RELIABLE MEMORY INJECTION INTO CORE VISUAL ATTENTION HEADS FOR OBJECT HALLUCINATION MITIGATION IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) often hallucinate objects that are absent from the input image during open-ended generation. A primary cause is the progressive weakening of visual evidence during autoregressive decoding, which allows language priors to dominate. Existing methods typically enhance visual information at the model or layer level while largely overlooking functional differences among attention heads. To address this limitation, we first analyze visual perception in LVLMs from the perspective of multihead attention and identify Core Visual Attention Heads (CVAHs) through statistics collected across multiple samples. We then introduce Core Visual Attention Head Memory Injection (CAMI), a training-free method. CAMI uses relative visual-attention decay as a state-dependent intervention scheduler, not as a classifier of hallucinated tokens. When triggered, it injects historical visual pattern memory to correct the spatial attention of CVAHs. Experiments on MSCOCO-CHAIR and AMBER-G show that CAMI consistently mitigates object hallucination and outperforms existing state-of-the-art baselines.

1 INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

In recent years, large vision-language models (LVLMs) have achieved substantial progress in image captioning, visual question answering, and multimodal dialogue. Representative models such as Qwen, LLaVA, and InstructBLIP have further advanced open-ended visual understanding and language generation (Bai et al., 2023; Liu et al., 2023; Dai et al., 2023). However, LVLMs remain prone to hallucination, which limits their deployment in applications that require high reliability (Li et al., 2023).

Prior studies attribute hallucination to two main sources: spurious co-occurrences and biases in visual instruction data, and the gradual decline in the model's reliance on visual inputs during autoregressive generation, which allows language priors to dominate the middle and later stages of decoding (Li et al., 2023; Zhou et al., 2024; Favero et al., 2024). Although existing mitigation methods operate through training, postprocessing, decoding, or internal attention intervention, they rarely model functional differences among attention heads jointly with changes in the model's visual state.

Motivated by this limitation, we analyze generation from a multihead attention perspective and find that visual evidence consistently concentrates in a small set of layer-specific Core Visual Attention Heads (CVAHs), whereas hallucinated object tokens receive weaker visual attention in early layers. Based on these observations, we propose Core Visual Attention Head Memory Injection (CAMI). At each autoregressive step, CAMI uses the decrease in the visual attention mass of CVAHs at a designated gating layer relative to its historical reference as an intervention-scheduling signal. This signal does not predict whether the current token is hallucinated; it controls whether visual memory is injected into CVAHs at the target layers. CAMI requires neither retraining nor an external detector and adapts to the current decoding state without changing the model architecture.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

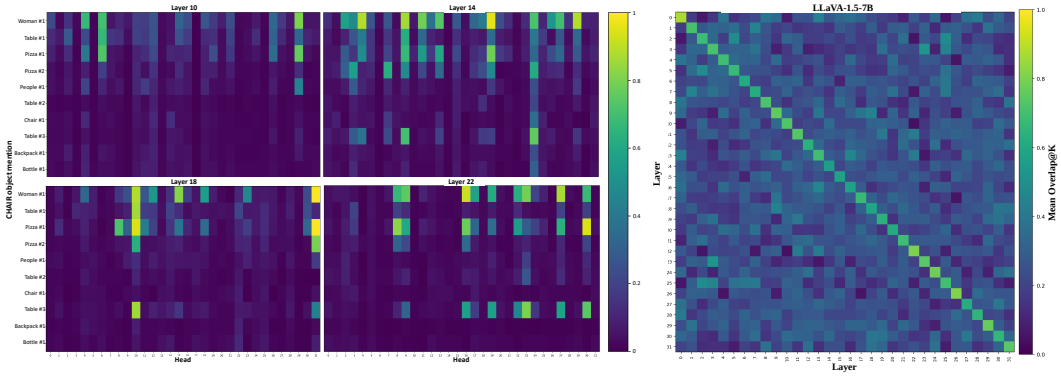

Figure 1: Heatmap of object mentions and attention heads (left), and mean cross-layer Overlap@8 on LLaVA-1.5-7B (right). Results for InstructBLIP-Vicuna-7B and Qwen2.5-VL-7B are provided in Appendix A.

Our contributions are threefold. First, we systematically analyze object generation in LVLMs from a multihead attention perspective and identify a sparse, stable, and layer-specific set of CVAHs through statistics collected across multiple samples.

Second, we reveal a difference in shallow-layer visual attention between faithful and hallucinated object mentions, with faithful mentions receiving stronger visual evidence.

Third, we propose the training-free CAMI framework, which dynamically intervenes in CVAHs through a visual-decay-based scheduler, reliable visual pattern memory, and text suppression in Auxiliary Visual Attention Heads (AVAHs).

2 Related Work

Existing approaches to hallucination mitigation in LVLMs broadly follow four directions: training, postprocessing, decoding, and attention intervention. Training methods reduce reliance on language priors by modifying model parameters (Yu et al., 2024; Xing et al., 2024; Yoon et al., 2025; Fu et al., 2025; Sun et al., 2024; Zhang et al., 2024; Yang et al., 2025). Although they can directly reshape model behavior, they require specialized data and additional training. Postprocessing methods use external tools to detect and correct hallucinated content after an initial response is generated (Zhou et al., 2024; Yin et al., 2024; Jeong et al., 2026; Lee et al., 2024; Sahu et al., 2024), introducing auxiliary models or iterative inference and thereby increasing latency. Decoding methods adjust candidate token probabilities during autoregressive generation through contrastive distributions, mutual information enhancement, retrospection penalties, or visual token sparsification (Favero et al., 2024; Leng et al., 2024; Huang et al., 2024; Zhuang et al., 2025; Dong et al., 2025; Wan et al., 2025; Huo et al., 2025; Chen et al., 2026). These methods typically leave model parameters unchanged but may require extra forward passes and can be sensitive to hyperparameters such as contrast strength and candidate truncation.

Attention based methods directly regulate how visual information is selected and propagated through the decoder. PAI, TAME, and CMAC apply corrections to overall visual dependence, intermediate layer visual enhancement, modality imbalance, and cross-modal positional bias (Liu et al., 2024; Yin et al., 2025; Sun et al., 2026; Li et al., 2026; Tang et al., 2025), but primarily operate on modalities, token groups, or layer level aggregate distributions. EAZY and MemVR recover visual evidence by removing image tokens that induce hallucination or reinjecting visual features (Zou et al., 2024; Che et al., 2025), without explicitly modeling functional differences among attention heads. ICT is the closest to our work; it selects attention heads that encode global or local visual information and intervenes in the model's internal representations (Chen et al., 2025).

As attention intervention becomes more granular, two problems remain insufficiently addressed. First, although existing methods that operate at the head level select particular attention heads, their head sets and intervention directions are typically determined by static rules or a limited calibration set. It

therefore remains unclear whether these heads consistently extract visual evidence across images and object generation positions. Second, most methods continuously enhance visual signals or use a fixed intervention strength, ignoring variations in visual information demands across decoding steps. Such intervention may induce object mentions without reliable visual support. We address these limitations by identifying stable core visual heads and introducing intervention that adapts to the current decoding state.

3 Understanding the Internal Mechanisms of Object Hallucination

This section first introduces the notation for autoregressive generation in LVLMs and defines visual attention mass. We then use case studies to examine multihead attention behavior and visual-state differences. The former identifies intervention targets, whereas the latter motivates a signal for scheduling intervention. These findings motivate the framework presented in Section 4.

3.1 Notation and Definition of Visual Attention Mass

Let the position sets of image and nonimage tokens in the fused input sequence be $\mathcal{V}$ and $\mathcal{T}$, respectively. At generation step t, let $\mathbf{z}_{\ell,h}^{t} \in \mathbb{R}^{N_t}$ denote the pre-softmax attention logits at the final query position for layer ℓ and attention head h. Applying a global softmax yields the complete attention distribution $\mathbf{a}_{\ell,h}^{t} = \mathrm{softmax}\Big(\mathbf{z}_{\ell,h}^{t}\Big)$. We define the total attention mass that this head assigns to all image tokens as

$$q_{\ell,h}^{t} = \sum_{j \in \mathcal{V}} a_{\ell,h,j}^{t}. \tag{1}$$

A larger value indicates that the head allocates more attention mass to image tokens at the current generation step. To characterize the relative allocation within the image token subset, we further define the visual attention pattern as the distribution normalized over image tokens:

$$\mathbf{p}_{\ell,h}^{t} = \mathrm{softmax}_{\mathcal{V}}\big(\mathbf{z}_{\ell,h,\mathcal{V}}^{t}\big)\,. \tag{2}$$

Here, $q_{\ell,h}^{t}$ measures how much the model attends to the image within the complete attention distribution, whereas $\mathbf{p}_{\ell,h}^{t}$ describes which image positions it attends to. We use these definitions throughout the remainder of the paper.

3.2 Differences in Multihead Attention Behavior

3.2.1 Analysis of Core Visual Attention Heads

During autoregressive generation in an LVLM, each self-attention head can theoretically read from the entire input sequence. This does not imply, however, that visual attention mass is distributed uniformly across heads. As our subsequent statistics show, when the model generates an object token, only a small subset of heads consistently assigns substantial attention mass to image tokens. The remaining heads allocate relatively little attention to the image and exhibit varying degrees of nonvisual preference.

To characterize these differences among heads, we analyze the object mentions identified by CHAIR. Let t_r be the generation step corresponding to CHAIR object mention r. For statistical analysis, we denote the visual attention mass of head h in layer ℓ at this object position by $q_{\ell,h}^{t_r}$. A larger value indicates stronger reliance on visual evidence when the head generates the current object mention.

Figure 1 shows clear vertical bands of high values within each decoder layer. This pattern indicates that visual information is not extracted uniformly by all heads; instead, a sparse subset repeatedly assumes this function. To reduce the influence of individual examples, we further collect statistics from 500 images and 3,095 aligned CHAIR object mentions. For object mention r, we rank the 32 heads in layer ℓ by $q_{\ell,h}^{t_r}$ and retain the Top-8 heads with the largest visual attention mass, denoted by $\mathcal{K}_{r,\ell}^{(8)}$.

We then compute the frequency $f_{\ell,h}$ with which each head appears in the Top-8 set. A head is classified as a Core Visual Attention Head (CVAH) if $f_{\ell,h} > 0.5$, as a Low Visual Attention Head (LVAH) if $f_{\ell,h} < 0.1$, and otherwise as an Auxiliary Visual Attention Head (AVAH). Uniformly selecting 8 of 32 heads yields an expected frequency of 0.25. The threshold of 0.5 therefore identifies visually dominant heads that recur across object mentions.

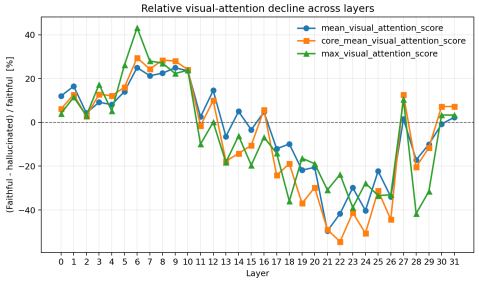

Figure 2: Layerwise visual attention differences between hallucinated and faithful object mentions.

On LLaVA-1.5-7B, every decoder layer contains a nonempty set of CVAHs, ranging from 5 to 9 heads with an average of 7.25. These results show that stable visual attention is concentrated in a small subset of heads. Moreover, CVAHs are layer specific, with limited correspondence across layers. Appendix A provides the complete distribution of CVAHs.

3.2.2 Stability Within and Across Layers

To assess the stability and layer specificity of CVAHs, we compute normalized Overlap@K, defined as the average size of the intersection between the Top-K head sets for two records divided by K. Within-layer overlap is averaged over pairs of records from the same layer, whereas cross-layer overlap is averaged over pairs from different layers.

On LLaVA-1.5-7B, the mean within-layer Overlap@8 is 0.690, compared with 0.250 for random selection and 0.245 across layers. The diagonal structure in Figure 1 indicates that CVAHs are stable within layers but vary across layers. Together with the results for the other two backbones in Appendix A, this pattern supports the stability and layer specificity of CVAHs across models.

3.3 Visual-Attention Differences for Intervention Scheduling

To characterize differences in visual dependence between faithful and hallucinated object mentions across decoder layers and to select an appropriate observation layer for gating, we compare their visual attention at each layer. Using visual attention mass $q^t_{\ell,h}$ defined in Section 3.1, we analyze its relationship to object hallucination. We consider three aggregation schemes that capture layerwise visual dependence from complementary perspectives: the mean visual attention over all heads, the mean over CVAHs, and the maximum visual attention of any single head in the layer:

$$
\begin{aligned}
s^{\text{mean}}_{\ell,t} &= \frac{1}{H}\sum_{h=1}^{H} q^t_{\ell,h}, \\
s^{\text{core}}_{\ell,t} &= \frac{1}{|\mathcal{C}_\ell|}\sum_{h\in\mathcal{C}_\ell} q^t_{\ell,h}, \\
s^{\max}_{\ell,t} &= \max_h q^t_{\ell,h}.
\end{aligned}
\tag{3}
$$

Here, $\mathcal{C}_\ell$ denotes the set of core attention heads with stable visual responses in layer ℓ. From descriptions generated for 500 images, we collect 3,593 CHAIR object mentions, comprising 3,095 faithful mentions and 498 hallucinated mentions. For each layer and aggregation scheme $\phi \in \{\text{mean}, \text{core}, \max\}$, let $\mathcal{F}$ and $\mathcal{H}$ denote the generation step sets associated with faithful and hallucinated mentions, respectively. We define the relative difference in visual attention between the two groups as

$$
R^\phi_\ell = \frac{\mathbb{E}_{t\in\mathcal{F}}\left[s^\phi_{\ell,t}\right] - \mathbb{E}_{t\in\mathcal{H}}\left[s^\phi_{\ell,t}\right]}{\mathbb{E}_{t\in\mathcal{F}}\left[s^\phi_{\ell,t}\right]}.
\tag{4}
$$

As shown in Figure 2, faithful object mentions receive greater visual attention in shallow layers, whereas the ordering reverses in several deeper layers. These offline population-level differences are used only to select the observation layer and signal direction; they do not define a token-level hallucination detector. At inference time, CAMI has no object or hallucination labels, and the

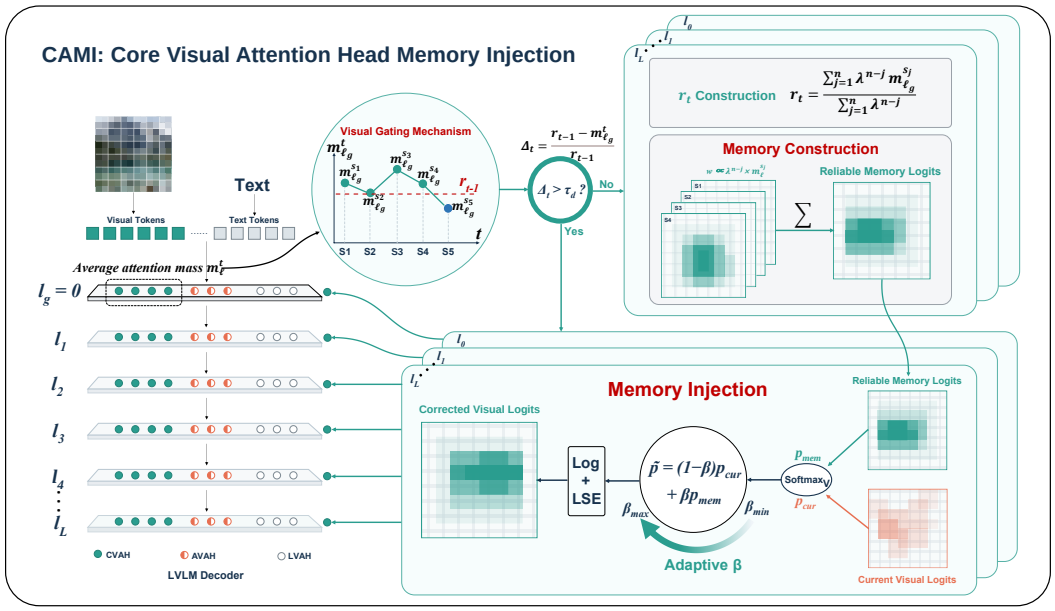

Figure 3: Overview of CAMI. A state-dependent gate uses relative visual-attention decay to schedule intervention; reference-consistent states form visual pattern memory, and adaptive injection restores visual evidence at triggered decoding steps.

relative decline in Section 4.1 acts only as a label-free control variable for scheduling intervention. Appendix B reports layerwise AUROC solely to support observation-layer selection.

4 CAMI

Based on the preceding observations, we propose Core Visual Attention Head Memory Injection (CAMI), which operates only during inference and requires neither model retraining nor an external detector. During autoregressive generation, CAMI dynamically monitors the attention mass assigned to image tokens by CVAHs at a designated gating layer. When this mass declines substantially relative to its historical reference, CAMI treats the current visual state as eligible for corrective intervention and adjusts the visual-token attention logits of CVAHs using historical visual pattern memory. CAMI consists of a visual decay gate, memory construction and injection, and text suppression in AVAHs.

4.1 Visual Decay Gating

We first define the mean visual attention mass of the CVAHs in layer ℓ and use the corresponding mean at a designated gating layer ℓ_g as the gating observation $m_{\ell_g}^t$:

$$m_\ell^t = \frac{1}{|\mathcal{C}_\ell|} \sum_{h \in \mathcal{C}_\ell} q_{\ell,h}^t. \tag{5}$$

To account for variation across samples and generation stages that would undermine a fixed reference, we maintain a visual attention reference r_{t-1} derived from reliable historical steps. The relative decline in visual attention at the current step, Δ_t, is defined as

$$\Delta_t = \frac{r_{t-1} - m_{\ell_g}^t}{r_{t-1}}. \tag{6}$$

The binary decision $g_t = \mathbb{I}[\Delta_t > \tau_d]$ schedules intervention rather than predicting hallucination. When $g_t = 1$, CAMI applies visual memory intervention. To evaluate step t, r_{t-1} is formed by exponentially weighting all non-triggered historical steps:

$$r_{t-1} = \frac{\sum_{j=1}^{n} \lambda^{n-j} m_{\ell_g}^{s_j}}{\sum_{j=1}^{n} \lambda^{n-j}}, \qquad s_1 < \cdots < s_n < t. \tag{7}$$

Here, $s_1, \ldots, s_n$ index the n non-triggered steps before t, and $\lambda \in [0, 1]$ controls exponential recency decay for both the gating reference and reliable visual memory. CAMI initializes the reference with the first observed visual attention mass without intervention. Thereafter, $m^t_{\ell_g}$ enters the history and updates r_t only when the gate is not triggered. Throughout, "reliable" denotes a non-triggered state consistent with the running reference, not verified factual correctness.

4.2 Reliable Visual Pattern Memory Construction

Reliable visual pattern memory stores the image token logits of CVAHs in the intervention layers when the gating layer exhibits no substantial visual decay. For a core visual head $h \in \mathcal{C}_\ell$ in intervention layer ℓ, CAMI writes the visual logits $\mathbf{u}^t_{\ell,h} := \mathbf{z}^t_{\ell,h,\mathcal{V}}$ to the history buffer only when step t does not activate the gate and satisfies the memory update interval I. It then constructs the aggregated memory $\mathbf{M}^t_{\ell,h}$ from all records collected under these update conditions:

$$\mathbf{M}^t_{\ell,h} = \frac{\sum_{j=1}^{n} \lambda^{n-j} m_\ell^{s_j} \mathbf{u}^{s_j}_{\ell,h}}{\sum_{j=1}^{n} \lambda^{n-j} m_\ell^{s_j}}, \qquad s_1 < \cdots < s_n \le t. \tag{8}$$

Here, $s_1, \ldots, s_n$ index all steps up to and including t that satisfy the memory update conditions, and n is the total number of collected records.

Memory construction uses the shared temporal decay coefficient λ defined in Section 4.1. The quantity $m_\ell^{s_j}$, defined above, is the mean visual attention mass of the CVAHs in layer ℓ at step s_j. If the current step does not satisfy the update conditions, the memory remains unchanged, $\mathbf{M}^t_{\ell,h} = \mathbf{M}^{t-1}_{\ell,h}$. The memory is updated only after computation at the current step, preventing triggered states from contaminating the historical pattern.

4.3 Reliable Visual Pattern Memory Injection

When the visual decay gate is triggered and reliable memory is available, CAMI injects it into the layer's CVAH image-token logits. To account for differences in offset and magnitude, CAMI normalizes current and memory logits over image tokens, blends their distributions in probability space, and restores the current visual log-partition when converting back to logits. This redistributes attention among image tokens while preserving the visual–text attention balance.

For a CVAH $h \in \mathcal{C}_\ell$ in layer ℓ, let $\mathbf{z}^t_{\ell,h,\mathcal{V}}$ denote the current visual logits and $\mathbf{M}^{t-1}_{\ell,h}$ the reliable visual memory. We normalize them over the visual-token subset as $\mathbf{p}^{\mathrm{cur},t}_{\ell,h} = \mathrm{softmax}_{\mathcal{V}}(\mathbf{z}^t_{\ell,h,\mathcal{V}})$ and $\mathbf{p}^{\mathrm{mem},t}_{\ell,h} = \mathrm{softmax}_{\mathcal{V}}(\mathbf{M}^{t-1}_{\ell,h})$, respectively. The fused distribution and the complete logits after memory injection but before text suppression are then computed as

$$\begin{aligned} \widetilde{\mathbf{p}}^t_{\ell,h} &= (1 - \beta^t_{\ell,h})\mathbf{p}^{\mathrm{cur},t}_{\ell,h} + \beta^t_{\ell,h}\mathbf{p}^{\mathrm{mem},t}_{\ell,h}, \\ c^t_{\ell,h} &= \log \sum_{k \in \mathcal{V}} \exp\big(z^t_{\ell,h,k}\big), \\ \widehat{z}^t_{\ell,h,j} &= \begin{cases} \log \widetilde{p}^t_{\ell,h,j} + c^t_{\ell,h}, & j \in \mathcal{V}, \\ z^t_{\ell,h,j}, & j \in \mathcal{T}. \end{cases} \end{aligned} \tag{9}$$

The LSE term restores the current visual log-partition, preserving the total visual attention mass and exactly recovering the original visual logits when $\beta^t_{\ell,h} = 0$. Here, $\beta^t_{\ell,h}$ dynamically controls the strength of reliable visual pattern memory injection. When visual decay is mild, $\beta^t_{\ell,h}$ approaches $\beta_{\min}$; as decay becomes more severe, the weight assigned to memory increases.

$$\beta^t_{\ell,h} = \beta_{\min} + (\beta_{\max} - \beta_{\min}) d^t_{\ell,h}. \tag{10}$$

$$d^t_{\ell,h} = \mathrm{clip}\left(\frac{\rho^{\mathrm{ref}}_{\ell,t-1} - q^t_{\ell,h}}{\rho^{\mathrm{ref}}_{\ell,t-1}}, 0, 1 \right). \tag{11}$$

$$\rho^{\mathrm{ref}}_{\ell,t-1} = \frac{\sum_{j=1}^{n} \lambda^{n-j} m_\ell^{s_j}}{\sum_{j=1}^{n} \lambda^{n-j}}, \qquad s_1 < \cdots < s_n < t. \tag{12}$$

Table 1: Comparison of different methods on MSCOCO-CHAIR. Lower CHAIR scores and higher recall are better.

Method	LLaVA-1.5-7B				Qwen2.5-VL-7B				InstructBLIP-Vicuna-7B			
	$C_S \downarrow$	$C_I \downarrow$	Recall ↑	Len	$C_S \downarrow$	$C_I \downarrow$	Recall ↑	Len	$C_S \downarrow$	$C_I \downarrow$	Recall ↑	Len
Base	46.8	12.6	76.1	90.7	28.2	7.8	58.9	145.4	46.2	12.7	73.6	101.4
VCD	51.4	16.1	76.3	92.9	29.6	8.6	59.2	150.5	52.4	14.9	75.5	98.9
OPERA	45.2	13.2	78.9	88.2	18.8	5.9	49.0	119.5	43.8	10.3	70.9	100.6
TAME	58.0	17.1	79.3	98.0	27.0	8.8	50.8	156.0	49.4	13.9	73.9	115.6
VTI	34.2	10.6	70.8	85.5	35.4	9.2	**62.4**	185.8	39.0	10.6	69.4	106.1
PAI	48.8	13.2	79.0	98.2	23.4	9.4	51.8	120.1	56.8	16.0	75.5	113.2
ONLY	57.6	17.8	75.6	87.7	19.4	6.6	52.1	113.7	55.4	17.4	71.5	99.9
MemVR	48.0	13.0	**80.0**	91.4	27.2	8.5	58.7	200.3	45.6	12.5	75.2	100.6
CMAC	36.0	9.2	76.3	95.1	27.2	8.5	57.1	144.6	59.4	16.5	**76.1**	101.2
EAZY	41.6	11.6	77.9	91.2	21.0	6.6	53.5	137.1	37.6	11.1	65.9	106.1
ICT	51.6	14.3	79.6	91.3	20.6	6.6	53.1	99.2	47.4	12.0	74.5	107.4
Ours	**29.6**	**8.4**	74.6	86.5	**18.2**	**5.6**	54.0	132.2	**36.4**	**9.7**	70.5	118.5

Table 2: Comparison of different methods on AMBER-G. Lower CHAIR, Hal, and Cog scores and higher Cover are better.

Method	LLaVA-1.5-7B				Qwen2.5-VL-7B				InstructBLIP-Vicuna-7B			
	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓
Base	6.7	50.2	28.6	4.3	4.0	56.1	21.4	1.6	7.1	54.0	34.6	3.7
VCD	7.9	50.6	36.0	5.4	4.7	54.0	21.4	2.0	8.1	54.2	40.2	3.8
OPERA	7.5	47.3	29.2	4.1	4.4	47.0	17.7	2.5	8.7	47.8	34.6	3.7
TAME	8.0	51.7	35.1	5.3	4.0	56.1	20.3	2.0	8.2	52.5	34.6	4.0
VTI	5.8	47.0	21.7	3.5	4.1	**62.2**	24.4	2.5	6.6	50.8	27.2	2.9
PAI	6.6	51.3	31.2	4.3	4.6	51.0	20.1	1.4	7.8	52.1	35.2	3.8
ONLY	10.1	47.5	38.5	4.9	4.1	56.1	23.6	2.1	10.7	49.6	43.4	5.2
MemVR	6.7	50.1	28.6	4.2	3.6	55.1	17.5	**1.2**	7.5	53.5	35.9	3.8
CMAC	6.3	**53.5**	32.1	4.0	3.5	53.6	17.1	1.3	8.8	**55.7**	44.6	6.7
EAZY	5.6	49.8	22.9	3.7	3.7	54.4	19.0	1.6	9.9	54.0	33.6	5.3
ICT	6.2	47.8	25.5	3.6	4.4	50.3	17.4	1.7	9.0	54.5	45.2	5.6
Ours	**5.5**	48.9	**20.8**	**3.2**	**3.1**	51.4	**14.5**	1.5	**6.3**	50.2	**26.9**	**2.8**

Here, $d^t_{\ell,h}$ quantifies visual decay, $q^t_{\ell,h}$ is the visual attention mass of the current head, and $\rho^{\text{ref}}_{\ell,t-1}$ is the layerwise historical reference maintained separately for each intervention layer. This reference and the gating reference r_{t-1} use the same weighted mechanism over reliable history, but differ in scope and purpose. The gating reference is constructed only for the gating layer to determine whether the current generation step requires intervention. The layerwise reference is constructed separately for each intervention layer and determines the adaptive memory injection weight of each CVAH after the gate is activated.

4.4 Text Suppression in Auxiliary Visual Attention Heads

To further reduce interference from language priors while limiting potential disruption to textual structure modeling, CAMI applies text suppression to AVAHs by default. Section 5.2 evaluates the effects of suppressing different types of attention heads through ablation studies.

Let $\mathcal{S}_\ell$ denote the set of heads to be suppressed in layer ℓ. For $h \in \mathcal{S}_\ell$, $\widehat{\mathbf{z}}^t_{\ell,h}$ denotes the complete attention logits before text suppression; if no memory injection is applied to this head, $\widehat{\mathbf{z}}^t_{\ell,h} = \mathbf{z}^t_{\ell,h}$.

Table 3: Summary of ablation results on MSCOCO-CHAIR with LLaVA-1.5-7B.

Ablation Dimension	Setting	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Recall ↑	Len
Base Model	No Intervention	46.8	12.6	76.1	90.7
Full Method	Visual Intervention on CVAHs ($r = 9.60\%$)	29.6	8.4	74.6	86.5
Head Selection for Gating and Memory Injection	Random Heads	45.8	12.5	77.9	90.9
	LVAHs	45.4	12.8	78.6	91.1
	AVAHs	46.2	12.7	78.1	89.3
	All Heads	43.8	16.2	61.2	83.7
Scheduling Policy	Always-On Intervention	44.0	17.1	62.4	82.9
	Without Memory Update Gate	32.8	9.4	73.7	86.5
	Random-Matched ($p = 0.0961$)	45.0	12.7	77.7	91.1
	Periodic-Matched ($P = 10$)	42.2	12.1	77.9	90.6
Memory Injection Mechanism	Fixed Visual Bias 0.3	47.0	12.5	79.3	92.2
	Fixed Visual Bias 0.5	47.6	12.6	79.9	92.7
Component Ablation	Memory Injection Only	33.2	9.5	73.6	83.5
	Text Suppression Only	45.6	12.3	78.7	91.3
Head Selection for Text Suppression	All Heads	38.0	10.2	74.8	86.9
	LVAHs	34.2	9.9	71.9	83.1
	Random Heads	38.2	11.0	73.6	84.7
	CVAHs	39.2	10.8	74.7	86.6

The nonvisual attention mass of this head is defined as $m_{\ell,h}^{\mathcal{T},t} = \sum_{j \in \mathcal{T}} \mathrm{softmax}(\widehat{\mathbf{z}}_{\ell,h}^t)_j$. When $m_{\ell,h}^{\mathcal{T},t}$ exceeds the target $\tau_{\mathcal{T}}$, we apply a negative correction to the logits at nonvisual positions:

$$\widetilde{z}_{\ell,h,j}^t = \begin{cases} \widehat{z}_{\ell,h,j}^t, & j \in \mathcal{V}, \\ \widehat{z}_{\ell,h,j}^t - \pi_{\ell,h}^t, & j \in \mathcal{T}, \end{cases} \qquad h \in \mathcal{S}_\ell. \tag{13}$$

Here, $\pi_{\ell,h}^t$ is an adaptive penalty that increases with the amount by which the nonvisual attention mass exceeds the target; when the nonvisual attention mass does not exceed the target, $\pi_{\ell,h}^t = 0$.

$$\pi_{\ell,h}^t = \mathrm{clip}\left(\log \frac{m_{\ell,h}^{\mathcal{T},t}(1 - \tau_{\mathcal{T}})}{\tau_{\mathcal{T}}(1 - m_{\ell,h}^{\mathcal{T},t})}, 0, \pi_{\max} \right). \tag{14}$$

5 Experiments

5.1 Comparisons with Existing Methods

Experimental Setup. We evaluate CAMI on LLaVA-1.5-7B, Qwen2.5-VL-7B, and InstructBLIP-Vicuna-7B, and compare it with VCD, OPERA, TAME, VTI, PAI, ONLY, MemVR, CMAC, EAZY, and ICT. All compared methods use the same task prompt and generate at most 256 new tokens. Baselines use their official default settings and hyperparameters. For each backbone, we use 500 images sampled from MSCOCO with random seed 142 exclusively as a calibration set for CVAH identification and hyperparameter selection. For CHAIR evaluation, we independently sample another 500 images with random seed 42. After selecting the core heads and setting the hyperparameters, we fix both and transfer them directly to AMBER-G without reselecting heads or adjusting parameters.

MSCOCO-CHAIR Results. As shown in Table 1, CAMI achieves the lowest CHAIRs and CHAIRi on all three models, demonstrating consistent suppression of object hallucination. Relative to the base models, CAMI reduces CHAIRs by 36.8%, 35.5%, and 21.2%, and CHAIRi by 33.3%, 28.2%, and 23.6%, respectively. CAMI also outperforms the strongest baselines on all three models. Recall and output length remain broadly comparable to those of the baselines and the unmodified models. Overall, CAMI consistently reduces hallucination across model architectures at a controlled cost in coverage.

AMBER-G Results. Table 2 reports the results of transferring CAMI to AMBER-G. We directly use the attention head sets and all hyperparameters determined on MSCOCO, without reselecting heads or tuning parameters on AMBER-G. CAMI achieves the lowest CHAIR and Hal scores across all three backbones. It also improves Cog on every model, indicating that CAMI not only reduces hallucinated objects but also decreases cognitive errors in generated descriptions. Overall, CAMI consistently improves multiple hallucination metrics without adaptation to the target dataset, demonstrating strong transferability across benchmarks.

5.2 Ablation Studies

We conduct ablation studies on MSCOCO-CHAIR using LLaVA-1.5-7B. The full method uses CVAHs in layer 0 as observations for the visual decay gate, constructs and injects reliable visual pattern memory in CVAHs across all decoder layers, and suppresses textual attention in AVAHs. The main hyperparameters remain fixed: the gating threshold τ_d is 0.04; the gating reference and memory share a temporal decay coefficient λ of 0.8; the memory update interval is 7; the adaptive injection weight ranges from 0.2 to 0.8; the nonvisual attention target $\tau_{\mathcal{T}}$ is 0.65; and the maximum text suppression strength $\pi_{\max}$ is 1.2. Because the gate is an intervention scheduler rather than a hallucination classifier, we evaluate whether it allocates a fixed intervention budget better than state-agnostic policies. CAMI triggers on 9.60% of steps; Random-Matched activates with probability 0.0961, and Periodic-Matched intervenes every 10 steps with a sample-specific random offset. All other components remain unchanged.

5.2.1 Analysis

First, head selection is critical: applying both gating and memory intervention to Random Heads, AVAHs, LVAHs, or All Heads raises CHAIRs to at least 43.8, confirming that gains stem from CVAHs' functional specificity rather than merely the number of intervened heads.

Second, the gating policy is important for intervention scheduling. Random-Matched and Periodic-Matched obtain CHAIR_s scores of 45.0 and 42.2 and CHAIR_i scores of 12.7 and 12.1, respectively, compared with 29.6 and 8.4 for CAMI. Although the two controls retain higher recall, they provide only modest hallucination reduction at comparable intervention frequencies. This matched-rate comparison directly validates the intervention utility of state-dependent scheduling; token-level detection accuracy is not its objective. Injecting at every step or removing the memory update gate increases CHAIR_s to 44.0 and 32.8, respectively, further confirming the importance of selective intervention and controlled memory updates.

Third, a fixed visual bias increases the model's visual attention mass by uniformly shifting the visual logits while preserving the relative distribution within the image token subset. The results show that globally increasing visual attention cannot replace selective restoration of reliable visual patterns.

Fourth, the component ablations show that memory injection provides the primary benefit, while text suppression in AVAHs yields complementary improvements. Applying text suppression to other types of attention heads consistently degrades performance.

6 Conclusion

We investigate object hallucination in LVLMs from a multihead attention perspective and find that visual evidence concentrates in a small set of stable, layer-specific CVAHs, while hallucinated object mentions exhibit insufficient visual attention in early layers. Based on these findings, we introduce CAMI, a training-free method that uses the visual attention decay of CVAHs relative to a historical reference to schedule intervention rather than classify hallucinated tokens. When triggered, CAMI injects historical visual pattern memory and suppresses excessive textual attention in AVAHs. Experiments on LLaVA-1.5-7B, Qwen2.5-VL-7B, and InstructBLIP-Vicuna-7B show that CAMI consistently reduces CHAIRs and CHAIRi and generalizes effectively to AMBER-G. Ablation studies further validate the importance of CVAH selection, dynamic gating, and reliable visual pattern memory injection.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21635–21644. IEEE, 2025.

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

Wei Chen, Xin Yan, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, and Long Chen. Decoupling contrastive decoding: Robust hallucination mitigation in multimodal large language models. *Advances in Neural Information Processing Systems*, 38:166734–166756, 2026.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2534–2544. IEEE, 2025.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312. IEEE, 2024.

Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong 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.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 24272–24295, 2025.

Sungheon Jeong, Ryozo Masukawa, Jihong Park, Sanggeon Yun, Wenjun Huang, Hanning Chen, Mahdi Imani, and Mohsen Imani. Draft and refine with visual experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18816–18826, 2026.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 391–404, 2024.

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

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, Ming Li, Xiaonan Luo, and Guanbin Li. Cross-modal attention calibration for lvlm hallucination mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40186–40196, 2026.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shi Liu, Kecheng Zheng, and Wei 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.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. Pelican: Correcting hallucination in vision-llms via claim decomposition and program of thought verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8228–8248, 2024.

Han Sun, Qin Li, Peixin Wang, and Min Zhang. Mitigating object hallucinations in lvlms via attention imbalance rectification. *arXiv preprint arXiv:2603.24058*, 2026.

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

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

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3225–3234. IEEE, 2025.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1167–1181, 2024.

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

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14625–14634, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Dokyoon Yoon, Youngsook Song, and Woomyong Park. Stop learning it all to mitigate visual hallucination, focus on the hallucination target. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4200–4208. IEEE, 2025.

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

Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Haocheng Feng, Jingdong Wang, et al. Automated multi-level preference for mllms. *Advances in Neural Information Processing Systems*, 37:26171–26194, 2024.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024.

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

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

A Cross-Architecture Analysis of CVAHs

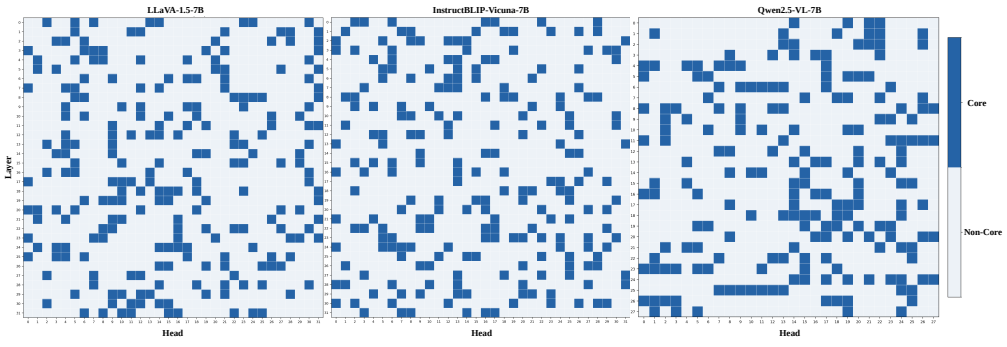

Figure A1: Distributions of Core Visual Attention Heads (CVAHs) in LLaVA-1.5-7B (left), InstructBLIP-Vicuna-7B (middle), and Qwen2.5-VL-7B (right).

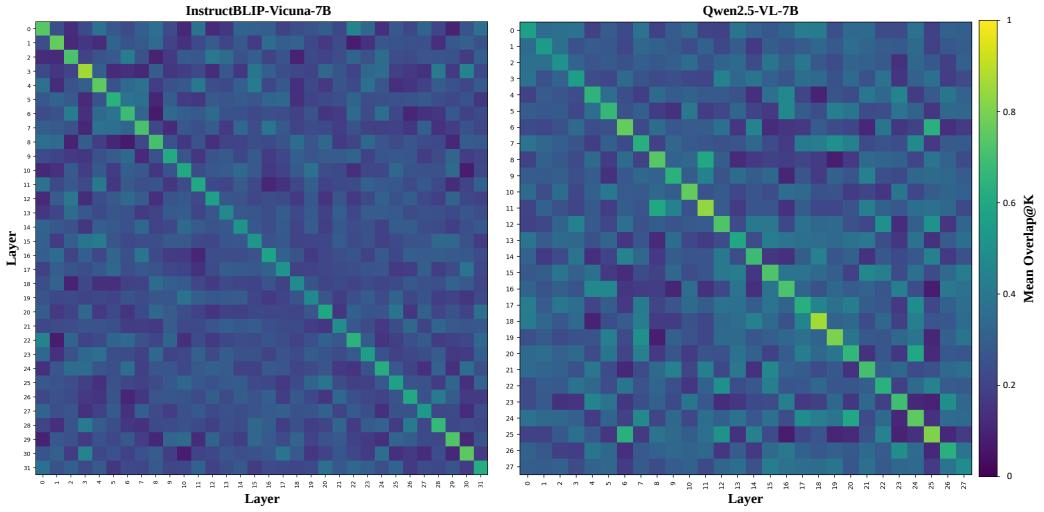

Figure A2: Mean cross-layer Overlap@8 for InstructBLIP-Vicuna-7B (left) and Qwen2.5-VL-7B (right). Diagonal entries measure within-layer stability, whereas off-diagonal entries measure cross-layer overlap.

Figure A1 presents the complete layerwise distributions of Core Visual Attention Heads (CVAHs) in the three backbones. CVAHs remain sparse: LLaVA-1.5-7B contains 5–9 CVAHs per layer (7.25 on average), while InstructBLIP-Vicuna-7B and Qwen2.5-VL-7B each contain 4–10 (7.00 and 7.14 on average, respectively). Across different architectures and attention-head configurations, stable visual attention thus concentrates in a small subset of heads.

Figure A2 reports cross-layer Overlap@8 matrices to assess within-layer stability and cross-layer specificity. Mean within-layer Overlap@8 is 0.677 for Qwen2.5-VL-7B and 0.622 for InstructBLIP-Vicuna-7B, versus mean cross-layer values of 0.285 and 0.239. The stronger diagonal structure shows that high-visual-attention head identities remain stable within layers but vary across layers, supporting the cross-architecture generality of layer-specific CVAHs.

B Rationale for Gating Layer Selection and Relative Visual-Decay Gating

B.1 Layerwise Visual Attention and Discriminative Power

Section 3.3 of the main paper uses the relative visual attention difference R_ℓ^ϕ to compare the visual attention received by faithful and hallucinated object mentions across decoder layers. Here, $\mathcal{F}$ and $\mathcal{H}$ denote the sets of generation steps associated with faithful and hallucinated object mentions,

respectively, and $\phi \in \{\text{mean}, \text{core}, \max\}$ indexes three aggregation schemes: the mean over all attention heads, the mean over CVAHs, and the maximum over individual heads. This statistic captures the direction of the average difference between the two classes at each layer. However, the difference in means alone does not reflect variance within each class or the degree of overlap between the two distributions.

To further evaluate the ability of individual decoder layers to distinguish faithful from hallucinated object mentions, we conduct a layerwise analysis of discriminative power using the area under the receiver operating characteristic curve (AUROC). For layer ℓ and aggregation scheme ϕ, we define

$$\begin{aligned} \mathrm{AUROC}_\ell^\phi = \Pr\left(s_{\ell,t_F}^\phi > s_{\ell,t_H}^\phi\right) \\ + \frac{1}{2}\Pr\left(s_{\ell,t_F}^\phi = s_{\ell,t_H}^\phi\right), \end{aligned} \tag{B1}$$

where $t_F \sim \mathcal{F}$ and $t_H \sim \mathcal{H}$ are samples drawn at random from the generation-step sets for faithful and hallucinated object mentions, respectively. When $\mathrm{AUROC}_\ell^\phi > 0.5$, faithful object mentions typically receive greater visual attention. Values below 0.5 indicate a reversal in the relative ordering of the two score distributions. If hallucination is treated as the positive class, $-s_{\ell,t}^\phi$ can instead be used as a hallucination-risk score, which yields the same AUROC value. Because AUROC depends only on sample rankings, it is independent of any particular decision threshold and invariant to differences in score scales across layers. It is therefore more appropriate than the absolute difference in means for selecting the gating layer.

Using 3,095 faithful object mentions and 498 hallucinated object mentions from the mechanism-analysis set, we compute the AUROC for the mean, core, and max visual attention scores at each of the 32 decoder layers. AUROC is computed using the average-rank formulation of the Mann–Whitney U statistic, with ties assigned a value of 0.5.

B.2 Layerwise AUROC and Layer 0 Selection

Figure B1 presents the results for all 32 layers. Layer 0 achieves the highest AUROC across all layers under both the core and max schemes, with values of 0.737 and 0.738, respectively. Layer 1 achieves the highest mean AUROC (0.757), but the corresponding core and max AUROCs are 0.732 and 0.712, respectively. Because CAMI uses the mean visual attention mass of the CVAHs at the gating layer as its gating observation, the core score is the discriminative metric that directly corresponds to the gate definition. Layer 0 attains the highest AUROC across all layers under this metric and is therefore selected as the gating layer, $\ell_g = 0$.

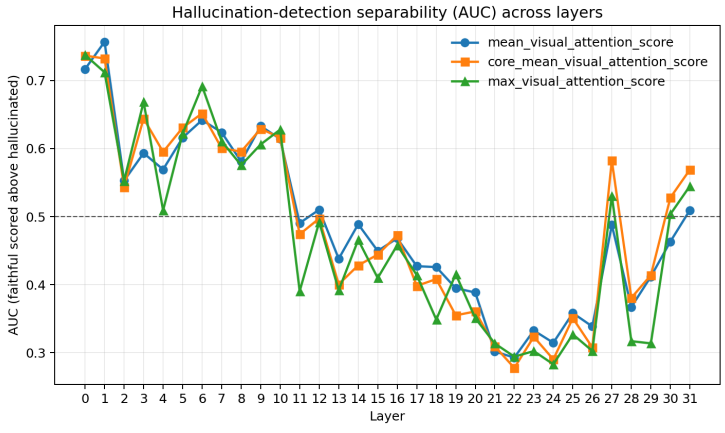

Figure B1: AUROC of the three visual attention aggregation scores across decoder layers. The dashed line indicates the chance level of 0.5; values below 0.5 indicate a reversal in score ordering between faithful and hallucinated object mentions.

At layer 0, Cohen's d is 0.84 for both the core and max schemes and 0.77 for the mean scheme, further demonstrating a substantial difference between the distributions of faithful and hallucinated object mentions in shallow layers. AUROC generally declines with depth and falls below 0.5 in several middle and later layers. This result is consistent with the R_ℓ^ϕ trend reported in Section 3.3 of the main paper: faithful object mentions receive stronger visual evidence in shallow layers, whereas the score ordering reverses in several middle and later layers. Although the reversed scores in middle and later layers still contain class information, their direction no longer satisfies the unified gating assumption that insufficient visual attention implies greater risk. Moreover, gating decisions at these depths occur too late to support timely intervention in subsequent layers.

The preceding R_ℓ^ϕ and AUROC analyses measure aggregate class differences between labeled object mentions, whereas CAMI has no access during inference to labels indicating whether an object mention is faithful or hallucinated. Therefore, Section 4.1 of the main paper does not use an absolute visual attention score directly as an online risk indicator. Instead, it compares the current mean visual attention mass of the CVAHs at the gating layer, $m_{\ell_g}^t$, with the reliable historical reference r_{t-1} from the same sample:

$$\begin{aligned} \Delta_t &= \frac{r_{t-1} - m_{\ell_g}^t}{r_{t-1}} = 1 - \frac{m_{\ell_g}^t}{r_{t-1}}, \\ \frac{\partial \Delta_t}{\partial m_{\ell_g}^t} &= -\frac{1}{r_{t-1}} < 0. \end{aligned} \tag{B2}$$

If $r_{t-1} > 0$, Equation (B2) shows that Δ_t is strictly decreasing in $m_{\ell_g}^t$; hence, lower visual attention at the gating layer produces a larger decline relative to the historical reference. Accordingly, R_ℓ^ϕ and the layerwise AUROC analysis jointly motivate the selection of the observation layer and the direction of the risk signal. They do not directly measure the discriminative performance of the final relative-decline score Δ_t, whose sample-specific historical reference varies across decoding steps. Instead, Δ_t operationalizes this offline observation as a label-free, within-sample gating criterion and is used to gate intervention rather than to predict hallucination.

AI USE STATEMENT

We used generative AI tools for language editing, code development, debugging, brainstorming, and feedback on experimental results. Generative AI was not used to draft the substantive content of the paper or to design the experiments.

All AI-assisted outputs were reviewed and verified by the authors. AI-assisted code was manually inspected and tested, AI-assisted edits and feedback were independently evaluated for technical accuracy and consistency with the experimental evidence. The authors take full responsibility for the final content of this work.