) and Adversarial Uncertainty Estimation (AUE) [14]. AUE estimates unreliable visual tokens through iterative adversarial optimization and suppresses their influence during visual encoding. We also compare with four decoder-side hallucination mitigation methods—OPERA, VCD, PAI, and Devils—and assess compatibility by reporting each method alone, with AUE, and with ours.

Implementation Details. We select low-attention visual tokens using the average received attention scores computed from the layers immediately preceding the focus phase. In the focus phase, we replace the selected value vectors with an image-level mean value vector, computed per image, layer, and attention head by averaging over all visual tokens. We use a fixed 25% ratio across all backbones for the main hallucination benchmarks without model-specific tuning, while the broader MMHal-Bench evaluation uses nearby ratios in the 20–25% range. Appendix C.2 shows that the effect is generally robust to nearby ratios, with larger ratios trading caption fidelity for stronger CHAIR reduction. All LLaVA-1.5-13B experiments in Table 3 and all MME experiments in Fig. 4 were run on an NVIDIA RTX PRO 6000 (96GB); the remaining experiments were run on an NVIDIA RTX 5090 (32GB).

Table 3: **Cross-model comparison and compatibility with existing mitigation methods.** Yellow cells indicate our method. Parentheses report the AUE-relative difference Δ = Ours − AUE, or +Ours − (+AUE) in (b). Lower CHAIR_S/CHAIR_I and higher F1/POPE F1 are better; blue deltas favor our method and red deltas favor AUE. Metric definitions are provided in Appendix A.

(a) Comparison with AUE across LVLMs, including inference time.

Benchmark	Metric	LLaVA-1.5-7B			LLaVA-1.5-13B			Qwen-2.5-VL			InternVL-2.5		
		Orig.	AUE	Ours (Δ)	Orig.	AUE	Ours (Δ)	Orig.	AUE	Ours (Δ)	Orig.	AUE	Ours (Δ)
	Time (s)↓	3.207	7.307	**3.034**	6.387	14.992	**6.366**	2.598	33.462	**2.708**	2.299	41.291	**2.441**
CHAIR	CHAIR-S↓	45.0	30.2	30.8(+0.6)	41.0	29.8	32.2(+2.4)	31.4	27.8	29.8(+2.0)	28.6	28.4	27.8(-0.6)
	CHAIR-I↓	13.3	10.4	9.6(-0.8)	11.8	9.8	9.9(+0.1)	7.6	7.4	6.9(-0.5)	7.1	7.0	6.9(-0.1)
	F1↑	74.2	71.8	74.1(+2.3)	75.3	71.4	73.8(+2.4)	75.5	75.8	75.9(+0.1)	76.3	76.6	76.3(-0.3)
POPE	Random↑	87.4	85.6	86.4(+0.8)	86.6	82.4	86.01(+3.61)	81.3	79.6	81.4(+1.8)	94.27	94.17	94.4(+0.23)
	Popular↑	84.3	83.0	84.0(+1.0)	83.9	82.4	83.47(+1.07)	80.8	79.2	80.8(+1.6)	88.68	88.53	89.3(+0.77)
	Adversarial↑	79.3	78.4	79.6(+1.2)	79.5	76.5	78.13(+1.63)	80.5	78.7	80.4(+1.7)	89.6	85.73	85.7(-0.03)

(b) Comparative performance and compatibility with decoder-side hallucination mitigation methods.

Benchmark	Metric	OPERA			VCD			PAI			Devils		
		Orig.	+AUE	+Ours (Δ)	Orig.	+AUE	+Ours (Δ)	Orig.	+AUE	+Ours (Δ)	Orig.	+AUE	+Ours (Δ)
CHAIR	CHAIR-S↓	44.2	29.8	32.6(+2.8)	46.2	34.2	39.2(+5.0)	32.4	19.4	22.6(+3.2)	23.2	13.0	13.4(+0.4)
	CHAIR-I↓	13.5	11.2	10.6(-0.6)	14.9	11.8	12.9(+1.1)	9.62	7.61	7.50(-0.11)	7.88	7.24	5.90(-1.34)
	F1↑	74.4	72.3	73.5(+1.2)	71.3	70.5	70.7(+0.2)	74.6	67.7	71.8(+4.1)	72.0	63.5	67.7(+4.2)
POPE	Random↑	86.3	84.5	85.5(+1.0)	83.7	78.9	82.1(+3.2)	87.4	85.8	86.1(+0.3)	87.7	84.5	86.0(+1.5)
	Popular↑	83.3	81.9	82.8(+0.9)	80.9	77.1	79.5(+2.4)	84.5	83.2	83.9(+0.7)	85.3	83.0	84.3(+1.3)
	Adversarial↑	79.3	77.5	78.4(+0.9)	77.3	73.3	76.0(+2.7)	79.7	78.8	79.4(+0.6)	79.9	78.7	79.8(+1.1)

Benchmarks and Metrics. We evaluate object hallucination using CHAIR [6] and POPE [27]. CHAIR evaluates hallucination in image captioning and reports sentence-level CHAIR_S, instance-level CHAIR_I, and F1 for caption fidelity. POPE evaluates object hallucination through binary object-existence questions, and we report F1 scores on the Random, Popular, and Adversarial splits.

For broader hallucination evaluation, we use MMHal-Bench [22], which covers attributes, relations, and reasoning. We report its average score, hallucination rate, and attribute- and relation-level scores. Finally, we use MME [28] to assess general multimodal capabilities across perception and cognition tasks. Detailed protocols and metric definitions are provided in Appendix A.

5.2 Experimental Results

Hallucination mitigation performance and compatibility. Table 3 summarizes the hallucination mitigation performance of our method across multiple LVLM backbones and its compatibility with existing decoder-side mitigation methods. As shown in Table 3 (a), across all four backbones, our method consistently reduces both CHAIR_S and CHAIR_I relative to the original models while largely preserving caption-level F1. This trend holds for both LLaVA-1.5-7B/13B and more recent architectures such as Qwen-2.5-VL and InternVL-2.5. Compared with AUE, the two methods exhibit different trade-offs: AUE often achieves stronger CHAIR_S reduction, whereas our method generally preserves F1 and POPE performance more effectively and frequently achieves comparable or lower CHAIR_I. For example, on LLaVA-1.5-7B, our method reduces CHAIR_I from 13.3 to 9.6 while keeping F1 nearly unchanged (74.2 → 74.1), whereas AUE reduces F1 to 71.8.

Table 3 (b) further shows that our vision-encoder intervention can be combined with decoder-side methods including OPERA, VCD, PAI, and Devils. Although the additional CHAIR reduction varies across combinations, +Ours improves hallucination metrics over the corresponding standalone baselines and consistently preserves higher F1 and POPE performance than +AUE across these settings. These results indicate that the proposed intervention can serve as a lightweight complementary module rather than being tied to a particular decoding strategy.

Efficiency. As shown in Table 3 (a), our method introduces only negligible inference overhead compared with the original models. In contrast, AUE relies on iterative adversarial optimization and therefore incurs substantial latency. On Qwen-2.5-VL, AUE increases inference time from 2.598 s to 33.462 s, whereas our method requires only 2.708 s. Similarly, on InternVL-2.5, AUE increases runtime from 2.299 s to 41.291 s, while ours requires 2.441 s. This substantial efficiency

Table 4: **Broader hallucination evaluation on MMHal-Bench.**

Model	Avg. Score ↑		Hallu. Rate ↓		Attribute ↑		Relation ↑	
	Orig.	Ours	Orig.	Ours	Orig.	Ours	Orig.	Ours
LLaVA-1.5-13B	2.04	**2.08**	0.604	**0.604**	2.42	**3.08**	1.58	**1.83**
Qwen-2.5-VL	4.18	**4.21**	0.260	**0.240**	3.83	**3.83**	4.50	**4.58**
InternVL-2.5	3.73	**3.83**	0.260	**0.229**	3.42	**3.67**	4.17	**4.17**

advantage stems from the single-pass nature of our intervention, which requires neither iterative optimization nor additional training. Its negligible overhead is particularly beneficial when combined with decoder-side mitigation methods, as it adds little additional latency on top of their existing inference cost.

Broader evaluation and capability preservation. To examine whether the effect extends beyond object-existence hallucination, we further evaluate our method on MMHal-Bench using LLaVA-1.5-13B, Qwen-2.5-VL-7B, and InternVL-2.5-8B. For this broader evaluation, we use low-attention token ratios ranging from 20% to 25%, while keeping the remaining intervention settings unchanged. As shown in Table 4, the average score improves across all three backbones, while all reported attribute- and relation-level scores are maintained or improved. The hallucination rate also decreases on Qwen-2.5-VL and InternVL-2.5 while remaining unchanged on LLaVA-1.5-13B. Together, these results indicate that the effect of our intervention extends beyond object-existence errors to broader attribute- and relation-related hallucination settings.

We additionally evaluate general multimodal capability on MME. As shown in Fig. 4, our method largely preserves the original category-level performance profiles of Qwen-2.5-VL and InternVL-2.5 across both perception- and cognition-oriented categories. Although small category-wise fluctuations are observed, the overall profiles remain similar, suggesting that the intervention mitigates hallucination without broadly degrading general multimodal capability.

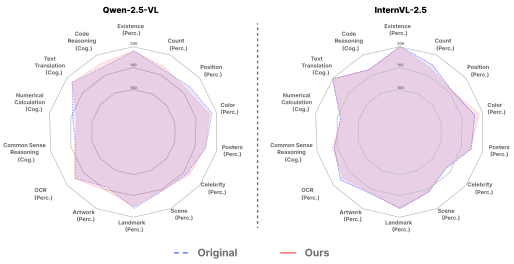

Figure 4: **Category-level MME comparison between the original models and our method.** Our method largely preserves the original MME performance profile across Qwen-2.5-VL and InternVL-2.5.

Overall, focus-phase low-attention value replacement provides a favorable balance of hallucination reduction, output fidelity, and inference efficiency across diverse LVLMs, while remaining compatible with decoder-side mitigation and largely preserving general multimodal capabilities.

6 Conclusion

In this paper, we studied object hallucination in LVLMs through the lens of attention-value dynamics inside vision encoders. Our analysis identified a consistent diffusion–focus–rediffusion structure and showed that the focus phase is the stage where attention most clearly separates strongly supported tokens from weakly supported ones. We further found that low-attention tokens in this phase can still exert non-negligible value-side influence relative to their small attention mass, revealing an attention-value mismatch that is not captured by attention scores alone. Through phase-wise value interventions, we showed that modifying focus-phase low-attention values reduces hallucination metrics while largely preserving caption fidelity. Token-level teacher-forcing analysis further demonstrated that this intervention decreases hallucinated object likelihood with much smaller effects on ground-truth object tokens. Motivated by these observations, we used a simple training-free intervention that replaces focus-phase low-attention values with an image-level mean value vector. Across multiple LVLM backbones, this intervention reduces object hallucination with negligible additional runtime. These findings suggest that hallucination mitigation can benefit from analyzing not only which visual tokens receive attention, but also how value-side signals from weakly supported tokens are propagated through the vision encoder.

Acknowledgments and Disclosure of Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2026-RS-2026-25615271, AI Star Fellowship Support Program). Computational resources were provided by "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT) and the Cluster Server for Computational Science at Pusan National University.

References

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 34892–34916, Curran Associates, Inc., 2023.

[2] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

[3] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, "Shikra: Unleashing multimodal llm's referential dialogue magic," *arXiv preprint arXiv:2306.15195*, 2023.

[4] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[6] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

[7] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng, "A survey on hallucination in large vision-language models," *arXiv preprint arXiv:2402.00253*, 2024.

[8] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, and P. He, "Dola: Decoding by contrasting layers improves factuality in large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[9] S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing, "Mitigating object hallucinations in large vision-language models through visual contrastive decoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

[10] Q. Huang, X. Dong, P. Zhang, B. Wang, C. He, J. Wang, D. Lin, W. Zhang, and N. Yu, "Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

[11] S. Liu, K. Zheng, and W. Chen, "Paying more attention to image: A training-free method for alleviating hallucination in lvlms," in *European Conference on Computer Vision*, pp. 125–140, Springer, 2024.

[12] Z. Jiang, J. Chen, B. Zhu, T. Luo, Y. Shen, and X. Yang, "Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25004–25014, 2025.

[13] C. Baek, J. Song, S. Kim, and K. Kong, "Agilepruner: An empirical study of attention and diversity for adaptive visual token pruning in large vision-language models," in *International Conference on Learning Representations* (C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, eds.), vol. 2026, pp. 66116–66144, 2026.

[14] H. Seo, D. U. Kang, H. Cho, J. Lee, and S. Y. Chun, "On epistemic uncertainty of visual tokens for object hallucinations in large vision-language models," *Advances in neural information processing systems*, 2025.

[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *The Sixth International Conference on Learning Representations*, 2018.

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[17] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[18] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *The Ninth International Conference on Learning Representations*, 2021.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 18–24 Jul 2021.

[21] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

[22] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell, "Aligning large multimodal models with factually augmented RLHF," in *Findings of the Association for Computational Linguistics: ACL 2024* (L.-W. Ku, A. Martins, and V. Srikumar, eds.), (Bangkok, Thailand), pp. 13088–13110, Association for Computational Linguistics, Aug. 2024.

[23] Z. Zhao, B. Wang, L. Ouyang, X. Dong, J. Wang, and C. He, "Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization," *arXiv preprint arXiv:2311.16839*, 2023.

[24] Y. Xie, G. Li, X. Xu, and M.-Y. Kan, "V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024.

[25] Z. Yang, X. Luo, D. Han, Y. Xu, and D. Li, "Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10610–10620, 2025.

[26] Y. Fu, R. Xie, X. Sun, Z. Kang, and X. Li, "Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization," in *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16563–16577, 2025.

[27] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 292–305, Association for Computational Linguistics, Dec. 2023.

[28] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our approach to hallucination mitigation is sufficiently described in the abstract and introduction. In particular, we summarize our contributions at the end of the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Appendix D

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs. The equations in the paper are used to define analysis metrics and intervention procedures, and their assumptions and definitions are described in the main text and supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide code for reproducibility in the supplementary material, and we describe the methodology, benchmarks, and baseline models used in our experiments in detail in Sec 5.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide our code and reproduction instructions in the supplementary material. All evaluation data are from publicly available benchmarks, including COCO val2014, POPE, MME, and MMHal-Bench, and we describe the data access, preprocessing, and evaluation settings needed to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the evaluated models, benchmarks, data splits, decoding settings, intervention hyperparameters, and runtime measurement details in Sec. 5.1. Since our method is training-free, no optimizer or training hyperparameters are required.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report statistical significance for the VAR analysis and provide 95% confidence intervals and two-sided p-values using paired image-level bootstrap resampling for matched mean-replacement controls C.1.1. We further report exact two-sided permutation Spearman tests for continuous value modulation and Holm-corrected significance results for matched amplification controls C.1.2. Other benchmark evaluations use frozen models, deterministic decoding, fixed random seeds, and fixed evaluation splits; therefore, we do not report run-to-run error bars for every benchmark table.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify the GPU specifications used for each experiment in main paper 5.1, where one RTX 5090 and one RTX PRO 6000 were used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss that reducing object hallucination can improve the reliability of LVLMs, while also noting risks such as over-reliance on imperfect mitigation and potential misuse in sensitive applications. These broader impacts are discussed in Appendix E, and limitations are discussed in Appendix D.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our paper does not introduce any new models or datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original papers for the existing LVLMs, baselines, and benchmark datasets used in our experiments. We use these assets for research evaluation only and do not redistribute modified versions of them.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide a README file in our code, included in the supplementary material, that describes the usage instructions, benchmarks, and models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: Our paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The development of our method does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Organization of the Appendices

These appendices provide detailed benchmark definitions, extended attention-value analyses, additional experimental validation, and discussions of limitations and broader impacts to complement the main manuscript.

The document is organized as follows:

- **Appendix A** describes detailed benchmark definitions used in our experiments.
- **Appendix B** provides details of the attention-value analysis, including phase boundary determination, the definition of the low-attention value influence metric, and additional influence plots on representative backbones.
- **Appendix C** presents additional analyses of focus-phase low-attention value intervention, including matched controls and statistical validation, continuous value modulation, sensitivity to the low-attention token ratio, and robustness to focus-phase boundaries.
- **Appendix D** discusses the limitations and future directions of our work.
- **Appendix E** discusses the broader impacts of our work.

A Benchmarks

CHAIR. CHAIR [6] evaluates object hallucination in image captioning by comparing generated object mentions with ground-truth object annotations. Following the standard protocol, we generate captions for 500 images sampled from COCO val2014 using the prompt "Please describe this image in detail." We report sentence-level CHAIR_S, instance-level CHAIR_I, and F1. CHAIR_S measures the proportion of generated captions that contain at least one hallucinated object, while CHAIR_I measures the proportion of hallucinated objects among all mentioned objects:

$$\text{CHAIR}_S = \frac{|\{\text{captions containing hallucinated objects}\}|}{|\{\text{all captions}\}|}, \quad \text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}. \tag{10}$$

Lower CHAIR_S and CHAIR_I indicate fewer hallucinations. We also report F1 to measure caption fidelity with respect to ground-truth object annotations.

POPE. POPE [27] evaluates object hallucination through binary object-existence questions. We follow the standard POPE protocol and report F1 scores on the Random, Popular, and Adversarial splits. The Random split samples negative objects randomly, the Popular split uses frequently occurring objects as negatives, and the Adversarial split selects more challenging negative objects that are likely to be confused with the image content. Higher F1 scores indicate stronger object-level grounding and better resistance to object hallucination.

MMHal-Bench. MMHal-Bench [22] evaluates hallucination in practical multimodal question-answering settings across diverse question categories and object topics. Following the standard evaluation protocol, open-ended model responses are assessed for both overall response quality and hallucination. We report the average response score, evaluated on a 0–6 scale, and the hallucination rate, which measures the frequency of hallucinated responses. To further examine broader hallucination behavior, we additionally report scores for attribute- and relation-related categories. Higher average, attribute, and relation scores and a lower hallucination rate indicate better performance.

MME. MME [28] evaluates general multimodal capabilities beyond hallucination-specific captioning benchmarks. We use MME to examine whether the proposed intervention preserves category-level performance across both perception-oriented and cognition-oriented tasks. The perception-oriented categories include object existence, count, position, color, OCR, posters, celebrity, scene, artwork, and landmark. The cognition-oriented categories include commonsense reasoning, numerical calculation, text translation, and code reasoning. We compare the category-level MME profiles of the original models and our method to verify that hallucination mitigation does not globally degrade general multimodal capability.

B Details of Attention-Value Analysis

B.1 Phase Boundary Determination

In Sec. 3.1 of the main manuscript, we empirically observed a consistent three-phase visual processing structure—Diffusion, Focus, and Rediffusion—across multiple LVLM vision encoders. To clarify how the phase boundaries used in our experiments are determined, we describe the procedure for identifying the focus-phase interval $\mathcal{F} = [\ell_{\text{start}}, \ell_{\text{end}}]$ based on the layer-wise dynamics of the attention concentration ratio.

The concentration ratio at layer ℓ is defined as

$$R^{(\ell)} = \frac{M^{(\ell)}}{H^{(\ell)}}, \tag{11}$$

where $M^{(\ell)}$ denotes the maximum attention score and $H^{(\ell)}$ denotes the attention entropy at layer ℓ. This ratio increases when attention becomes concentrated on a small subset of tokens and decreases when attention is more evenly distributed.

B.1.1 Onset of the Focus Phase

During the initial diffusion phase, attention is broadly distributed and the layer-wise change of the concentration ratio remains relatively small. To identify the transition into the focus phase, we examine the discrete forward difference

$$\Delta R^{(\ell)} = R^{(\ell)} - R^{(\ell-1)}. \tag{12}$$

The start layer ℓ_{start} is determined as the first layer where the increase in concentration exceeds the baseline variation observed in the early layers:

$$\ell_{\text{start}} = \min \left\{ \ell \mid \Delta R^{(\ell)} > \mu_{\text{base}} + \lambda \cdot \sigma_{\text{base}} \right\}, \tag{13}$$

where μ_{base} and σ_{base} denote the mean and standard deviation of $\Delta R^{(\ell)}$ computed from the early diffusion layers, e.g., the first 25% of the network depth, and λ is a confidence multiplier. We set $\lambda = 2$ for all evaluated models; using $\lambda = 3$ yields identical focus-phase intervals. This criterion identifies the layer where attention concentration begins to increase beyond the relatively stable baseline behavior of the diffusion stage.

B.1.2 Offset of the Focus Phase

After identifying ℓ_{start}, we determine the end layer of the focus phase by using a proportional window over the vision encoder depth. Specifically, the focus-phase length is defined as

$$K = \text{round}(0.3L), \tag{14}$$

where L denotes the total number of vision-encoder layers. The end layer is then given by

$$\ell_{\text{end}} = \ell_{\text{start}} + K - 1. \tag{15}$$

This proportional window captures the main high-concentration region following the onset of the focus phase, while avoiding model-specific tuning based on downstream hallucination metrics. Layers before ℓ_{start} are treated as the diffusion phase, and layers after ℓ_{end} are treated as the rediffusion phase.

B.1.3 Phase Boundaries Across Evaluated Models

Applying the above procedure yields the layer intervals used in our experiments, as visualized in Fig. 5. The boundaries are determined from attention concentration dynamics, not from validation performance or hallucination metrics.

LLaVA-1.5-7B and LLaVA-1.5-13B ($L = 24$). For both LLaVA-1.5 backbones, the concentration change $\Delta R^{(\ell)}$ remains close to zero in early layers and exhibits a clear increase around layer 11. With $\ell_{\text{start}} = 11$ and $K = \text{round}(0.3L) = 7$, the focus-phase interval becomes layers 11–17.

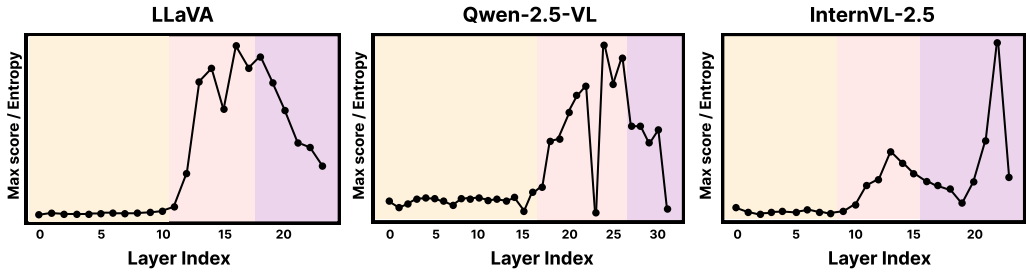

Figure 5: **Layer-wise attention dynamics across evaluated LVLM backbones.** The solid lines indicate the ratio of maximum attention score to entropy ($R^{(l)}$). The shaded regions represent the identified Diffusion (yellow), Focus (red), and Rediffusion (purple) phases based on our gradient-based boundary formulation.

Qwen-2.5-VL ($L = 32$). For Qwen-2.5-VL, the concentration change exceeds the early baseline around layer 17. With $\ell_{\text{start}} = 17$ and $K = \text{round}(0.3L) = 10$, the focus-phase interval becomes layers 17–26, covering the dominant high-concentration region before the concentration ratio decreases. The sharp dips at indices 7, 15, 23, and 31 coincide with the only layers that use full attention; the remaining layers use window attention. Full attention operates over a broader set of tokens and can temporarily reduce the measured concentration. As seen around indices 22–24, the concentration recovers in neighboring window-attention layers. We therefore interpret these dips as local effects of the attention-pattern switch rather than additional phase boundaries.

InternVL-2.5 ($L = 24$). For InternVL-2.5, the first consistent increase in $\Delta R^{(\ell)}$ occurs around layer 9. With $\ell_{\text{start}} = 9$ and $K = \text{round}(0.3L) = 7$, the focus-phase interval becomes layers 9–15. Although a secondary spike appears in later layers, it is largely confined to CLS-to-patch attention and diminishes when averaged over patch queries. Because InternVL-2.5 discards the CLS token before projecting patch features to the language model, we use the first dominant concentration region, which marks the initial transition from broadly distributed attention to the focus phase.

B.2 Definition of Low-Attention Value Influence

We provide the detailed formulation of the low-attention value influence metric used in Sec. 3.2. For layer ℓ and head h, let $\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{N\times N}$ denote the visual self-attention matrix and $\mathbf{V}^{(\ell,h)} \in \mathbb{R}^{N\times d_h}$ denote the corresponding value matrix. The received attention mass of token j is defined as

$$s_j^{(\ell)} = \frac{1}{|\mathcal{H}|N} \sum_{h\in\mathcal{H}} \sum_{i=1}^{N} A_{i,j}^{(\ell,h)}. \tag{16}$$

We define the low-attention token set $\mathcal{L}^{(\ell)}$ as the bottom 25% of tokens according to $s_j^{(\ell)}$.

The attention mass assigned to the low-attention group is

$$\alpha_{\mathcal{L}}^{(\ell)} = \frac{1}{|\mathcal{H}|N} \sum_{h\in\mathcal{H}} \sum_{i=1}^{N} \sum_{j\in\mathcal{L}^{(\ell)}} A_{i,j}^{(\ell,h)}. \tag{17}$$

The attention-output contribution from the low-attention group is

$$\mathbf{O}_{\mathcal{L}}^{(\ell,h)} = \mathbf{A}_{:,\mathcal{L}}^{(\ell,h)} \mathbf{V}_{\mathcal{L}}^{(\ell,h)}. \tag{18}$$

We normalize this contribution by the full attention output:

$$\eta_{\mathcal{L}}^{(\ell)} = \frac{\sum_{h\in\mathcal{H}} \left\| \mathbf{O}_{\mathcal{L}}^{(\ell,h)} \right\|_F}{\sum_{h\in\mathcal{H}} \left\| \mathbf{A}^{(\ell,h)} \mathbf{V}^{(\ell,h)} \right\|_F + \epsilon}. \tag{19}$$

Finally, the influence per attention mass is defined as

$$G_{\mathcal{L}}^{(\ell)} = \frac{\eta_{\mathcal{L}}^{(\ell)}}{\alpha_{\mathcal{L}}^{(\ell)} + \epsilon}. \tag{20}$$

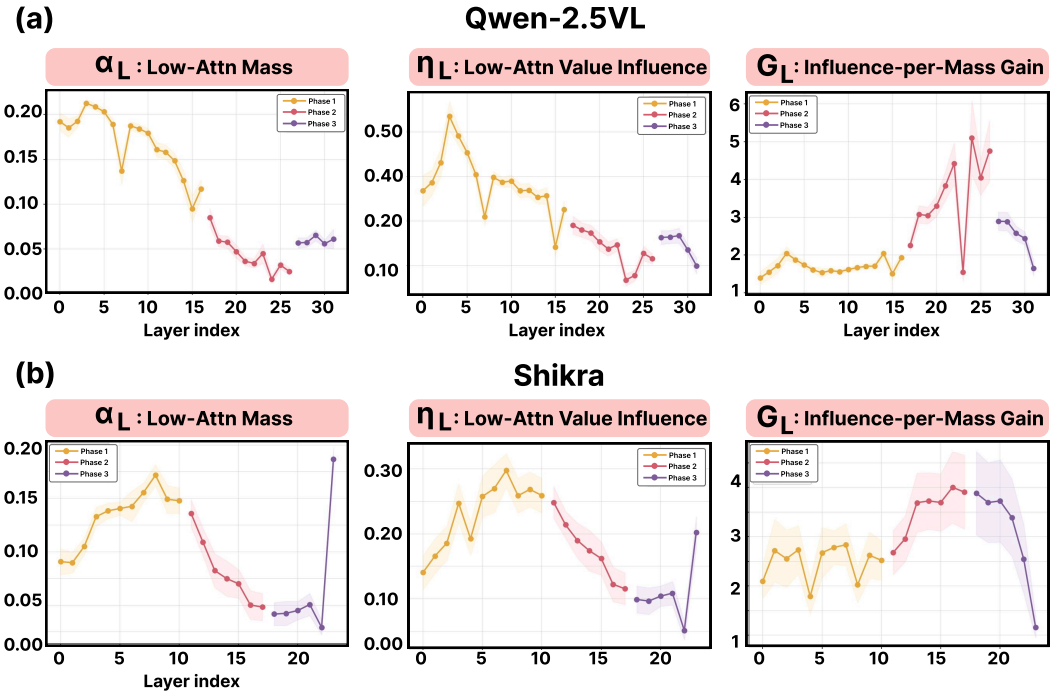

Figure 6: **Additional low-attention value influence analysis.** We report $\alpha_{\mathcal{L}}$, $\eta_{\mathcal{L}}$, and $G_{\mathcal{L}}$ for Qwen-2.5-VL and Shikra-7B. Both models show a similar qualitative trend to the main LLaVA analysis: around the focus phase, the attention mass assigned to low-attention tokens decreases, while their normalized value contribution remains non-negligible, increasing the influence per attention mass. The strength of the effect varies across architectures.

B.3 Additional Low-Attention Value Influence Analysis

As shown in Fig. 6, we provide additional low-attention value influence plots for Qwen-2.5-VL and Shikra-7B. Both models show the same qualitative behavior as the main LLaVA analysis: the attention mass assigned to low-attention tokens decreases around the focus phase, while their normalized value contribution remains non-negligible, leading to an increased influence per attention mass. The magnitude of this effect varies across architectures, and we therefore use these plots as diagnostic evidence rather than as a strict universality claim.

C Additional Analysis and Robustness of Focus-Phase Low-Attention Intervention

C.1 Specificity and Statistical Validation of Low-Attention Value Intervention

C.1.1 Mean-replacement matched controls

In Sec. 4.1 of the main paper, we showed that hallucination is consistently reduced when the token-specific content of focus-phase low-attention values is removed or disrupted using different intervention strategies. This suggests that the observed effect does not depend on a particular replacement value. Here, we further examine whether the selection of low-attention tokens itself is important for this effect.

To this end, we perform matched control experiments in which the intervention phase, token budget, and replacement operation are fixed, while only the selected visual-token group is varied. Specifically, within each phase, we apply the same mean-replacement operation to the same 25% token budget and compare three token-selection strategies: *Low*, which selects the bottom 25% of visual tokens according to received attention mass; *Random*, which uniformly samples 25% of the visual tokens; and *High*, which selects the top 25% according to received attention mass. We perform this comparison

Table 5: **Mean-replacement matched controls across intervention phases.** We fix the intervention phase, token budget (25%), and mean-replacement operation, and vary only the selected visual-token group. Random results are reported as mean ± standard deviation. Δ denotes Low minus the corresponding control.

Phase	Metric	Baseline	Low	Random	High	Low–Random Δ [95% CI], p	Low–High Δ [95% CI], p
P1	$\text{CHAIR}_S \downarrow$	45.0	44.0	45.8 ± 0.3	44.0	-1.8 [-5.4, $+1.9$], $p = .32$	$+0.0$ [-4.6, $+4.4$], $p = 1.00$
	$\text{CHAIR}_I \downarrow$	13.3	16.2	15.2 ± 0.3	14.6	$+1.0$ [-0.6, $+2.6$], $p = .22$	$+1.7$ [-0.3, $+3.6$], $p = .10$
P2	$\text{CHAIR}_S \downarrow$	45.0	**30.8**	39.6 ± 2.3	42.4	$\mathbf{-8.8}$ [$\mathbf{-12.0}$, $\mathbf{-5.6}$], $p < .001$	$\mathbf{-11.6}$ [$\mathbf{-16.6}$, $\mathbf{-6.6}$], $p < .001$
	$\text{CHAIR}_I \downarrow$	13.3	**9.6**	12.6 ± 0.5	18.1	$\mathbf{-2.9}$ [$\mathbf{-4.1}$, $\mathbf{-1.8}$], $p < .001$	$\mathbf{-8.5}$ [$\mathbf{-10.7}$, $\mathbf{-6.3}$], $p < .001$
P3	$\text{CHAIR}_S \downarrow$	45.0	45.2	43.9 ± 1.2	49.4	$+1.3$ [-1.5, $+4.1$], $p = .36$	-4.2 [-8.4, $+0.0$], $p = .06$
	$\text{CHAIR}_I \downarrow$	13.3	13.2	13.2 ± 0.3	14.7	$+0.0$ [-0.9, $+1.0$], $p = .95$	-1.5 [-3.0, $+0.1$], $p = .06$

Table 6: **Continuous modulation of low-attention value contribution.** We amplify the contribution of focus-phase low-attention values by a scalar α while keeping the original value vectors and attention weights unchanged. As α increases, CHAIR_S and CHAIR_I increase, while F1 decreases.

Metric	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 2.0$	$\alpha = 2.5$	$\alpha = 3.0$	$\alpha = 4.0$	$\alpha = 5.0$
$\text{CHAIR}_S \downarrow$	45.0	47.4	48.4	49.0	48.0	54.2	57.6
$\text{CHAIR}_I \downarrow$	13.3	14.2	14.7	15.8	18.4	29.0	46.5
F1 ↑	74.2	73.3	72.8	70.8	67.9	56.3	34.7

independently in the diffusion (P1), focus (P2), and rediffusion (P3) phases. For the Random control, we report the mean and standard deviation across repeated runs. For the difference between Low and each matched control, we compute 95% confidence intervals and two-sided p-values using paired image-level bootstrap resampling.

As shown in Table 5, a significant advantage of low-attention selection emerges only in the focus phase (P2). For CHAIR_S, Low achieves 30.8, which is 8.8 points lower than Random at 39.6 ± 2.3 (95% CI: $[-12.0, -5.6]$) and 11.6 points lower than High at 42.4 (95% CI: $[-16.6, -6.6]$); both differences are significant with $p < 0.001$. For CHAIR_I, Low achieves 9.6, which is 2.9 and 8.5 points lower than Random at 12.6 ± 0.5 and High at 18.1, respectively. Both differences are again statistically significant ($p < 0.001$).

In contrast, in the diffusion (P1) and rediffusion (P3) phases, Low does not show a statistically significant advantage over either the Random or High control. These matched controls therefore indicate that the hallucination reduction observed in the main paper is not a generic consequence of applying mean replacement to a fixed number of visual tokens, but is closely associated with selecting low-attention tokens specifically during the focus phase.

C.1.2 Continuous Modulation of Low-Attention Value Signals

In the main manuscript, we show that neutralizing focus-phase low-attention values reduces hallucination while largely preserving caption fidelity. Here, we perform a complementary diagnostic experiment: instead of suppressing or replacing these values, we progressively amplify their contribution to the attention output and examine how hallucination metrics change.

We conduct this experiment on LLaVA-1.5-7B using 500 images from COCO [29] val2014. We identify the low-attention token group using the same criterion as in the main experiments: the bottom 25% visual tokens according to pre-focus received attention mass, with the CLS token protected. The modulation is then applied to the focus-phase target layers.

Unlike replacement-based interventions, this experiment keeps the original value vectors unchanged. We only modulate the contribution of the low-attention group to the attention output by a scalar α:

$$\mathbf{O}(\alpha) = \mathbf{A}_{\mathcal{H}}\mathbf{V}_{\mathcal{H}} + \alpha \cdot \mathbf{A}_{\mathcal{L}}\mathbf{V}_{\mathcal{L}}, \tag{21}$$

where $\mathcal{H}$ denotes the remaining visual tokens and $\mathcal{L}$ denotes the low-attention token group. The attention weights and all value vectors are kept unchanged; only the gain applied to the low-attention value contribution is varied. We sweep $\alpha \in \{1, 1.5, 2, 2.5, 3, 4, 5\}$, where $\alpha = 1$ corresponds to the original model.

Table 7: **Matched amplification controls.** Differences in CHAIR_S between low-attention amplification and norm-matched high-attention or random-token amplification. Positive values indicate a larger hallucination increase for the low-attention group.

α	Low–High Δ CHAIR_S [95% CI]	Low–Random Δ CHAIR_S [95% CI]
1.25	$+1.8$ [-1.6, $+5.2$]	$+0.9$ [-1.7, $+3.4$]
1.50	$+\mathbf{6.6}$ [$+2.8$, $+10.6$]	$+2.9$ [$+0.3$, $+5.9$]
1.75	$+\mathbf{6.4}$ [$+2.0$, $+11.2$]	$+\mathbf{5.2}$ [$+2.3$, $+8.3$]
2.00	$+\mathbf{12.6}$ [$+8.2$, $+17.2$]	$+\mathbf{5.7}$ [$+2.6$, $+8.9$]

Table 8: **Sensitivity to the low-attention token ratio.** We vary the fraction of low-attention tokens used for focus-phase value replacement and report CHAIR_S, CHAIR_I, and F1. Larger ratios can further reduce CHAIR metrics, but may introduce a stronger fidelity trade-off depending on the model. We use 25% as the default ratio because it provides a conservative balance between hallucination reduction and caption fidelity without model-specific tuning.

Ratio	LLaVA-1.5-7B			Qwen-2.5-VL			InternVL-2.5		
	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	F1 $\uparrow$	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	F1 $\uparrow$	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	F1 $\uparrow$
Baseline	45.0	13.3	74.2	31.4	7.6	75.5	28.6	7.1	76.3
20%	36.2	11.3	74.0	30.0	7.4	75.2	29.0	7.5	76.3
25%	30.8	9.6	74.1	29.8	6.9	75.9	27.8	6.9	76.3
30%	27.8	9.4	72.5	28.4	6.7	76.1	27.2	6.6	76.3
35%	25.4	8.8	70.4	26.8	6.7	76.0	26.0	6.8	76.2
40%	22.6	8.0	69.1	27.8	6.8	76.4	27.2	7.3	75.8

As shown in Table 6, increasing the contribution of low-attention values generally worsens hallucination while reducing caption fidelity. As α increases from 1 to 5, CHAIR_S rises from 45.0 to 57.6 and CHAIR_I increases more sharply from 13.3 to 46.5, while F1 decreases from 74.2 to 34.7. Across the seven amplification settings, exact two-sided permutation Spearman tests confirm that these trends are statistically significant: CHAIR_S shows $\rho = 0.893$ ($p = 0.012$), CHAIR_I shows $\rho = 1.000$ ($p < 0.001$), and F1 shows $\rho = -1.000$ ($p < 0.001$). These results demonstrate a consistent relationship between increasing low-attention value contribution and hallucination, providing complementary evidence to the value-replacement experiments in the main paper.

We next perform matched amplification controls to examine whether this sensitivity is specific to the low-attention group. We amplify the contributions of high-attention and random-token groups while norm-matching the perturbation magnitude to that of the low-attention condition. As shown in Table 7, low-attention amplification induces a larger increase in CHAIR_S than high-attention and random-token amplification as α increases. After Holm correction, the low-attention condition produces a significantly larger increase in hallucination than the high-attention condition for $\alpha \geq 1.5$, and than the random-token condition for $\alpha \geq 1.75$.

These results indicate that the observed sensitivity is not merely a generic effect of amplification, but is closely associated with focus-phase low-attention value signals. Importantly, this experiment does not modify the token content itself; it only changes the gain applied to their contribution to the attention output. Thus, the results support the interpretation that focus-phase low-attention value signals can influence hallucination behavior without making the stronger claim that all low-attention tokens are inherently harmful.

C.2 Sensitivity to the Low-Attention Token Ratio

In the main experiments, we use the bottom 25% visual tokens as the low-attention group for focus-phase value replacement. To examine how the intervention behaves under different token ratios, we vary the replacement ratio from 20% to 40% and evaluate CHAIR_S, CHAIR_I, and F1 across multiple LVLM backbones.

As shown in Table 8, increasing the ratio generally strengthens hallucination suppression, but the fidelity trade-off differs across models. On LLaVA-1.5-7B, larger ratios further reduce hallucination,

Table 9: **Robustness to focus-phase boundaries.** We perturb the default focus-phase interval by contracting or expanding both boundaries by one layer while keeping all other intervention settings fixed.

Model	Setting	Target layers	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	F1 $\uparrow$
LLaVA-1.5-7B	No intervention	–	45.0	13.3	74.2
	Contracted	[12–16]	33.0	10.3	73.6
	Default	[11–17]	30.8	9.6	74.1
	Expanded	[10–18]	32.0	10.0	72.3
Qwen-2.5-VL-7B	No intervention	–	31.4	7.6	75.5
	Contracted	[18–25]	28.2	7.1	76.1
	Default	[17–26]	29.8	6.9	75.9
	Expanded	[16–27]	26.0	6.6	76.7
InternVL-2.5-8B	No intervention	–	28.6	7.1	76.3
	Contracted	[10–14]	27.4	6.9	76.3
	Default	[9–15]	27.8	6.9	76.3
	Expanded	[8–16]	27.8	6.9	75.8

with CHAIR_S decreasing from 30.8 at 25% to 22.6 at 40%, and CHAIR_I decreasing from 9.6 to 8.0. However, this stronger hallucination reduction comes with a clear F1 drop from 74.1 to 69.1, indicating that aggressive replacement may neutralize value signals that remain useful for caption generation.

For Qwen-2.5-VL, CHAIR metrics remain lower than or close to the original baseline across all ratios, while F1 changes only marginally. On InternVL-2.5, ratios between 25% and 35% consistently reduce CHAIR metrics while maintaining F1 close to the baseline; at 40%, the CHAIR_I improvement becomes weaker. These results suggest that the intervention is not dependent on a finely tuned ratio, although overly aggressive replacement can introduce a stronger fidelity trade-off.

Based on these observations, we use 25% as the default ratio in the main experiments. Although larger ratios can sometimes produce stronger CHAIR reduction, especially for LLaVA-1.5-7B and Qwen-2.5-VL, they may also increase the risk of fidelity degradation depending on the model. The 25% setting therefore provides a conservative balance between hallucination reduction and caption fidelity without model-specific ratio tuning.

C.3 Robustness to Focus-Phase Boundaries

In the main paper, we define the focus phase of each model based on its attention dynamics and apply the low-attention value intervention within the corresponding layer range. Here, we examine whether the observed hallucination reduction is sensitive to the exact focus-phase boundaries by shifting the start and end boundaries of the default focus interval inward or outward by one layer. All other token-selection and intervention settings are kept identical to those used in the main experiments.

As shown in Table 9, the effect of the intervention is largely preserved under these boundary perturbations. For LLaVA-1.5-7B, both CHAIR_S and CHAIR_I remain lower than the no-intervention baseline not only for the default interval $[11, 17]$, but also for the contracted $[12, 16]$ and expanded $[10, 18]$ settings. The same trend is observed for Qwen-2.5-VL-7B and InternVL-2.5-8B, where both CHAIR metrics remain below the corresponding baselines across all tested boundary settings.

Performance under the contracted and expanded intervals also remains broadly comparable to that of the default setting, while F1 is generally maintained at a similar level to the original models. These results indicate that the effectiveness of the proposed intervention does not depend strongly on a single exact layer boundary and remains stable under small perturbations around the identified focus phase.

D Limitations and Future Directions

Despite its favorable hallucination reduction–efficiency trade-off, our method has several limitations. First, focus-phase low-attention value replacement can weaken visual signals that are useful for fine-grained perception. As shown in the ratio analysis, increasing the replacement ratio can further reduce CHAIR metrics but may also decrease F1, indicating a trade-off between hallucination suppression and caption fidelity. Therefore, a fixed replacement ratio may not be optimal for all images, prompts, or downstream tasks. Developing adaptive image- or prompt-dependent replacement strategies is an important direction for future work.

Second, our analysis provides diagnostic evidence rather than a formal causal proof of hallucination mechanisms. Although the diffusion–focus–rediffusion pattern is observed across multiple LVLM backbones, the phase boundaries are determined by an empirical attention-concentration criterion rather than a theoretical definition. A more principled characterization of phase transitions inside vision encoders could further clarify why the focus phase is particularly sensitive to low-attention value-side signals.

Third, our evaluations mainly focus on image-based object hallucination. In more complex settings such as video understanding, 3D perception, multi-image reasoning, and dense spatial reasoning, low-attention tokens may contain small, rare, or contextually important visual evidence. In such cases, more careful adaptive interventions may be required to reduce hallucination while preserving task-relevant information.

Finally, improvements in generative hallucination metrics do not always directly translate to improvements in binary object-existence evaluation. For example, POPE changes can be mixed across splits, suggesting that reducing hallucinated object mentions and improving yes/no object grounding are related but not identical objectives. Future work could study how value-side interventions affect calibration and object-existence decision boundaries.

E Broader impacts

Our work aims to improve the reliability of LVLMs by reducing object hallucination in image-grounded generation. This can have positive impacts in applications that require faithful visual grounding, such as assistive technologies, educational tools, robotics, and human-AI decision support systems. By reducing unsupported object mentions, hallucination mitigation can make LVLM outputs easier to interpret and less likely to mislead users.

However, our method does not eliminate hallucinations completely and should not be viewed as a guarantee of factual correctness. Users may over-rely on model outputs if hallucination mitigation is mistaken for complete reliability, especially in sensitive domains such as medical, legal, robotics, or safety-critical decision-making. In addition, deployment of LVLMs in surveillance or automated monitoring can raise privacy and fairness concerns that are outside the scope of our method. Thus, the proposed intervention should be used with appropriate human oversight, task-specific validation, and safeguards in high-stakes settings.