

# Vocabulary Hijacking in LVLMs: Unveiling Critical Attention Heads by Excluding Inert Tokens to Mitigate Hallucination

Yangneng Chen<sup>1</sup> Junlin Li<sup>1</sup> Weijun Yao<sup>2</sup> Xilai Ma<sup>1</sup>  
Guodong Du<sup>3</sup> Wenya Wang<sup>4</sup> Jing Li<sup>1</sup>✉

<sup>1</sup>Harbin Institute of Technology (Shenzhen), China <sup>2</sup>Huawei Technologies Co., Ltd.  
<sup>3</sup>The Hong Kong Polytechnic University <sup>4</sup>Nanyang Technological University  
yangnengchen8@gmail.com jingli.phd@hotmail.com

## Abstract

Large Vision-Language Models (LVLMs) have achieved remarkable progress in multimodal tasks, yet their reliability is persistently undermined by hallucinations—generating text that contradicts visual input. Recent studies often attribute these errors to inadequate visual attention. In this work, we analyze the attention mechanisms via the logit lens, uncovering a distinct anomaly we term **Vocabulary Hijacking**. We discover that specific visual tokens, defined as **Inert Tokens**, disproportionately attract attention. Crucially, when their intermediate hidden states are projected into the vocabulary space, they consistently decode to a fixed set of unrelated words (termed **Hijacking Anchors**) across layers, revealing a rigid semantic collapse. Leveraging this semantic rigidity, we propose **Hijacking Anchor-Based Identification (HABI)**, a robust strategy to accurately localize these Inert Tokens. To quantify the impact of this phenomenon, we introduce the **Non-Hijacked Visual Attention Ratio (NHAR)**, a novel metric designed to identify attention heads that remain resilient to hijacking and are critical for factual accuracy. Building on these insights, we propose **Hijacking-Aware Visual Attention Enhancement (HAVA)**, a training-free intervention that selectively strengthens the focus of these identified heads on salient visual content. Extensive experiments across multiple benchmarks demonstrate that HAVA significantly mitigates hallucinations with **no additional computational overhead**, while preserving the model’s general capabilities. Our code is publicly available at <https://github.com/lab-klc/HAVA>.

## 1 Introduction

Large Vision-Language Models (LVLMs), extending large language models (LLMs) to process visual inputs, represent a major advancement in artificial intelligence (Liu et al., 2023; Chen et al.,

(a) Average Head Selection (b) HAVA Head Selection

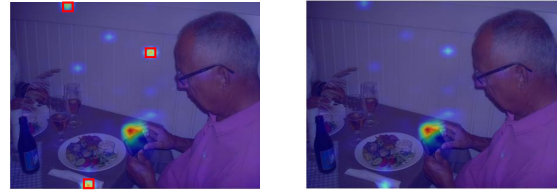


Figure 1: Attention maps for generating “phone”. (a) The average of all heads exhibits the Vocabulary Hijacking phenomenon, with focus scattered on the background. (b) In contrast, heads selected by HAVA concentrate attention precisely on the target object.

2023; Zhu et al., 2023; Wang et al., 2024). Nevertheless, LVLMs remain susceptible to hallucinations (Sun et al., 2024; Zhou et al., 2024; Huang et al., 2024; Bai et al., 2024), where generated text fails to align with visual content. Such inconsistencies undermine LVM reliability in multimodal tasks, limiting their practical deployment.

Substantial research on mitigating hallucinations in LVLMs has produced various intervention strategies (Yu et al., 2024a; Leng et al., 2024; Chen et al., 2024; Li et al., 2025). **A critical consensus has emerged:** insufficient attention to visual tokens during generation is a primary driver of hallucinations. This insight has inspired numerous attention-based methods (Jiang et al., 2025; Liu et al., 2024; Yang et al., 2025). However, a fundamental question remains unresolved: *Which specific attention heads are actually critical?*

To demystify attention mechanisms, (Jiang et al., 2025) applied Logit Lens to vision tokens, identifying a transition from *Visual Information Enrichment* to *Semantic Refinement*. Adopting this framework, we define the **Trace** as the layer-wise sequence of vocabulary words decoded from a token’s hidden states. Typically, a functional Trace evolves from meaninglessness to rich semantics; for instance, Token #395 in Figure 2 shifts from nonsensical outputs to image-aligned concepts like “phone” after layer 20.

✉ Corresponding author.



head, while preserving general capabilities.

## 2 Related Work

**Reducing Hallucinations in MLLMs.** Mitigating hallucinations in MLLMs is primarily addressed through fine-tuning (Yu et al., 2024a,b) or training-free inference interventions. Prominent strategies include contrastive decoding (Leng et al., 2024; Chen et al., 2024), activation steering (Li et al., 2025), and attention intervention (Liu et al., 2024; Jiang et al., 2025; Yang et al., 2025). While attention-based methods are most relevant to our work, they face a critical unresolved problem: **determining precisely which heads require intervention**. In the absence of a principled identification mechanism, these approaches are often forced to rely on heuristics. Our work bridges this gap by proposing a precise metric to target the specific heads essential for factual grounding.

**Attention Sink in Language Models.** The ‘‘Attention Sink’’ phenomenon, where vacuous tokens monopolize attention, is established in LLMs (Xiao et al., 2023) and recently extended to LVLMs as Visual Attention Sinks (VAS) (Kang et al., 2025). Along similar lines, works like AVISC (Woo et al., 2025) have also investigated comparable attention anomalies to mitigate hallucinations. However, the underlying mechanisms of these phenomena remain under-explored. We argue that our proposed *Vocabulary Hijacking* offers a more precise characterization of anomalous attention patterns in LVLMs compared to the VAS framework.

## 3 Characterizing and Identifying Vocabulary Hijacking

### 3.1 Preliminaries

**Autoregressive Generation in LVLMs.** Large Vision-Language Models (LVLMs) generate responses autoregressively by modeling the conditional probability of the next token. At each timestep  $k$ , the model predicts token  $y_k$  based on the preceding context, which comprises a sequence of image tokens  $\mathcal{I}_v$ , a text prompt  $\mathcal{I}_t$ , and previously generated tokens  $\mathcal{I}_o$ . These components are concatenated to form a single input sequence  $\mathcal{I}$ .

**Attention Mechanism in LVLMs.** The core component enabling token interaction is Multi-Head Attention (MHA). Following (Elhage et al., 2021), at layer  $\ell$ , the representation for a token  $\mathbf{x}_i^{\ell-1}$  is updated by attending to all previous tokens

$\mathbf{X}_{\leq i}^{\ell-1} = \{\mathbf{x}_0^{\ell-1}, \dots, \mathbf{x}_i^{\ell-1}\}$  as follows:

$$\begin{aligned} \text{MHA}^{\ell,h}(\mathbf{x}_i^{\ell-1}) &= \sum_{j \leq i} \mathbf{A}_{i,j}^{\ell,h} \mathbf{x}_j^{\ell-1} \mathbf{W}_{OV}^{\ell,h}, \\ \mathbf{A}_i^{\ell,h} &= \text{softmax} \left( \frac{1}{\sqrt{D_k}} (\mathbf{x}_i^{\ell-1} \mathbf{W}_Q^{\ell,h}) \right. \\ &\quad \left. \cdot (\mathbf{X}_{\leq i}^{\ell-1} \mathbf{W}_K^{\ell,h})^\top \right). \end{aligned} \quad (1)$$

Here,  $\mathbf{W}_Q^{\ell,h}, \mathbf{W}_K^{\ell,h} \in \mathbb{R}^{D \times D_k}$  are the query and key projection matrices, and  $\mathbf{W}_{OV}^{\ell,h} \in \mathbb{R}^{D \times D}$  is the output-value projection matrix. The attention weight  $\mathbf{A}_{i,j}^{\ell,h}$  quantifies the contribution of token  $\mathbf{x}_j^{\ell-1}$  to the updated representation of token  $\mathbf{x}_i^{\ell-1}$ . Our analysis focuses on the cases where  $i \in \mathcal{I}_o$  and  $j \in \mathcal{I}_v$ , which correspond to the attention paid to visual tokens during the generation process.

**Logit Lens.** To investigate the model’s internal processing of visual information, we employ Logit Lens (nostalgebraist, 2020; Jiang et al., 2025). It projects an intermediate visual hidden state  $\mathbf{v}_i^\ell$  directly onto the vocabulary distribution  $\Sigma$  by applying the model’s final unembedding matrix,  $\mathbf{W}_\Sigma \in \mathbb{R}^{|\Sigma| \times d}$ :

$$\mathbf{p}(\Sigma | \mathbf{v}_i^\ell) = \text{softmax}(\mathbf{W}_\Sigma \cdot \mathbf{v}_i^\ell) \in \mathbb{R}^{|\Sigma|}, \quad (2)$$

where  $p_j(\Sigma | \mathbf{v}_i^\ell)$  denotes the probability of the  $j$ -th vocabulary token. We identify the token with the highest probability from this distribution as the semantic interpretation of the hidden state  $\mathbf{v}_i^\ell$ .

### 3.2 Setup for Exploratory Analysis

Following (Jiang et al., 2025), we conduct our analysis on 500 random images from the COCO 2014 validation set (Lin et al., 2014), utilizing LLaVA-1.5 (7B/13B), Shikra-7B, MiniGPT4-7B, and Qwen2-VL-7B. Descriptions are generated via **greedy decoding** with the prompt: ‘‘Please help me describe the image in detail’’. Let  $\mathcal{V}$  denote the collective set of vision tokens from these images. We categorize generated objects into real ( $\mathcal{O}_{\text{real}}$ ) and hallucinated ( $\mathcal{O}_{\text{hal}}$ ) sets based on ground-truth annotations of COCO.

### 3.3 Understanding Vocabulary Hijacking via Logit Lens

**Trace Analysis & Hijacking Score.** For a vision token  $v_i \in \mathcal{V}$ , we define its **Trace**,  $\mathcal{T}_{v_i}$ , as the sequence of words decoded from its layer-wise

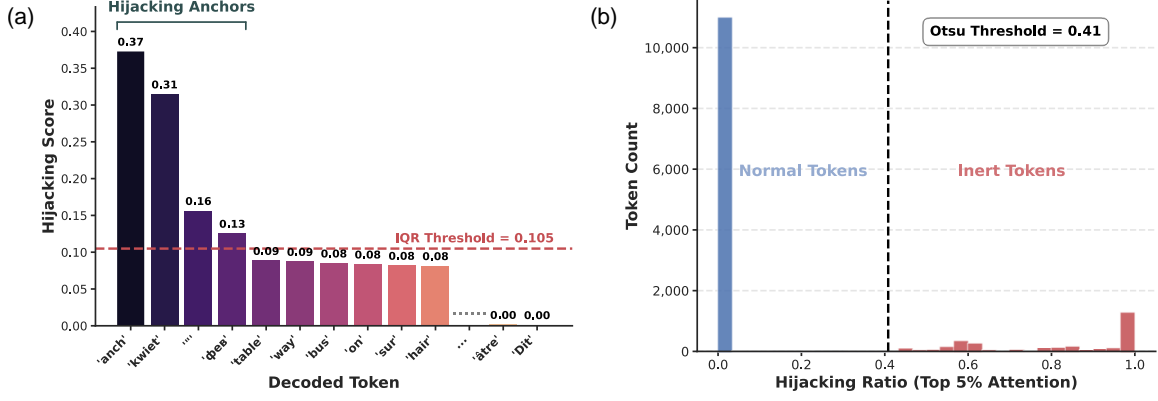


Figure 3: Empirical basis for the HABI method on LLaVA-1.5 7B. **(a)** The long-tailed distribution of Mean Hijacking Scores, illustrating the identification of **Hijacking Anchors** via an IQR-based outlier threshold. **(b)** The distinct bimodal distribution of the Hijacking Ratio for salient visual tokens (top 5% attention mass), demonstrating the automatic isolation of **Inert Tokens** using Otsu’s method.

hidden states via Logit Lens. We identify **Inert Tokens** as specific high-attention tokens that are rigidly dominated by fixed, meaningless words. To identify them, we designate the most frequent token in  $\mathcal{T}_{v_i}$  as the **Anchor** ( $a_{v_i}$ ) and propose a composite **Hijacking Score** based on three dimensions: (1) **Dominance** ( $\mathcal{D}$ ): the proportion of the Anchor  $a_{v_i}$  within the Trace  $\mathcal{T}_{v_i}$ ; (2) **Frequency** ( $\mathcal{F}$ ): the log-smoothed frequency of  $a_{v_i}$  among all Anchors in  $\mathcal{V}$ ; and (3) **Attention** ( $\mathcal{A}$ ): the log-smoothed average attention received by  $v_i$  during the first 10 decoding steps. The score is formulated as:

$$S_{\text{hijack}}(v_i) = \mathcal{D}(v_i) \cdot \mathcal{F}(v_i) \cdot \mathcal{A}(v_i). \quad (3)$$

**Defining Hijacking Anchors.** We shift focus to the vocabulary level  $\Sigma$  to identify systemic distractors. For a unique token  $w \in \Sigma$ , we compute its **Mean Hijacking Score**,  $\bar{S}(w)$ , by averaging  $S_{\text{hijack}}$  across all vision tokens anchored to it:

$$\bar{S}(w) = \frac{1}{|\{v \in \mathcal{V} \mid a_v = w\}|} \sum_{v \in \mathcal{V}, a_v = w} S_{\text{hijack}}(v). \quad (4)$$

The distribution of  $\bar{S}(w)$  exhibits a long-tailed pattern (Figure 3 (a)). Using an outlier threshold  $\tau_s = Q_3 + 1.5 \cdot \text{IQR}$ , we formally define the set of **Hijacking Anchors**,  $\mathcal{A}_{\text{hijack}}$ , as:

$$\mathcal{A}_{\text{hijack}} = \{w \in \Sigma \mid \bar{S}(w) > \tau_s\}. \quad (5)$$

This set characterizes a systemic model property, indicating that vision tokens whose Traces contain these Anchors are significantly more prone to hijacking the model’s attention. To further corroborate the validity of the Hijacking Score and the rationality of the IQR thresholding, we present a supplementary clustering analysis in Appendix G.1.

### 3.4 Hijacking Anchor-Based Identification

Leveraging the identified set  $\mathcal{A}_{\text{hijack}}$ , we propose *Hijacking Anchor-Based Identification* (HABI) to detect Inert Tokens. We first define the **Hijacking Ratio** ( $r_{\text{hijack}}$ ) to quantify the extent to which a token’s Trace is dominated by Hijacking Anchors:

$$r_{\text{hijack}}(v_i) = \frac{1}{|\mathcal{T}_{v_i}|} \sum_{t \in \mathcal{T}_{v_i}} \mathbb{I}(t \in \mathcal{A}_{\text{hijack}}). \quad (6)$$

To establish a robust decision boundary, we analyze the distribution of  $r_{\text{hijack}}$ . While the global distribution is long-tailed, restricting analysis to salient tokens (top 5% attention mass, given the sparsity of visual attention) reveals a distinct **bimodal structure** (Figure 3 (b)). This bimodality enables the use of **Otsu’s method** (Otsu, 1979) to automatically derive an optimal separation threshold  $\tau_r$ . Accordingly, we formally define the set of **Inert Tokens**,  $\mathcal{I}_{\text{inert}}$ , as the subset of vision tokens exceeding this threshold:

$$\mathcal{I}_{\text{inert}} = \{v_i \in \mathcal{I}_v \mid r_{\text{hijack}}(v_i) > \tau_r\}. \quad (7)$$

This definition explicitly isolates the specific subset of tokens responsible for hijacking the model’s attention mechanism. The specific algorithmic procedure is detailed in Algorithm 1. Distributions of Hijacking Scores and Hijacking Ratios for additional models are presented in Appendix B.1. Furthermore, we investigate the spatial distribution of Inert Tokens in Appendix B.3 and provide an attention analysis in Appendix B.4.

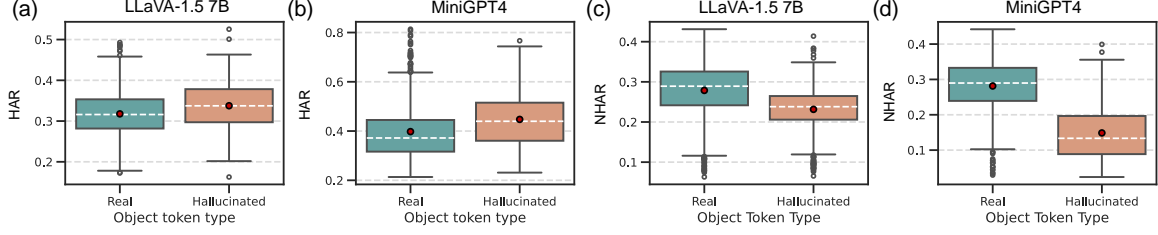


Figure 4: Statistical distributions of attention metrics. (a, b) present the distributions of the Hijacked Attention Ratio (HAR) for LLaVA-1.5 7B and MiniGPT-4, respectively. (c, d) illustrate the Non-Hijacked Visual Attention Ratio (NHAR) distributions for the same models.

## 4 From Vocabulary Hijacking to Hallucination

### 4.1 Relationship with Visual Attention Sink

As discussed in Section 2, the *Visual Attention Sink* (VAS) (Kang et al., 2025) is characterized by three key features: persistent attention attraction, background localization, and negligible contribution. While Vocabulary Hijacking shares these fundamental properties, we demonstrate in Appendix C that our framework offers a significantly more precise mechanism for identifying and characterizing these anomalous attention patterns.

### 4.2 Linking Vocabulary Hijacking to Hallucination

**Diagnostic Analysis using HAR.** Acknowledging the functional specialization of attention heads (Deiseroth et al., 2023; Zhang et al., 2024; Ge et al., 2024; Zheng et al., 2024), we first focus our analysis on the top-300 heads with the **highest total visual attention**. We identify this subset as the primary conduits for visual processing, serving as the basis to probe the hijack-hallucination link. We introduce the **Hijacked Attention Ratio (HAR)** to quantify severity of hijacking within this critical visual stream:

$$\text{HAR}^{(\ell,h)} \triangleq \frac{\sum_{v \in \mathcal{I}_{\text{inert}}} A_v^{(\ell,h)}}{\sum_{v \in \mathcal{I}_v} A_v^{(\ell,h)}}. \quad (8)$$

Our results reveal a strong correlation: **hallucinated tokens consistently exhibit higher HAR** (Figure 4 (a, b)), indicating that when these critical visual heads are “hijacked,” the resulting loss of grounding precipitates hallucinations.

**From Diagnosis to Intervention: The NHAR Metric.** While HAR diagnoses distraction, selecting effective heads requires a dual filter: ensuring high visual intensity *and* low hijacking. To

identify heads that actively encode meaningful features using a single standard, we propose the **Non-Hijacked Visual Attention Ratio (NHAR)**:

$$\text{NHAR}^{(\ell,h)}(y_k) \triangleq \sum_{v \in \mathcal{I}_v \setminus \mathcal{I}_{\text{inert}}} A_{k,v}^{(\ell,h)}. \quad (9)$$

Unlike HAR, NHAR acts as a **unified density metric**, directly measuring the global attention budget allocated to valid visual content. Validating on the top-450 NHAR-ranked heads reveals a stark contrast: **Real Object Tokens correlate with high NHAR, whereas Hallucinated Tokens cluster in low-NHAR regions** (Figure 4 (c, d)). This confirms NHAR as a robust discriminator for factual grounding, justifying its role as our primary selection criterion. Additional statistics for other models are provided in Appendix B.2.

## 5 Hijacking-Aware Visual Attention Enhancement

Leveraging the **NHAR** metric defined in Section 4.2, we propose **Hijacking-Aware Visual Attention Enhancement (HVAE)**. This training-free framework operates in two stages to precisely identify and reinforce the attention heads critical for factual grounding, incurring zero additional computational overhead.

**Stage 1: Principled Head Selection.** We first operationalize the NHAR metric to curate a set of critical heads. For each head  $(\ell, h)$ , we compute a global stability score,  $\overline{\text{NHAR}}$ , by averaging its performance across the real-object tokens ( $\mathcal{O}_{\text{real}}$ ) defined in Section 3.2:

$$\overline{\text{NHAR}}^{(\ell,h)} = \frac{1}{|\mathcal{O}_{\text{real}}|} \sum_{y_k \in \mathcal{O}_{\text{real}}} \text{NHAR}^{(\ell,h)}(y_k). \quad (10)$$

The top- $K$  heads ranked by  $\overline{\text{NHAR}}$  constitute the target set  $\mathcal{H}_{\text{target}}$ , representing the model’s primary channels for reliable visual processing.

Model	Method	CHAIR			POPE		POPE Chat	
		CHAIR <sub>s</sub> ↓	CHAIR <sub>i</sub> ↓	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
LLaVA-1.5-7B	Greedy	48.2	14.2	76.4	84.8	85.5	85.5	83.4
	PAI	23.8	6.2	76.8	85.9	86.0	85.5	83.4
	Devils	27.2	7.0	76.1	85.5	85.8	87.6	86.9
	VISTA	15.6	5.2	67.3	83.1	84.6	—	—
	<b>HAVAEOurs</b>	<b>18.2</b> <sup>-23.5%</sup>	<b>3.8</b> <sup>-38.7%</sup>	76.7	<b>86.2</b> <sup>+0.3%</sup>	<b>86.3</b> <sup>+0.3%</sup>	<b>88.0</b> <sup>+0.5%</sup>	<b>87.0</b> <sup>+0.1%</sup>
MiniGPT-4-7B	Greedy	28.2	8.8	73.7	76.8	76.6	77.7	76.9
	PAI	22.6	7.6	72.9	74.7	76.3	79.1	78.8
	Devils	21.9	7.9	71.5	72.3	75.9	79.4	78.7
	VISTA	18.0	4.3	68.3	76.6	77.4	—	—
	<b>HAVAEOurs</b>	<b>21.8</b> <sup>-0.5%</sup>	<b>6.9</b> <sup>-9.2%</sup>	72.5	<b>76.9</b> <sup>+0.1%</sup>	<b>77.6</b> <sup>+0.3%</sup>	<b>80.2</b> <sup>+1.0%</sup>	<b>80.2</b> <sup>+1.8%</sup>
Shikra-7B	Greedy	56.8	14.8	75.4	81.0	81.7	76.4	78.3
	PAI	36.1	9.8	75.4	81.3	81.1	76.5	77.5
	Devils	26.2	9.3	73.0	80.5	80.4	75.7	77.7
	VISTA	32.8	9.8	73.4	81.3	81.9	—	—
	<b>HAVAEOurs</b>	<b>15.8</b> <sup>-39.7%</sup>	<b>5.0</b> <sup>-46.2%</sup>	71.8	<b>81.6</b> <sup>+0.4%</sup>	<b>82.1</b> <sup>+0.2%</sup>	<b>76.7</b> <sup>+0.3%</sup>	<b>78.6</b> <sup>+0.4%</sup>
LLaVA-1.5-13B	Greedy	41.6	11.1	79.3	<b>82.6</b>	84.5	85.4	83.2
	Devils	29.0	8.6	79.9	71.4	77.2	87.8	86.4
	<b>HAVAEOurs</b>	<b>21.8</b> <sup>-24.8%</sup>	<b>5.0</b> <sup>-41.9%</sup>	79.8	82.5 <sup>-0.1%</sup>	<b>84.7</b> <sup>+0.2%</sup>	<b>87.9</b> <sup>+0.1%</sup>	<b>86.6</b> <sup>+0.2%</sup>

Table 1: Performance of **HAVAEOurs** against baselines. Best results are in **bold**. Pink cells mark potentially unreliable CHAIR scores. Superscripts show the % change vs. the best baseline (excluding unreliable scores).

Metric	Method	LLaVA 1.5-7B	MiniGPT-4 7B	Shikra 7B	LLaVA 1.5-13B
Perception	Vanilla	1472.5	731.9	962.0	1515.0
	<b>HAVAEOurs</b>	<b>1483.9</b>	<b>744.8</b>	<b>976.0</b>	<b>1528.9</b>
Cognition	Vanilla	322.5	173.0	250.4	290.6
	<b>HAVAEOurs</b>	<b>327.9</b>	<b>183.9</b>	<b>272.5</b>	<b>295.4</b>

Table 2: MME Benchmark results. Comparison between vanilla models and **HAVAEOurs** on Perception and Cognition subsets.

## Stage 2: Collective Attention Reinforcement.

During inference, we selectively intervene on  $\mathcal{H}_{\text{target}}$ . To amplify their focus without introducing external noise, we augment their attention weights using the layer-wise mean attention magnitude:

$$\mathbf{A}_{k,i}^{(\ell,h)} \leftarrow \mathbf{A}_{k,i}^{(\ell,h)} + \alpha \frac{1}{H} \sum_{h'=1}^H \left| \mathbf{A}_{k,i}^{(\ell,h')} \right|, \quad \forall v_i \in \mathcal{I}_v, \quad (11)$$

where  $\alpha$  controls the enhancement strength. By injecting this collective visual signal into the non-hijacked heads, HAVAEOurs effectively counteracts Vocabulary Hijacking. The complete procedure is summarized in Algorithm 2.

## 6 Experiments

### 6.1 Experimental Setup

**Models.** We evaluate our method on LLaVA-1.5 (7B, 13B) (Liu et al., 2023), Shikra (7B) (Chen et al., 2023), MiniGPT-4 (7B) (Zhu et al., 2023),

Model	Method	CHAIR			POPE		MME
		CHAIR <sub>s</sub> ↓	CHAIR <sub>i</sub> ↓	F1 ↑	Acc. ↑	F1 ↑	All ↑
Qwen2-VL	Vanilla	27.6	8.8	79.2	89.1	88.1	2268.4
	<b>HAVAEOurs</b>	<b>22.8</b>	<b>6.2</b>	<b>79.4</b>	<b>89.3</b>	<b>88.9</b>	<b>2290.2</b>

Table 3: Performance comparison on Qwen2-VL across CHAIR, POPE, and MME benchmarks.

and Qwen2-VL (7B) (Wang et al., 2024) to assess its generalizability and scalability.

**Evaluation Benchmarks.** Following prior work (Liu et al., 2024; Jiang et al., 2025), we evaluate our method on several hallucination benchmarks. We use **CHAIR** (Rohrbach et al., 2018) for captioning evaluation and **POPE** (Li et al., 2023) for query-based object probing, along with its conversational variant, **POPE-Chat**. To assess out-of-domain performance and fine-grained errors, we also employ **AMBER** (Wang et al., 2023). Finally, to investigate the impact on general capabilities, we conduct evaluations on **MME** (Fu et al., 2023). Further details are provided in Appendix E.1.

**Baselines.** We evaluate HAVAEOurs against representative training-free hallucination mitigation strategies. Our primary comparisons focus on attention-based interventions, specifically **PAI** (Liu et al., 2024) and **Devils** (Jiang et al., 2025), which are most relevant to our approach. Additionally, we evaluate against **VISTA** (Li et al., 2025), a leading method for activation steering.

**Implementation Details.** We set the head count

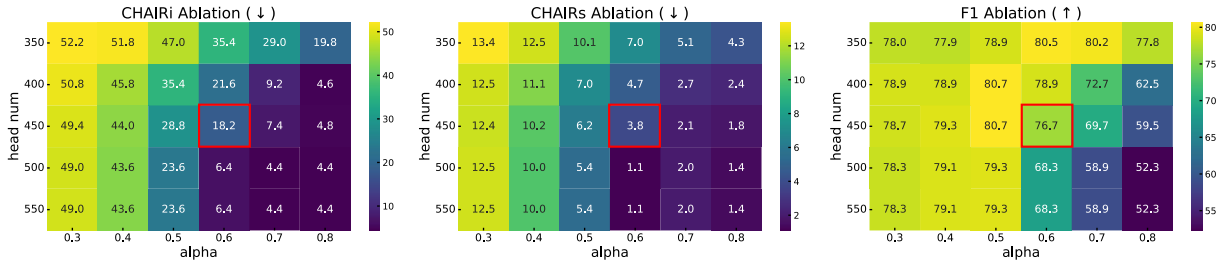


Figure 5: Ablation on hyperparameters  $\alpha$  and  $K$  for LLaVA-1.5 7B. Red boxes highlight the parameter combinations we used.

Model	Selecting Strategy	CHAIR			POPE		POPE Chat		MME	
		CHAIR <sub>s</sub> ↓	CHAIR <sub>i</sub> ↓	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Per. ↑	Cog. ↑
LLaVA-1.5-7B	Max Attention	7.8	4.4	65.8	85.9	85.6	86.0	85.5	1399.0	277.0
	HAVAEOurs	18.2	3.8	76.7	<b>86.2</b>	<b>86.3</b>	<b>88.0</b>	<b>87.0</b>	<b>1483.9</b>	<b>327.9</b>

Table 4: Comparison of head selection strategies on LLaVA-1.5 7B.

$K = 300$  for Qwen2-VL and  $K = 450$  for all other models. The enhancement strength  $\alpha$  defaults to 0.1, but for long-context tasks (Captioning, POPE-Chat), it is increased to 0.6 (LLaVA-1.5-7B, MiniGPT-4, Qwen2-VL) or 0.7 (LLaVA-1.5-13B, Shikra). We adopt greedy decoding as the default strategy for our main experiments. Comprehensive results using alternative decoding methods are provided in Appendix F.1.

## 6.2 Main Results

**Superior Performance and Generalization.** As detailed in Table 1, HAVAEOurs demonstrates clear superiority across all evaluated tasks. In standard settings, it excels in both long-form (CHAIR) and short-form (POPE) generation, reducing the CHAIR<sub>i</sub> score by 38.7% on LLaVA-1.5-7B and 46.2% on Shikra-7B relative to the strongest reliable baseline. Crucially, this effectiveness extends to scenarios beyond the calibration domain. When applied to the AMBER benchmark—which contains images distinct from the COCO source used for head selection—HAVAEOurs maintains significant gains (Table 14, see Appendix F.2). Furthermore, even in the absence of ground-truth annotations for calibration (analyzed in Appendix G.2), our method sustains high performance. These results confirm that HAVAEOurs captures intrinsic, model-specific anti-hallucination mechanisms rather than overfitting to specific datasets or supervision signals.

**Preserving General Capabilities.** To verify that HAVAEOurs preserves fundamental capabilities, we evaluate on the comprehensive MME benchmark.

As shown in Table 2, HAVAEOurs consistently surpasses vanilla baselines across all tested architectures. This suggests that by suppressing the attention noise of Inert Tokens, our method mitigates hallucinations while maintaining—and often enhancing—general visual perception and reasoning. **Scalability and Architectural Generalization.** To assess the robustness of our approach, we extend our evaluation to larger scales and advanced architectures. As shown in Table 1, HAVAEOurs yields significant improvements on LLaVA-1.5 13B, confirming its scalability. Furthermore, validation on the advanced Qwen2-VL 7B (Table 3) demonstrates consistent gains over baselines. These combined results confirm that HAVAEOurs effectively generalizes across diverse model sizes and structural designs.

## 6.3 Ablation Studies and Analysis

**Impact of Hyperparameters  $\alpha$  and  $K$ .** We investigate  $\alpha \in [0.3, 0.8]$  and  $K \in [350, 550]$ , observing that precise tuning is particularly critical in long-context scenarios. Results on LLaVA-1.5 7B (Figure 5) reveal a trade-off: aggressive settings effectively reduce hallucinations (lower CHAIR) but degrade generation quality (lower F1). Consequently, we select parameters to balance suppression with fluency. Additional ablations are provided in Appendix F.3.

**Superiority of NHAR for Head Selection.** We demonstrate NHAR’s superiority by ablating its core component: hijacking-awareness. As shown in Table 4, selecting heads based only on total visual attention ( $\mathcal{I}_{\text{inert}} = \emptyset$ ) **significantly lowers F1 scores** across all benchmarks on LLaVA-1.5 7B.

Please help me describe the image in detail.



#### LLaVA-1.5 7B

The image features a woman playing tennis on a clay court. She is in a ready position, holding a tennis racket and preparing to hit the ball. The woman is wearing a red shirt and appears focused on the game. **The tennis ball is in the air**, close to the woman, and she is poised to make contact with it. The scene captures the intensity and skill of the sport, as the woman is fully engaged in the game.

#### HAVAe (ours)

The woman is wearing a red shirt and a white hat while playing tennis on a clay court. She is holding a **tennis racket** and is in a ready position to hit the ball. The court is brown and the woman is in the middle of the court.

Figure 6: A case study demonstrates HAVAe correcting a baseline hallucination by redirecting the model’s focus from irrelevant Inert tokens (as shown in the heatmap) back to the salient visual target.

Num	CHAIR		POPE		
	CHAIR <sub>s</sub>	CHAIR <sub>i</sub>	F1	Acc	F1
10	18.8	3.7	76.5	86.1	86.2
100	18.8	3.7	76.5	85.9	86.1
300	18.6	4.8	76.9	85.9	86.1
500	18.2	3.7	76.7	86.1	86.2
1000	18.2	3.7	76.7	86.1	86.2

Table 5: Ablation study on the sample size for LLaVA-1.5 7B.

This confirms high raw visual attention is a flawed proxy for effective grounding, easily captured by redundant Inert Tokens. NHAR’s crucial advantage lies in filtering this noise, isolating heads that truly focus on salient visual content.

**Ablation on Sample Size.** We performed an ablation on the number of samples used for our analysis in Section 3.2. As shown in Table 5, key performance metrics across all benchmarks stabilize once the sample size reaches 500. Therefore, we adopt a sample size of 500 for all experiments, as this provides a robust estimate of performance without incurring unnecessary computational costs.

**Effect of Penalizing Attention to Inert Tokens.** Positing that **Inert Tokens** are redundant, we hypothesized that explicitly penalizing their attention might improve performance. We introduced a penalty factor  $\beta$ , where  $\beta = 0$  corresponds to the standard HAVAe configuration. As  $\beta$  increases, the suppression of attention on Inert Tokens progressively intensifies. However, results in Table 6 indicate that this strategy offers no benefit and is in fact detrimental, causing notable performance degradation on CHAIR. This observation, consistent with prior work (Gu et al., 2025), confirms that suppressing attention to these tokens impairs generation, thereby validating our design

$\beta$	CHAIR		POPE		
	CHAIR <sub>s</sub>	CHAIR <sub>i</sub>	F1	Acc	F1
0.0	18.2	3.7	76.7	86.1	86.2
0.1	20.2	4.5	76.8	86.1	86.2
0.3	19.0	4.1	77.0	86.2	86.2
0.6	19.8	4.7	76.8	86.1	86.2
0.9	19.2	4.7	76.9	86.2	86.3

Table 6: Ablation on the sink token penalty factor  $\beta$  for LLaVA-1.5 7B.

choice to employ positive attention reinforcement rather than negative suppression.

**Origins of Vocabulary Hijacking.** We hypothesize that this vocabulary hijacking is an artifact deeply rooted in the initial vision-language alignment process. When we analyzed the base LLaVA-1.5 model (which only underwent multimodal pre-training without subsequent instruction tuning), we observed that the vocabulary hijacking phenomenon was significantly more severe. In this early stage, almost all attention weights were monopolized by these Inert tokens. This suggests that during pre-training, the model may learn a shortcut to dump unaligned or redundant visual features into specific vocabulary anchors.

## 6.4 Hyperparameter Sensitivity Analysis

To quantitatively evaluate the stability of our Hijacking Anchor-Based Identification (HABI) metric and the overall HAVAe method, we conduct ablation studies by varying the identification thresholds  $\tau_r$  and  $\tau_s$ . As shown in Table 7 and Table 8, we scale the thresholds from  $0.8\times$  to  $1.2\times$  of our empirically established optimal values ( $\tau^*$ ).

The experimental results clearly indicate that the performance remains highly stable under these fluctuations. This analysis confirms that the proposed

$\tau_r/\tau_r^*$	CHAIR		POPE		
	CHAIR <sub>s</sub>	CHAIR <sub>i</sub>	F1	Acc	F1
0.8	18.5	4.3	76.9	86.3	86.2
0.9	18.0	4.0	76.9	86.1	86.4
<b>1.0 (Ours)</b>	18.2	3.8	76.7	86.2	86.3
1.1	17.9	3.9	76.8	86.1	86.3
1.2	18.3	4.1	76.6	86.2	86.3

Table 7: Sensitivity Analysis of  $\tau_r$ . HVAE exhibits strong robustness to variations in the  $\tau_r$  threshold.

method is robust and not overly sensitive to minor variations in hyperparameter selection.

Due to space limitations, additional experimental results and analyses are presented in Appendix G.

## 6.5 Case Study

A case study in Figure 6 (d) illustrates our method’s effectiveness. While the baseline hallucinates that “The tennis ball is in the air”, our method provides a factually accurate description. The attention visualizations reveal the mechanism behind this correction: the baseline, when generating “ball”, suffers from severe **Vocabulary Hijacking**, where its attention is captured by **Inert Tokens** and scattered onto irrelevant background. In contrast, the heads selected by HVAE demonstrate a precise focus on the target object when generating “racket”, thereby preventing the hallucination. Additional case studies are presented in Appendix H.

## 7 Conclusion

This paper provides the first mechanistic analysis of the **Vocabulary Hijacking** phenomenon in LVLMs. We reveal a core property where specific visual tokens, termed **Inert Tokens**, hijack the model’s attention and consistently decode to a fixed set of **Hijacking Anchors**. This discovery forms the foundation for our **Hijacking Anchor-Based Identification (HABI)** method, which reliably localizes these **Inert Tokens**. Building on this, we propose **Hijacking-Aware Visual Attention Enhancement (HVAE)**, a training-free intervention that leverages our **Non-Hijacked Visual Attention Ratio (NHAR)** metric to identify hallucination-critical heads and selectively strengthen their focus on salient visual content. Our experimental results demonstrate that HVAE significantly reduces hallucination with **zero computational overhead**, while preserving the model’s general capabilities. We believe our work contributes to a deeper, mechanistic understanding of attention anomalies

$\tau_s/\tau_s^*$	CHAIR			POPE	
	CHAIR <sub>s</sub>	CHAIR <sub>i</sub>	F1	Acc	F1
0.8	18.6	4.5	76.6	86.2	86.3
0.9	17.8	3.9	76.8	86.3	86.2
<b>1.0 (Ours)</b>	18.2	3.8	76.7	86.2	86.3
1.1	18.1	4.1	77.0	86.3	86.2
1.2	18.4	4.2	76.9	86.1	86.3

Table 8: Sensitivity Analysis of  $\tau_s$ . HVAE exhibits strong robustness to variations in the  $\tau_s$  threshold.

and their causal link to hallucination, offering a new direction for improving the reliability of LVLMs.

## Acknowledgements

This work was supported in part by National Natural Science Foundation of China (62476070), Shenzhen Science and Technology Program (JCYJ2024 1202123503005, GXWD20231128103232001, Z DSYS20230626091203008, KQTD20240729102 154066), Department of Science and Technology of Guangdong (2024A1515011540), National Key R&D Program of China (SQ2024YFE0200592) and Suzhou Science and Technology Program (SYG2025072).

## Limitations

First, while this work uncovers the phenomenon of Vocabulary Hijacking, its mechanistic origin remains to be fully elucidated. Specifically, the precise training dynamics that precipitate this behavior remain unclear. Moreover, understanding why a distinct subset of semantically meaningless tokens crystallize into Hijacking Anchors is a complex question we leave for future investigation. Second, constrained by computational resources, our experimental validation is currently limited to models with sizes up to 13B parameters. Extending this analysis to larger-scale foundational models represents a vital direction for follow-up research. Finally, exploring training-based strategies to mitigate Vocabulary Hijacking constitutes another critical avenue for future work.

## Ethical Considerations

Our method, HVAE, is a training-free intervention validated on standard publicly available benchmarks (e.g., COCO, POPE). We explicitly state that the efficacy of our findings may be influenced by the distribution of these evaluation datasets. Consequently, the performance of our method on highly

specialized domains or privacy-sensitive data remains unverified. Thus, potential risks exist when deploying HVAE in high-stakes applications (e.g., medical or legal consultation). We strongly advise users to exercise caution and maintain human oversight to verify the factual correctness of the model’s responses, as the mitigation of hallucination does not guarantee the complete elimination of errors.

## References

- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.
- Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Björn Deiseroth, Mayukh Deb, Samuel Weinbach, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2023. Atman: Understanding transformer predictions through memory efficient attention manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 63437–63460. Curran Associates, Inc.
- Guodong Du, Zitao Fang, Jing Li, Junlin Li, Runhua Jiang, Shuyang Yu, Yifei Guo, Yangneng Chen, Sim Kuan Goh, Ho-Kin Tang, Daojing He, Honghai Liu, and Min Zhang. 2025. [Neural parameter search for slimmer fine-tuned models and better transfer](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Guodong Du, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim Kuan Goh, Ho-Kin Tang, Daojing He, and Min Zhang. 2024. Parameter competition balancing for model merging. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.
- Guodong Du, Zhuo Li, Xuanning Zhou, Junlin Li, Zesheng Shi, Wanyu Lin, Ho-Kin Tang, Xiucheng Li, Fangming Liu, Wenya Wang, Min Zhang, and Jing Li. 2026. Knowledge fusion of large language models via modular skillpacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. 2024. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.
- Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. 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 (CVPR)*, pages 25004–25014.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhen-ting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. 2025. The hidden

- life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Forty-second International Conference on Machine Learning (ICLR)*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2024. Paying more attention to image: A training-free method for alleviating hallucination in vlms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 125–140.
- nostalgebraist. 2020. Interpreting gpt: The logit lens. <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>.
- Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4035–4045.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13088–13110.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2025. Don’t miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.
- Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024a. 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 (CVPR)*, pages 13807–13816.
- Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.
- Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2024. Tell your model where to attend: Post-hoc attention steering for LLMs. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*.
- Yiyang Zhou, Chenhao Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

## Appendix

<b>A</b>	<b>The Use of Large Language Models</b>	<b>12</b>
<b>B</b>	<b>More Statistical Results</b>	<b>12</b>
B.1	Empirical Results for HABI on Additional Models . . . . .	12
B.2	NHAR and HAR Distributions on Additional Models . . . . .	12
B.3	Positional Distribution of Inert Tokens . . . . .	14
B.4	Quantifying the Attention Attraction of Inert Tokens . . . . .	14
<b>C</b>	<b>Comparison with Visual Attention Sink</b>	<b>15</b>
C.1	Verification of General Characteristics . . . . .	15
C.2	Analysis of Critical Properties . . . . .	15
<b>D</b>	<b>Comparison with Concurrent Work: AVISC</b>	<b>17</b>
D.1	Fundamental Mechanistic Differences . . . . .	17
D.2	Empirical Superiority and Computational Efficiency . . . . .	18
<b>E</b>	<b>Additional Experimental Setups</b>	<b>18</b>
E.1	Detailed Benchmark and Evaluation Metrics . . . . .	18
<b>F</b>	<b>Supplementary Experiments Results</b>	<b>20</b>
F.1	Performance with Alternative Decoding Strategies . . . . .	20
F.2	AMBER Results on Additional Models . . . . .	21
F.3	Supplementary Ablation Results . . . . .	21
<b>G</b>	<b>Supplementary Experiments and Analysis</b>	<b>21</b>
G.1	Qualitative Validation via Clustering and Visualization . . . . .	21
G.1.1	High-Attention Manifold: Anchors vs. Content . . . . .	21
G.1.2	High-Frequency Manifold: Noise vs. Substrate . . . . .	22
G.1.3	Long-Tail & Generic Groups . . . . .	22
G.2	Robustness to the Absence of Ground-Truth Annotations . . . . .	23
<b>H</b>	<b>Additional Case Studies</b>	<b>23</b>
H.1	Qualitative Error Analysis . . . . .	23
<b>I</b>	<b>Algorithms</b>	<b>23</b>

## A The Use of Large Language Models

Throughout the preparation of this manuscript, large language models were employed exclusively for light stylistic refinement and the occasional grammatical adjustment. Every conceptual insight, analytical thread, and interpretive conclusion emerged from the authors themselves; no algorithmic assistance was solicited for the framing, design, or substance of the work, and full scientific responsibility rests with the human contributors alone.

## B More Statistical Results

### B.1 Empirical Results for HABI on Additional Models

This section validates the universality of our HABI identification strategy across diverse architectures, including MiniGPT-4, Shikra, LLaVA-1.5 13B, and Qwen2-VL-7B. As visualized in Figure 7, all models exhibit statistical behaviors consistent with LLaVA-1.5 7B: a **long-tailed distribution** of Hijacking Scores and a distinct **bimodal distribution** of Hijacking Ratios among salient tokens. These ubiquitous patterns confirm that the Vocabulary Hijacking phenomenon is intrinsic to LVLMs, thereby justifying the robustness of our statistical thresholds for reliably isolating **Inert Tokens**.

We note specific adaptation details for distinct architectures: due to the distribution characteristics of Qwen2-VL-7B, we employed the third quartile ( $Q_3$ ) as the threshold for identifying Hijacking Anchors, rather than the standard  $Q_3 + 1.5 \times IQR$ . Additionally, for MiniGPT-4, given its limited visual sequence length (only 32 tokens), we calculated the Hijacking Ratio using the entire set of visual tokens, bypassing the top 5% attention filtering.

### B.2 NHAR and HAR Distributions on Additional Models

Figure 8 (a) illustrates the HAR distributions for LLaVA-1.5 7B and 13B, confirming that hallucinated tokens are consistently associated with elevated HAR scores. By synthesizing these results with those in the main text (Figure 4), we observe the varying severity of Vocabulary Hijacking across the tested models: Shikra exhibits the most pronounced susceptibility, while LLaVA-1.5 7B shows the mildest effect.

To substantiate this observation, we present a visualization of Vocabulary Hijacking on Shikra in Figure 9. The attention map reveals that a predominant portion of the model’s focus is captured by

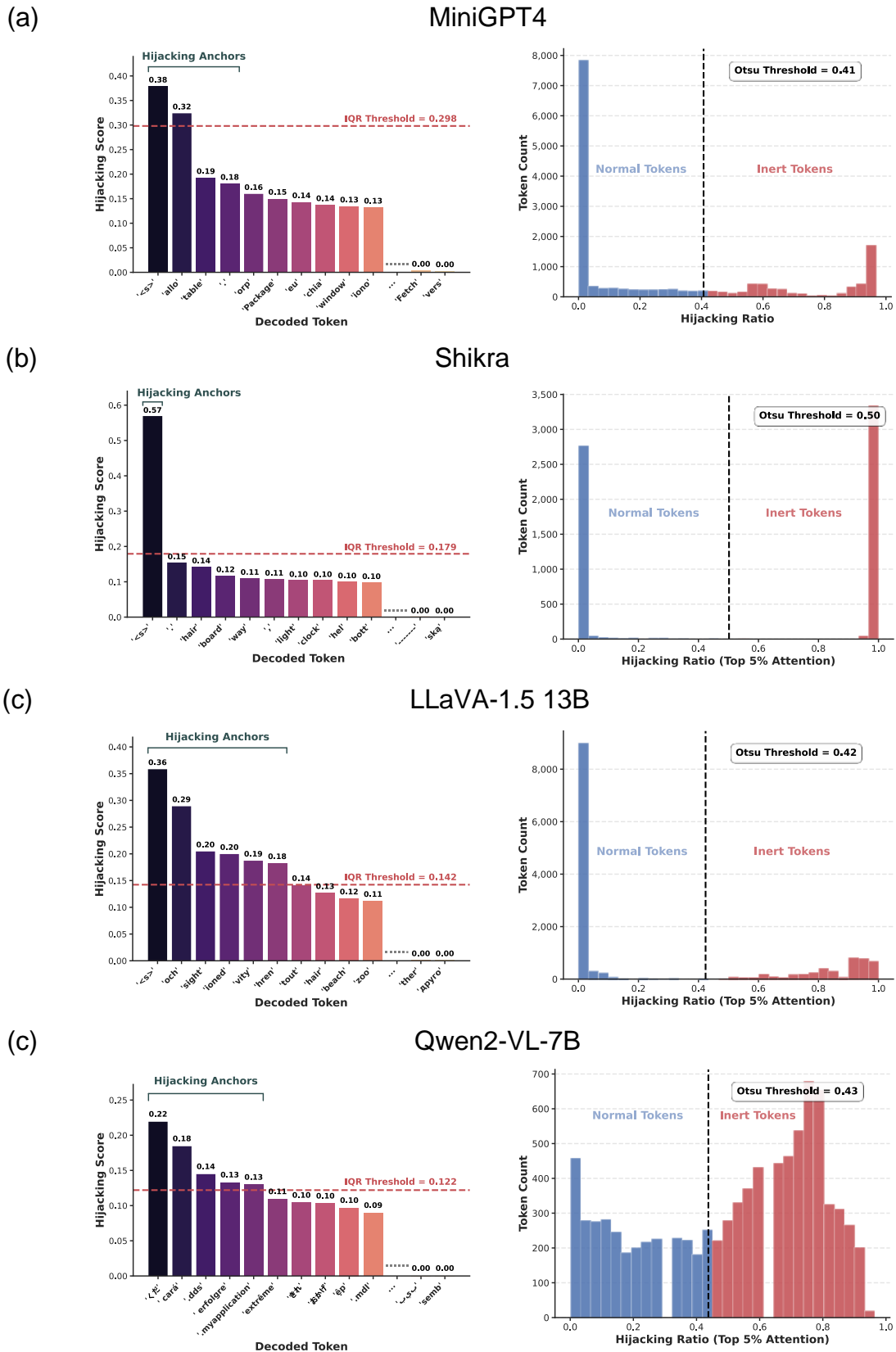


Figure 7: Validation of the HABI identification strategy across diverse LVLm architectures. Consistent with the findings on LLaVA-1.5 7B, all tested models exhibit the characteristic long-tailed distribution of Hijacking Scores (enabling **Hijacking Anchor** identification) and the bimodal distribution of Hijacking Ratios (facilitating **Inert Token** isolation). The figure displays results for: (a) MiniGPT-4, (b) Shikra, (c) LLaVA-1.5 13B, and (d) Qwen2-VL-7B.



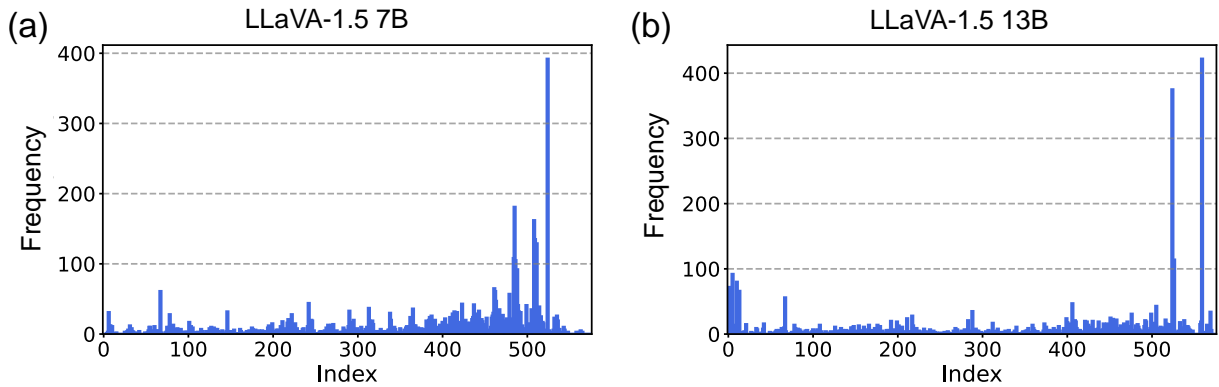


Figure 10: Positional distribution of Inert tokens for LLaVA-1.5 7B and 13B within the 576-token visual sequence. The plots show that Inert tokens are not uniformly distributed, but are instead highly concentrated in specific index ranges. This suggests that the Vocabulary Hijacking phenomenon may be intrinsically linked to the mechanics of the attention mechanism.

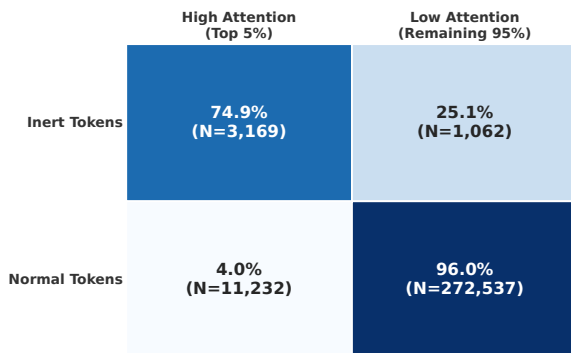


Figure 11: **Conditional Probability of Attention Distribution.** Although Inert Tokens (top row) constitute only  $\sim 1\%$  of the total visual tokens, over 75% of them are concentrated in the High Attention (Top 5%) region. In contrast, Normal Tokens (bottom row) follow a standard distribution. This confirms the hijacking nature of Inert Tokens, demonstrating their capacity to actively capture the model’s focus.

model’s focus despite their scarcity.

## C Comparison with Visual Attention Sink

In this section, we clarify the relationship between Vocabulary Hijacking and the Visual Attention Sink (VAS) phenomenon. While both share certain behavioral characteristics, we demonstrate that our proposed method offers a more precise identification of underlying attention anomalies compared to the activation-based method (Kang et al., 2025).

### C.1 Verification of General Characteristics

We first verify that the **Inert Tokens** identified by HABI share the defining characteristics typically associated with the Visual Attention Sink (VAS) phenomenon: negligible functional contribution

and background localization.

**Low Functional Contribution (Zero-Ablation Study).** To assess the functional importance of these tokens, we conducted a zero-ablation experiment on a subset of the POPE benchmark. As shown in Figure 12 (a), masking out either the **Inert Tokens** (identified by HABI) or the VAS tokens (identified by Massive Activation) has a negligible impact on the F1 score. This outcome confirms that both token categories correspond to informationally sparse regions that are functionally redundant for the generation task.

**Background Localization.** We quantitatively analyze the spatial distribution of these tokens relative to semantic regions. By utilizing ground-truth segmentation masks, we define the image background as the complementary area to all annotated objects. As visualized in Figure 12 (b), both **Inert Tokens** and VAS tokens exhibit a significantly higher probability of residing in background regions compared to a random baseline. This empirical evidence confirms that both token types are predominantly localized within the background.

### C.2 Analysis of Critical Properties

While background localization and functional redundancy are observable symptoms, we argue that they are merely *secondary* characteristics. The **definitive feature** of this phenomenon is **Persistent Attention Attraction**—the capacity of specific tokens to consistently monopolize attention across multiple generation steps, disregarding the evolving context.

To quantitatively evaluate which identification paradigm better captures this core anomaly, we

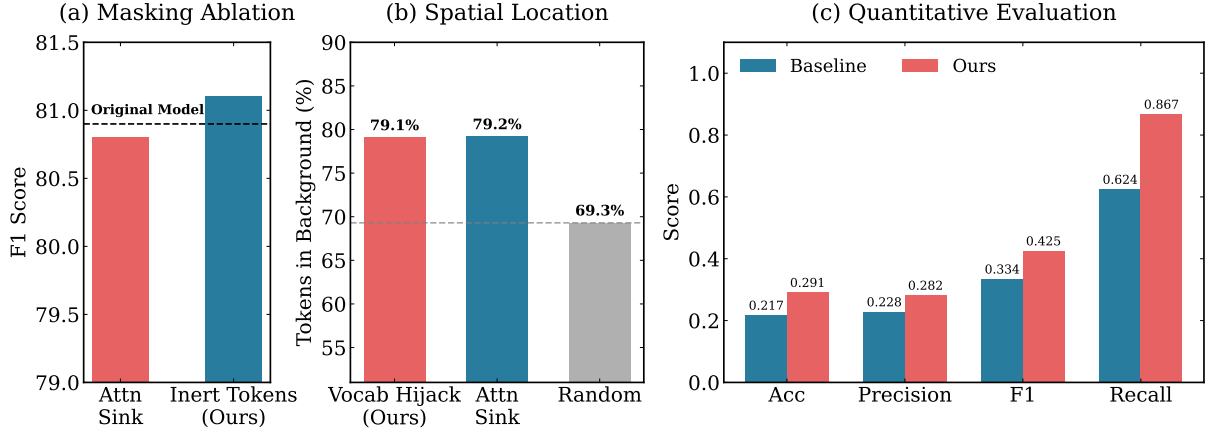


Figure 12: **Comparative analysis of attention anomaly identification methods.** (a) **Zero-Ablation Verification:** Changes in POPE F1 scores when masking out specific token sets (Massive Activation vs. HABI), used to quantify the functional redundancy of the identified tokens. (b) **Spatial Alignment:** The proportion of tokens falling within image background regions (defined by segmentation masks) across different selection methods (Random, Massive Activation, HABI). (c) **Identification Accuracy:** Evaluation of the alignment with the behavioral ground truth ( $\mathcal{G}_{\text{persist}}$ ). The chart reports Precision, Recall, and F1 scores, measuring the capability of each method to retrieve tokens that exhibit persistent attention patterns.

constructed a behavior-based ground truth and performed a comparative analysis between our **Hijacking Anchor-Based Identification (HABI)** and the **Massive Activation method** used in prior work (Kang et al., 2025).

### Constructing the Behavioral Ground Truth.

Given the absence of explicit annotations for attention anomalies, we establish a behavioral approximation of the ground truth. **Strictly speaking, this constructed set represents a subset of the true ground truth.** We hypothesize that a true hijacking token must consistently attract high attention during the initial generation phase. For each image, we define the **Persistent Attention Set**, denoted as  $\mathcal{G}_{\text{persist}}$ , by computing the intersection of the top- $K$  attended vision tokens across the first  $T$  decoding steps. Formally, let  $\mathcal{S}_t$  be the set of indices for vision tokens with the top- $K$  attention weights at step  $t$ . The ground truth is defined as:

$$\mathcal{G}_{\text{persist}} = \bigcap_{t=1}^T \mathcal{S}_t. \quad (12)$$

In our experiments, we set  $T = 5$  and  $K = 10$ , aiming to capture tokens that act as stable attention sinks during the critical initial captioning phase.

**Comparative Results.** We evaluate two identification methods against  $\mathcal{G}_{\text{persist}}$ :

1. **Massive Activation (Baseline):** Identifies tokens where specific dimensions of the hidden

states exhibit anomalously high activation values (Kang et al., 2025).

2. **HABI (Ours):** Identifies **Inert Tokens** using the proposed Hijacking Ratio thresholding ( $r_{\text{hijack}} > \tau_r$ ).

We calculate the Accuracy, Precision, Recall, and F1-score of both methods in retrieving the tokens in  $\mathcal{G}_{\text{persist}}$ . As shown in Figure 12 (c), our HABI method significantly outperforms the Massive Activation baseline across all evaluated metrics.

Crucially, to quantify the distinctness of the identified anomalies, we highlight that HABI achieves a Unique Ratio of 42.93%, meaning that nearly half of the Inert Tokens we identify are completely overlooked by the activation-based baseline.

This substantial divergence suggests that **activation magnitude is an insufficient proxy for attention anomalies.** While many sink tokens do possess large activations, a substantial portion of “hijackers” (which persistently attract attention) do not exhibit massive norms but are accurately detected by our HABI. Thus, HABI provides a more robust and mechanically grounded identification of the attention hijacking phenomenon.

### Case Study: Failure of the Activation-based Method.

To illustrate the limitation of the baseline, we revisit the example from Figure 2, where the model generated the token “phone”. Among

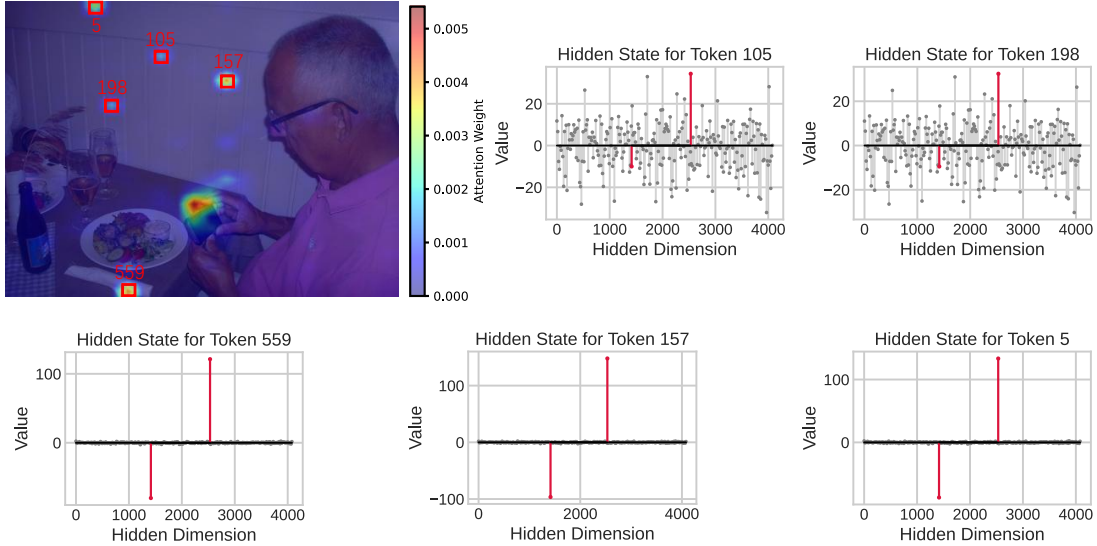


Figure 13: **Visualization of hidden state distributions for five identified Inert Tokens at layer 5.** The red vertical lines indicate the specific dimensions (1415 and 2533) used by the Massive Activation baseline to detect anomalies. Crucially, while some tokens exhibit the expected spikes, **tokens 105 and 198 do not show massive activation in the designated dimensions.** This absence demonstrates why the activation-based method fails to detect these potent hijackers, whereas our HABI method successfully identifies them.

the top-10 most attended visual tokens, five (indices 5, 105, 157, 198, and 559) were identified as clear persistent sinks. We visualize the layer-5 hidden state distributions for these five tokens in Figure 13, specifically highlighting the massive activation dimensions  $\mathcal{D}_{\text{sink}} = \{1415, 2533\}$  identified by (Kang et al., 2025). The visualization reveals a critical discrepancy: while several tokens exhibit the massive activation pattern, **tokens 105 and 198 do not show massive activation in the designated dimensions.**

Consequently, the activation-based method fails to identify these two potent attention hijackers. In contrast, HABI successfully flags them due to their rigid Traces, demonstrating superior sensitivity to the phenomenon’s true nature.

## D Comparison with Concurrent Work: AVISC

Following our analysis distinguishing Vocabulary Hijacking from the Visual Attention Sink (VAS) phenomenon in Appendix C, we now clarify the relationship between our method and a concurrent work, AVISC (Woo et al., 2025). While both works aim to mitigate LVLMM hallucinations, our concept of *Vocabulary Hijacking* (driven by Inert Tokens) and AVISC’s *Blind Tokens* represent fundamentally distinct phenomena. We elaborate on these differences from two primary perspectives: mechanistic behavior and empirical performance.

Comparison Pair	Jaccard Dist	Unique Ratio
Inert Tokens vs. Attn Sink	0.5596	0.4291
Inert Tokens vs. Blind Tokens	0.8302	0.0926
Attn Sink vs. Blind Tokens	0.8852	0.2942

Table 9: Three-Way Set Divergence Analysis. A higher Jaccard Distance indicates less overlap between the identified token sets.

### D.1 Fundamental Mechanistic Differences

Beyond the divergence from VAS discussed previously, quantitative set analysis demonstrates that AVISC’s Blind Tokens and our Inert Tokens capture entirely disjoint subsets of visual features.

**Set Divergence.** We expanded our set divergence analysis to include Blind Tokens, measuring the overlap between different token sets via Jaccard Distance and Unique Ratio. As reported in Table 9, the remarkably high Jaccard Distance (0.8302) between Inert Tokens and Blind Tokens confirms that they represent highly distinct subsets. Furthermore, the distance between VAS (Attn Sink) and Blind Tokens is even higher (0.8852), indicating that Blind Tokens are distinct from both Vocabulary Hijacking and the attention sink phenomenon.

**Background Distribution Discrepancy.** In Appendix C, we empirically established that both Inert Tokens and VAS tokens are predominantly localized within the image background (approximately

Token Type	Total Count	In Background	BG Ratio (%)
Inert Tokens	16,438	12,972	78.91
Attn Sink	15,030	11,903	79.19
Blind Tokens	8,785	6,127	69.75
Random	20,000	13,798	68.99

Table 10: Background Distribution Analysis. Unlike Inert Tokens and Attention Sinks, which show a strong tendency to concentrate in the visual background, Blind Tokens align closely with a random spatial distribution.

79% for both). In stark contrast, as detailed in Table 10, the background ratio of Blind Tokens (69.75%) is merely on par with that of randomly sampled tokens (68.99%). Because Blind Tokens are heuristically defined by simple attention-score thresholding, they are prone to mistakenly selecting foreground semantic targets. Conversely, Inert Tokens represent a highly specific, mechanistically driven semantic collapse firmly rooted in non-semantic background regions.

## D.2 Empirical Superiority and Computational Efficiency

The precision of our mechanistic identification directly translates into performance and efficiency gains. As shown in Table 11, our HAVAE method consistently outperforms AVISC across standard hallucination benchmarks, including POPE, MME, and AMBER.

To ensure a rigorous and fair setup during our evaluation, we set the maximum generation length to 512 tokens for AVISC, which accounts for the numerical differences compared to their original reported results. Despite this standardized setting, HAVAE consistently achieves higher Accuracy and F1 scores.

Crucially, HAVAE maintains a significantly lower computational footprint. Unlike AVISC, which relies on contrastive decoding and thus requires an additional forward pass, HAVAE is a single-pass intervention that selectively reinforces attention heads. Consequently, our method achieves superior hallucination mitigation while consuming approximately half the inference computing resources required by AVISC.

## E Additional Experimental Setups

### E.1 Detailed Benchmark and Evaluation Metrics

**CHAIR (Rohrbach et al., 2018).** The Caption Hallucination Assessment with Image Relevance

(CHAIR) metric quantifies hallucination in image captions by comparing generated object mentions against a pre-compiled set of ground-truth objects for each image. An object is considered a hallucination if it is mentioned in the caption but is absent from this ground-truth set. It comprises two scores: instance-level (CHAIR<sub>I</sub>) and sentence-level (CHAIR<sub>S</sub>), calculated as follows:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, \quad (13)$$

$$\text{CHAIR}_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}. \quad (14)$$

Our evaluation is conducted on 500 randomly sampled instances from the MSCOCO 2014 validation set. To specifically assess long-form generation, we adopt the setup from PAI (Liu et al., 2024) and Devils (Jiang et al., 2025), generating descriptions with a ‘max\_new\_tokens’ of 512 using the prompt: “Please help me describe the image in detail.”.

**POPE (Li et al., 2023).** The Polling-based Object Probing Evaluation (POPE) is a benchmark designed within the VQA paradigm to assess object hallucination. It evaluates LLMs by posing binary questions about object presence, such as “Is there a <object> in the image?”. The questions are constructed using objects from three distinct sampling strategies to test different aspects of model knowledge: *random* (objects chosen randomly from the dataset), *popular* (frequently occurring objects), and *adversarial* (objects semantically related to those present in the image). *Experimental Setup.* We evaluate on 500 images from the COCO test set, with 6 questions per split for each image, reporting both Accuracy and F1 scores. Furthermore, following (Liu et al., 2024), to comprehensively examine performance in conversational contexts, we extend the evaluation to include both single-turn and multi-turn dialogues, a setup we term **POPE-Chat**.

**MME (Fu et al., 2023).** The MME benchmark is a comprehensive evaluation suite designed to assess the diverse capabilities of LLMs. It comprises 14 subtasks organized into two primary capability types: *Perception* (e.g., existence, counting, position, color, OCR) and *Cognition* (e.g., commonsense reasoning, numerical calculation). The benchmark employs a concise Yes/No question format to minimize instruction-following variance and

Method	POPE		MME		AMBER		
	Acc. ↑	F1 ↑	Per. ↑	Cog. ↑	CHAIR ↓	Acc. ↑	F1 ↑
AVISC	83.2	84.1	1379.3	321.4	12.6	70.7	75.2
<b>HVAE(Ours)</b>	<b>86.2</b>	<b>86.3</b>	<b>1483.8</b>	<b>327.9</b>	<b>3.6</b>	<b>78.6</b>	<b>82.7</b>

Table 11: Comprehensive performance comparison between AVISC and our proposed HVAE across standard hallucination benchmarks.

Model	Method	CHAIR			POPE		POPE Chat	
		CHAIR <sub>s</sub> ↓	CHAIR <sub>i</sub> ↓	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
LLaVA-1.5-7B	Beam Search	47.6	13.0	79.0	84.7	85.4	85.3	83.2
	PAI	21.6	6.2	75.8	85.1	85.7	88.1	87.0
	Devils	29.0	6.8	80.1	84.9	85.6	<b>88.4</b>	<b>87.6</b>
	VISTA	9.8	5.7	54.3	83.1	84.6	—	—
	<b>HVAE(Ours)</b>	<b>20.0</b> <sup>-7.4%</sup>	<b>5.8</b> <sup>-6.5%</sup>	78.7	<b>86.0</b> <sup>+1.1%</sup>	<b>86.2</b> <sup>+0.6%</sup>	88.1 <sup>-0.3%</sup>	87.1 <sup>-0.6%</sup>
MiniGPT-4-7B	Beam Search	29.0	8.9	72.9	76.9	76.9	77.9	78.0
	PAI	23.0	7.5	72.8	75.5	76.9	78.5	78.6
	Devils	20.2	6.9	71.9	67.2	74.3	79.3	79.5
	VISTA	15.4	4.6	67.4	76.3	77.2	—	—
	<b>HVAE(Ours)</b>	<b>20.0</b> <sup>-1.0%</sup>	<b>6.6</b> <sup>-4.3%</sup>	74.3	<b>77.1</b> <sup>+0.3%</sup>	<b>78.1</b> <sup>+1.2%</sup>	<b>79.7</b> <sup>+0.5%</sup>	<b>80.4</b> <sup>+1.1%</sup>
Shikra-7B	Beam Search	56.6	14.1	77.0	81.1	81.6	77.3	78.5
	PAI	35.4	9.2	77.1	<b>82.0</b>	81.3	77.4	77.3
	Devils	21.2	8.1	73.6	80.3	80.5	76.9	78.0
	VISTA	31.4	10.7	74.4	81.3	82.1	—	—
	<b>HVAE(Ours)</b>	<b>15.0</b> <sup>-29.2%</sup>	<b>3.4</b> <sup>-58.0%</sup>	72.2	81.7 <sup>-0.4%</sup>	<b>82.2</b> <sup>+0.1%</sup>	<b>77.5</b> <sup>+0.1%</sup>	<b>78.9</b> <sup>+0.5%</sup>

Table 12: Performance of **HVAE** against baselines using beam search decoding. Best results are in **bold**. Pink cells mark potentially unreliable CHAIR scores. Superscripts show the % change vs. the best baseline.

facilitate quantitative analysis. We conduct evaluations on the full MME dataset. To provide a holistic view of model performance, we report the aggregate scores for the two overarching categories: **Perception** and **Cognition**. This split evaluation allows us to verify that our intervention maintains robust performance across both fundamental visual recognition and high-level reasoning tasks.

**AMBER (Wang et al., 2023)**. AMBER is a comprehensive benchmark designed to evaluate multiple facets of hallucination, including object, attribute, and relation errors, across both discriminative and generative tasks. While its discriminative tasks are evaluated using standard metrics (e.g., Accuracy, F1 Score), its generative tasks employ a suite of four specific metrics to assess the quality and faithfulness of model responses. Let  $R_{\text{obj}}$  be the set of objects mentioned in the model’s response,  $G_{\text{obj}}$  be the set of ground-truth objects, and  $H_{\text{obj}}$  be a pre-annotated set of common human hallucinations. The generative metrics are defined as follows:

- **CHAIR**: Evaluates the proportion of hallucinated objects among all objects mentioned by the model. Note: This is equivalent to the instance-level CHAIR, and is referred to as CHAIR<sub>i</sub> in our main text.

$$\text{CHAIR} = 1 - \frac{|R_{\text{obj}} \cap G_{\text{obj}}|}{|R_{\text{obj}}|}. \quad (15)$$

- **Cover**: Measures the proportion of ground-truth objects that are correctly mentioned in the model’s response (i.e., object recall).

$$\text{Cover} = \frac{|R_{\text{obj}} \cap G_{\text{obj}}|}{|G_{\text{obj}}|}. \quad (16)$$

- **Hal**: A binary metric that indicates whether any hallucination occurred in the response.

$$\text{Hal} = \begin{cases} 1, & \text{if CHAIR} > 0 \\ 0, & \text{otherwise} \end{cases}. \quad (17)$$

- **Cog**: Assesses the similarity between the model’s hallucinations and those common to

Model	Method	CHAIR			POPE		POPE Chat	
		CHAIR <sub>s</sub> ↓	CHAIR <sub>i</sub> ↓	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
LLaVA-1.5-7B	Sample	48.2	15.2	73.8	83.2	84.0	85.1	83.1
	PAI	41.6	11.3	71.7	83.5	84.2	87.0	85.9
	Devils	31.8	7.1	79.9	83.7	<b>84.3</b>	<b>87.3</b>	<b>86.1</b>
	VISTA	16.0	7.9	65.7	82.6	84.0	—	—
	<b>HAVAE(Ours)</b>	<b>24.8</b> <sup>-22.0%</sup>	<b>5.5</b> <sup>-22.5%</sup>	77.2	<b>84.0</b> <sup>+0.4%</sup>	<b>84.3</b> <sup>+0.0%</sup>	87.0 <sup>-0.3%</sup>	<b>86.1</b> <sup>+0.0%</sup>
MiniGPT-4-7B	Sample	33.8	10.4	71.4	67.2	68.2	74.2	74.2
	PAI	28.4	12.2	69.1	65.9	68.9	75.4	75.4
	Devils	22.2	<b>7.7</b>	71.9	63.0	68.0	75.0	74.5
	VISTA	17.4	4.8	67.7	66.9	68.1	—	—
	<b>HAVAE(Ours)</b>	<b>22.0</b> <sup>-0.9%</sup>	<b>7.8</b> <sup>+1.3%</sup>	72.9	<b>67.4</b> <sup>+0.3%</sup>	<b>69.1</b> <sup>+0.3%</sup>	<b>75.7</b> <sup>+0.4%</sup>	<b>76.7</b> <sup>+1.7%</sup>
Shikra-7B	Sample	57.4	16.1	73.7	79.7	80.7	<b>75.7</b>	77.7
	PAI	41.6	11.4	72.4	80.1	80.2	75.6	76.7
	Devils	25.0	8.8	73.3	78.6	79.4	75.2	77.3
	VISTA	32.6	11.0	72.5	<b>80.2</b>	81.0	—	—
	<b>HAVAE(Ours)</b>	<b>18.2</b> <sup>-27.2%</sup>	<b>4.6</b> <sup>-47.7%</sup>	71.3	80.1 <sup>-0.1%</sup>	<b>81.1</b> <sup>+0.1%</sup>	75.5 <sup>-0.3%</sup>	<b>78.0</b> <sup>+0.4%</sup>

Table 13: Performance of **HAVAE** against baselines using sample as decoding strategy. Best results are in **bold**. Pink cells mark potentially unreliable CHAIR scores. Superscripts show the % change vs. the best baseline.

Model	Method	Generative				Discriminative		AMBER
		CHAIR <sub>i</sub> ↓	Cover ↑	Hal ↓	Cog* ↓	Acc. ↑	F1 ↑	Score ↑
LLaVA-1.5-7B	Greedy	6.0	50.6	27.4	2.8	74.8	77.6	85.8
	PAI	5.0	46.2	20.5	1.7	78.0	81.2	88.1
	Devils	3.8	46.0	20.7	<b>1.2</b>	77.8	81.3	88.8
	<b>HAVAE(Ours)</b>	<b>3.6</b> <sup>-5.3%</sup>	<b>51.7</b> <sup>+2.2%</sup>	<b>20.2</b> <sup>-1.5%</sup>	1.3 <sup>+8.3%</sup>	<b>78.6</b> <sup>+0.8%</sup>	<b>82.7</b> <sup>+1.7%</sup>	<b>89.6</b> <sup>+0.9%</sup>
MiniGPT-4-7B	Greedy	15.3	<b>63.3</b>	65.2	11.0	64.9	65.1	74.9
	PAI	12.3	60.8	51.3	7.2	61.4	61.3	74.5
	Devils	11.5	58.8	49.2	6.4	58.0	56.4	72.5
	<b>HAVAE(Ours)</b>	<b>11.2</b> <sup>-2.6%</sup>	61.1 <sup>-3.5%</sup>	<b>48.9</b> <sup>-0.6%</sup>	<b>6.1</b> <sup>-4.7%</sup>	<b>65.0</b> <sup>+0.2%</sup>	<b>65.2</b> <sup>+0.2%</sup>	<b>77.0</b> <sup>+2.8%</sup>
Shikra-7B	Greedy	11.2	<b>50.9</b>	49.7	5.6	78.5	82.1	85.5
	PAI	7.2	49.3	34.3	3.0	78.0	82.0	87.4
	Devils	6.7	45.3	22.5	1.6	71.1	74.1	83.7
	<b>HAVAE(Ours)</b>	<b>3.6</b> <sup>-46.3%</sup>	48.9 <sup>-3.9%</sup>	<b>17.8</b> <sup>-20.9%</sup>	<b>1.0</b> <sup>-37.5%</sup>	<b>78.6</b> <sup>+0.1%</sup>	<b>82.3</b> <sup>+0.2%</sup>	<b>89.4</b> <sup>+2.3%</sup>

Table 14: AMBER benchmark results on LLaVA-1.5-7B, MiniGPT-4-7B, and Shikra-7B. Best results are in **bold**. Superscripts show the % change vs. the best baseline.

humans.

$$\text{Cog} = \frac{|R_{\text{obj}} \cap H_{\text{obj}}|}{|R_{\text{obj}}|}. \quad (18)$$

Finally, to provide a single, unified measure of performance, AMBER also proposes the **AMBER Score**, which combines the F1 score from discriminative tasks and the CHAIR score from generative tasks:

$$\text{AMBER Score} = \frac{1}{2} \times (1 - \text{CHAIR}_i + \text{F1}). \quad (19)$$

The final reported scores are the average values of these metrics across all queries in the benchmark (Du et al., 2025, 2026).

## F Supplementary Experiments Results

### F.1 Performance with Alternative Decoding Strategies

In this subsection, we present additional experimental results for the beam search and nucleus sampling decoding strategies, which were omitted from the main text due to space constraints. For beam search, we set the beam size to 5, and for nucleus sampling, the temperature is set to 0.5. The detailed results are presented in Table 12 and Table 13, respectively. Overall, the findings are consistent with the conclusions drawn from the greedy decoding experiments in the main text: our method generally outperforms all baseline approaches across these different strategies.

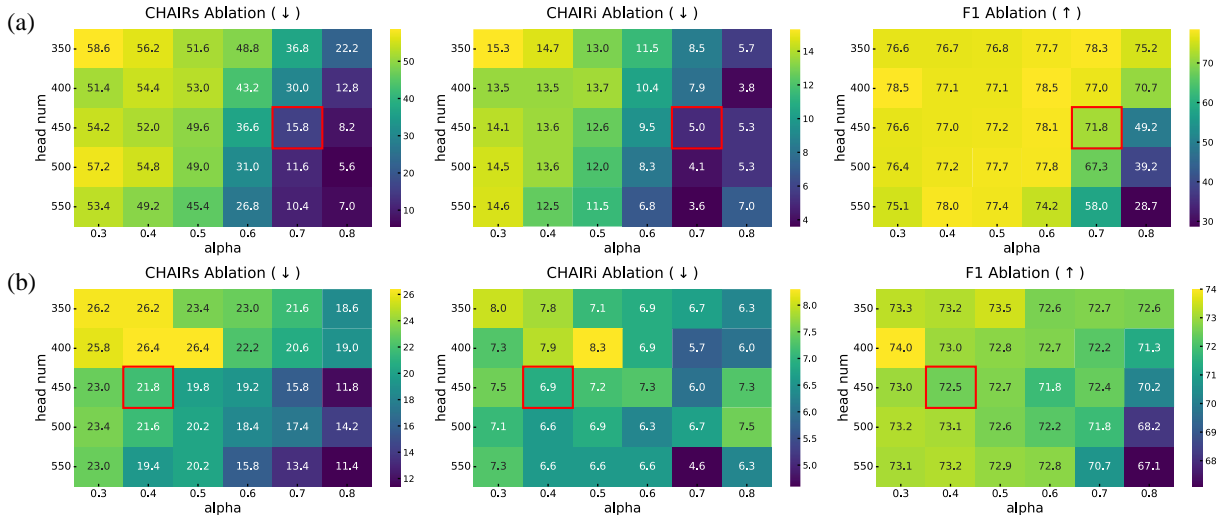


Figure 14: Ablation on hyperparameters  $\alpha$  and  $K$  for Shikra (a) and MiniGPT4 (b). Red boxes highlight the parameter combinations we used.

## F.2 AMBER Results on Additional Models

In this section, we present the comprehensive experimental results on the AMBER benchmark. As shown in Table 14, the findings align with the conclusions in the main text: **HVAE** consistently demonstrates clear superiority over baseline approaches, validating its robust generalization capabilities.

## F.3 Supplementary Ablation Results

In this subsection, we present the supplementary hyperparameter ablation studies for the remaining models, which were omitted from the main text due to space constraints. Specifically, Figure 14 illustrates the results for Shikra and MiniGPT4. These findings align with the trade-offs discussed in the main paper, further verifying the criticality of precise parameter tuning in long-context scenarios.

## G Supplementary Experiments and Analysis

### G.1 Qualitative Validation via Clustering and Visualization

In the main text, we utilized a composite quantitative metric (Hijacking Score) combined with IQR thresholding to identify **Hijacking Anchors**.

To corroborate the distinct properties of these Anchors, we perform an unsupervised K-Means clustering analysis on the visual tokens  $\mathcal{V}$ , using LLaVA-1.5 7B. Each visual token is represented in the feature space defined by (Dominance, Frequency, Attention). Subsequent to clustering, we assign each unique **vocabulary item** (Anchor) to a

specific cluster based on the majority distribution of its corresponding visual tokens.

The corresponding word cloud visualizations for each cluster—comprising the Anchors assigned to them—are presented in Figure 15. The results validate our approach by showing that the feature space naturally disentangles vocabulary items into distinct functional roles (Du et al., 2024).

#### G.1.1 High-Attention Manifold: Anchors vs. Content

The critical validation lies in the contrast between **Cluster 4** and **Cluster 2**. Both clusters are composed of Anchors associated with high-attention visual tokens, but they exhibit opposite representational stabilities.

**Cluster 4: Hijacking Anchors (Ours).** *Signature: [Highest Dom, High Freq, Highest Attn]. Anchors: kwiet, anch, ", фев.*

This cluster corresponds precisely to our identified **Hijacking Anchors**. The visual tokens mapped to these Anchors peak in both Dominance and Attention. This unique combination confirms their role as “**Active Sinks**”: they actively monopolize attention yet remain representationally rigid. Composed largely of meaningless artifacts and delimiters, these Anchors serve as fixed decoding endpoints rather than contextualized meanings.

**Cluster 2: Semantic & Spatial Concepts.** *Signature: [Low Dom, Avg Freq, High Attn]. Anchors: intersection, row, living, Airlines.*

While sharing the high-attention profile of Cluster 4, the visual tokens associated with these Anchors



Figure 15: **Word cloud visualization of token clusters identified via K-Means clustering on LLaVA-1.5 7B.** Tokens are grouped based on their behavioral signatures in the feature space (Dominance, Frequency, Attention). **Cluster 4 (Hijacking Anchors)** distinctively isolates the rigid, high-attention artifacts (e.g., kwiet, ") identified by our HABI method, separating them from the active semantic content in **Cluster 2** and the ignored environmental noise in **Cluster 5**. This qualitative separation confirms that our quantitative metrics align with the intrinsic semantic structure of the model’s vocabulary.

exhibit **Low Dominance**. This signature indicates **active processing**: the representations undergo significant shifts as the model extracts meaning. The presence of complex spatial and semantic concepts (intersection, row) confirms that the model correctly distinguishes meaningful content from rigid Hijacking Anchors in the feature space.

### G.1.2 High-Frequency Manifold: Noise vs. Substrate

This group differentiates frequent Anchors based on the attention levels of their corresponding visual tokens.

**Cluster 5: Ignored / Noise Artifacts.** *Signature: [High Dom, Highest Freq, Lowest Attn]. Anchors: trees, dust, olas.*

This cluster captures Anchors associated with environmental noise. The corresponding visual tokens exhibit High Frequency but Lowest Attention. Their High Dominance here reflects “**Pasive Stability**”—these tokens remain static because the model ignores them (bypassing computation), which is mechanically distinct from the “Active Rigidity” of Cluster 4.

**Cluster 0: Syntactic / Environmental Markers.** *Signature: [Low Dom, High Freq, Low Attn]. Anchors: ., and, in, &=\.*

Unlike Cluster 5, the corresponding visual tokens in this cluster show **Low Dominance**. This suggests that these tokens (decoding to syntactic markers like ., and, in) actively participate in local structure formation (“syntactic glue”) without capturing the global attention reserved for Hijacking Anchors or semantic targets.

### G.1.3 Long-Tail & Generic Groups

The final category encompasses the broad spectrum of standard semantic content. It distinguishes between specific, context-dependent entities and generic background elements based on their frequency and specificity.

**Cluster 1: Long-Tail Entities.** *Signature: [Lowest Dom, Lowest Freq, Avg Attn]. Anchors: EU, Portuguese, Flags.*

These Anchors represent rare entities and proper nouns. The **Lowest Dominance** of their corresponding visual tokens indicates **extensive semantic evolution**. Unlike common concepts, these rare entities require significant layer-wise refinement

Method	CHAIR			POPE	
	CHAIR <sub>s</sub> ↓	CHAIR <sub>i</sub> ↓	F1 ↑	Acc. ↑	F1 ↑
HVAE	18.2	3.7	76.7	86.1	86.2
w/o GT	18.6	5.0	78.1	85.9	86.0

Table 15: Ablation study on the necessity of ground-truth (GT) annotations. “w/o GT” denotes the variant where head selection relies solely on generated objects. Results confirm HVAE’s robustness to the absence of external supervision.

and context aggregation to be correctly resolved, preventing their representations from stabilizing early.

**Cluster 3: Generic Background Concepts.** *Signature: [Low Dom, Avg Freq, Low Attn]. Anchors: on, table, plate, way.*

These Anchors represent the statistical median. While containing common visual objects (table, plate) and prepositions (on), the Low Attention of their corresponding visual tokens suggests they often serve as common background elements rather than primary attentional focal points.

## G.2 Robustness to the Absence of Ground-Truth Annotations

To validate the applicability of our method in scenarios lacking ground-truth (GT) object annotations, we conducted a comparative ablation study. Specifically, we modified the head selection criterion in equation (10) by substituting the ground-truth object set  $\mathcal{O}_{\text{real}}$  with the entire set of generated objects (i.e., the union  $\mathcal{O}_{\text{real}} \cup \mathcal{O}_{\text{hal}}$ ). For this setting, we adjusted the hyperparameters to  $\alpha = 0.5$  and the number of selected heads to 550, while keeping all other configurations unchanged. The comparative results on the CHAIR and POPE benchmarks are presented in Table 15. The empirical evidence demonstrates that HVAE maintains superior performance even without reliance on external ground-truth annotations, thereby confirming its robustness and potential for broader scalability.

## H Additional Case Studies

In this section, we present several additional case studies to qualitatively demonstrate how our HVAE method concentrates attention on key visual content and effectively mitigates hallucination. Figure 16 visualizes four distinct examples. For each example, we contrast the attention map for a hallucinated object token generated by the baseline

model with the attention map for a corresponding real object token from our HVAE-enhanced model. It is important to note a key methodological difference in these visualizations: the baseline maps represent the average attention across **all** heads, whereas the maps for our method show the average attention of the **450 heads selected by HVAE**.

Additionally, for three of these examples, we provide logit lens visualizations in Figure 17 (following the approach in Figure 2) to further illustrate the dynamics of **Vocabulary Hijacking**. These visualizations confirm that the model allocates disproportionately high attention to specific **Hijacking Anchors** when a hallucination is produced.

From these examples, we can clearly observe a consistent pattern: when a hallucination occurs, the baseline model’s attention is often hijacked by irrelevant **Inert Tokens**. The effect of HVAE is visually evident, as it successfully refocuses the model’s attention onto the grounded target object.

### H.1 Qualitative Error Analysis

While our method successfully mitigates hallucinations driven by Inert Tokens, we provide a qualitative error analysis to discuss a limitation of our current observational framework.

Revisiting the third case presented in Figure 17, we observe an intriguing anomaly: the 321<sup>st</sup> visual token captures an anomalously high proportion of the attention weights, despite being completely semantically irrelevant to the currently generated text token. Interestingly, however, this specific token does *not* exhibit the characteristic Vocabulary Hijacking phenomenon.

This counter-example indicates that while Vocabulary Hijacking is a highly prevalent and actionable indicator of LLM hallucinations, it cannot deterministically localize *all* instances of attention misallocation. There remain atypical failure modes where the model assigns incorrect attention without triggering the semantic collapse associated with Inert Tokens, suggesting that complementary mechanisms may be required to comprehensively resolve these diverse attention errors in future work.

## I Algorithms

For clarity and to facilitate reproducibility, this section provides the detailed algorithmic procedures for our proposed frameworks. We present the **Hijacking Anchor-Based Identification (HABI)**

framework in Algorithm 1 and the **Hijacking-Aware Visual Attention Enhancement (HVAE)** framework in Algorithm 2.

Please help me describe the image in detail.



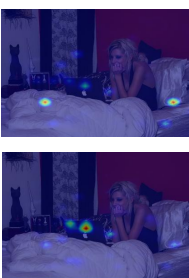
LLaVA-1.5 7B

The image features a small TV sitting on a wooden desk in a room. The TV is positioned towards the right side of the desk, and it appears to be turned off. The desk is also equipped with a remote control, which is placed near the left side of the TV. In addition to the TV and remote, there are two books on the desk, one located closer to the left side and the other towards the right side. A **person** can be seen in the background, partially visible on the left side of the image.

HAVAe (ours)

The image features a small TV sitting on a wooden desk in a room. The **TV** is placed on a silver stand, and it is situated next to a window. The television is turned off, and the room appears to be dimly lit. There are several remote controls on the desk. The room has a brown table, and the television is the main focus of the scene.

Please help me describe the image in detail.



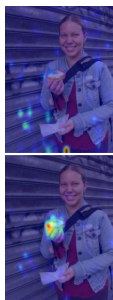
LLaVA-1.5 7B

The image features a woman sitting on a bed in a bedroom. She is wearing a black dress and appears to be looking at something on her laptop, which is placed on her lap. The bed is covered with a white comforter, and there are two pillows on it. In the room, there are several items scattered around. A cat is present, located near the left side of the bed. A cell **phone** can be seen on the bed, closer to the center. Additionally, there are two bottles, one near the left edge of the bed and the other on the right side. A vase is also visible, placed near the left edge of the bed.

HAVAe (ours)

The image shows a woman sitting on a bed with a **laptop** computer in front of her. She is wearing a black shirt and is looking at the camera. The bed is covered with a white comforter, and the woman is sitting on a white bed. The room is decorated with a red wall and a black and white curtain. The woman is looking at the camera while using her laptop.

Please help me describe the image in detail.



LLaVA-1.5 7B

The image features a woman standing outside, holding a donut in her hand. The donut is decorated with red, white, and blue colors, making it visually appealing. The woman is wearing a red shirt and appears to be smiling as she poses with the donut. In the background, there is a backpack placed on the **ground**, possibly belonging to the woman. The scene seems to be set in an urban environment, with a building visible in the background.

HAVAe (ours)

The image features a woman wearing a red shirt and holding a **donut** with red, white, and blue frosting. She is smiling and holding the donut in front of her face. The woman is wearing a backpack, and she is standing in front of a building. The scene is set in a city, and the woman is holding a donut in her hand.

Please help me describe the image in detail.



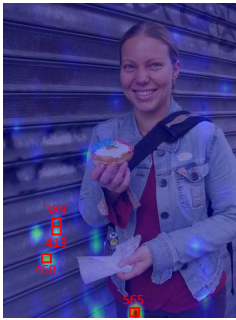
LLaVA-1.5 7B

The image features a group of three giraffes standing together in a dirt area. They are positioned close to each other, with one giraffe on the left, another in the middle, and the third on the right. The giraffes are standing under a tree, which provides shade and a natural environment for them. The scene also includes a **bench** located in the background, possibly for visitors to sit and observe the giraffes. The overall atmosphere of the image is serene and peaceful, showcasing the beauty of these majestic animals in their natural habitat.

HAVAe (ours)

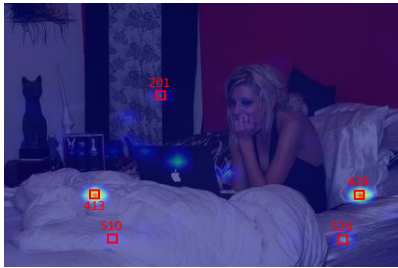
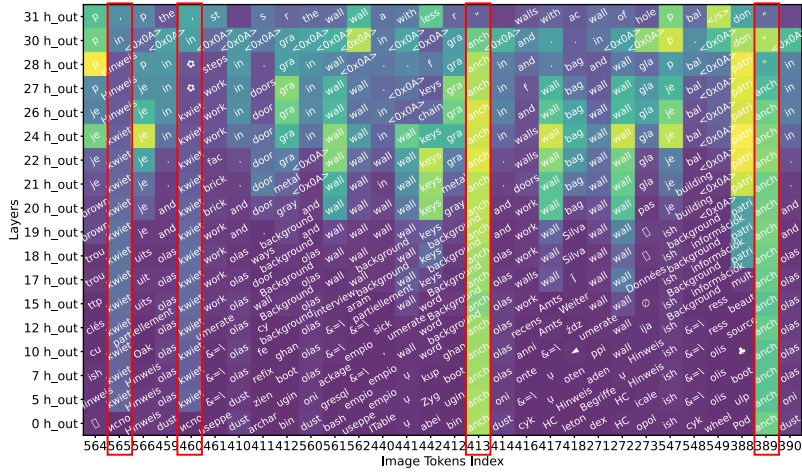
The image features a group of three **giraffes** standing together in a zoo enclosure. They are surrounded by trees and plants, with two of the giraffes eating leaves from a tree. The giraffes are standing in a dirt area, which is their natural habitat. The giraffes are positioned in a way that they are facing each other, creating a sense of unity among them. The scene is set in a forest-like environment, with the giraffes standing near a tree and a fence.

Figure 16: Additional qualitative comparison of attention maps. For each case, we contrast the attention map for a hallucinated token from the baseline model (top row) with a corresponding real object token from our HAVAe-enhanced model (bottom row), demonstrating HAVAe’s ability to refocus attention on salient objects.



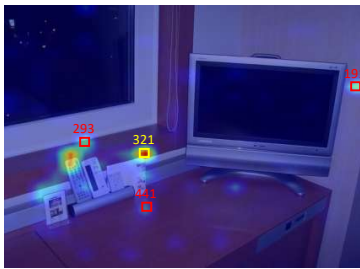
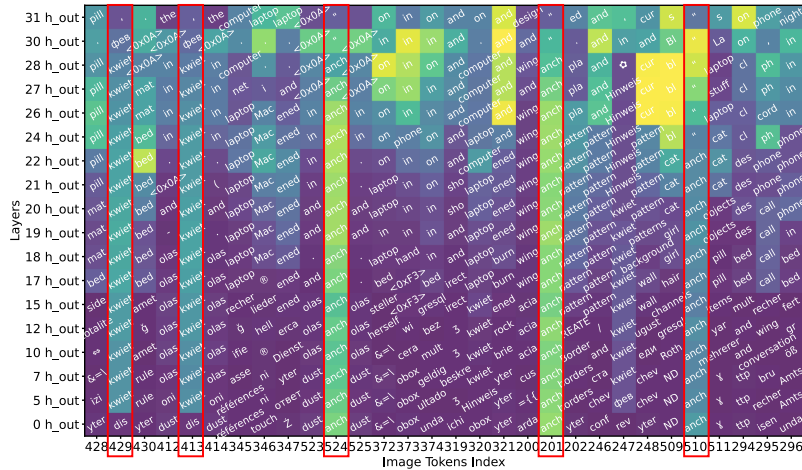
Please help me describe the image in detail.

... In the background, there is a backpack placed on the **ground**, possibly belonging to the woman...



Please help me describe the image in detail.

... A cell **phone** can be seen on the bed, closer to the center. Additionally, there are two bottles...



Please help me describe the image in detail.

... A **person** can be seen in the background, partially visible on the left side of the image.

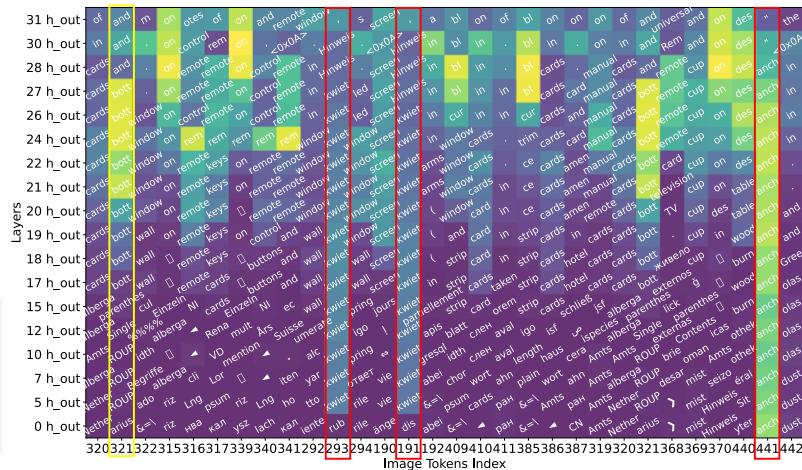


Figure 17: Logit lens visualization for three case studies from Appendix H. These plots provide further qualitative evidence that the model allocates disproportionately high attention to the identified **Inert Tokens** during hallucination.

---

**Algorithm 1** Hijacking Anchor-Based Identification (HABI)

---

**Require:** LVLm model  $M$ ; Calibration dataset  $\mathcal{D}$  (e.g., 500 COCO images).

**Ensure:** The set of Inert Tokens  $\mathcal{I}_{\text{inert}}$  for a given input image’s visual tokens  $\mathcal{I}_v$ .

```
1:
2: /* Phase 1: Global Calibration & Anchor Discovery (Corresponds to Sec. 3.2 & 3.3) */
3:  $\mathcal{V} \leftarrow \emptyset$ ,  $AnchorMap \leftarrow \{\}$ ,  $AllScores \leftarrow []$ 
4: for each image  $I \in \mathcal{D}$  do
5:    $\mathcal{I}_v \leftarrow \text{GetVisionTokens}(M, I)$ 
6:   for each vision token  $v_i \in \mathcal{I}_v$  do
7:      $\mathcal{T}_{v_i} \leftarrow \text{LogitLensDecode}(M, v_i)$  {Get Trace}
8:      $a_{v_i} \leftarrow \text{MostFrequentToken}(\mathcal{T}_{v_i})$  {Identify Trace Anchor}
9:     Calculate metrics: Dominance  $\mathcal{D}(v_i)$ , Frequency  $\mathcal{F}(v_i)$ , Attention  $\mathcal{A}(v_i)$ 
10:     $S_{\text{hijack}}(v_i) \leftarrow \mathcal{D}(v_i) \cdot \mathcal{F}(v_i) \cdot \mathcal{A}(v_i)$ 
11:    Append( $AllScores$ ,  $S_{\text{hijack}}(v_i)$ ) {Collect individual scores}
12:    Append( $AnchorMap[a_{v_i}]$ ,  $S_{\text{hijack}}(v_i)$ ) {Group scores by anchor word}
13:     $\mathcal{V} \leftarrow \mathcal{V} \cup \{v_i\}$ 
14:   end for
15: end for
16:
17: /* Phase 2: Threshold Determination (Corresponds to Sec. 3.3 & HABI) */
18: /* 2.1 Identify Hijacking Anchors */
19:  $\tau_s \leftarrow Q_3(AllScores) + 1.5 \cdot \text{IQR}(AllScores)$  {Outlier threshold based on global score distribution}
20:  $\mathcal{A}_{\text{hijack}} \leftarrow \emptyset$ 
21: for each unique anchor  $w \in \text{Keys}(AnchorMap)$  do
22:   if  $\text{Mean}(AnchorMap[w]) > \tau_s$  then
23:      $\mathcal{A}_{\text{hijack}} \leftarrow \mathcal{A}_{\text{hijack}} \cup \{w\}$ 
24:   end if
25: end for
26:
27: /* 2.2 Determine Hijacking Ratio Threshold using Otsu’s Method */
28:  $Ratios \leftarrow []$ 
29: for each  $v_i \in \mathcal{V}$  if  $v_i \in \text{Top5\%Attention}$  do
30:    $r_{\text{hijack}}(v_i) \leftarrow \frac{1}{|\mathcal{T}_{v_i}|} \sum_{t \in \mathcal{T}_{v_i}} \mathbb{I}(t \in \mathcal{A}_{\text{hijack}})$ 
31:   Append( $Ratios$ ,  $r_{\text{hijack}}(v_i)$ )
32: end for
33:  $\tau_r \leftarrow \text{OtsuMethod}(Ratios)$  {Find separation threshold in bimodal distribution}
34:
35: /* Phase 3: HABI Inference Function (Corresponds to HABI Implementation) */
36: function IdentifyInertTokens( $\mathcal{I}_v$ ,  $M$ ,  $\mathcal{A}_{\text{hijack}}$ ,  $\tau_r$ )
37:  $\mathcal{I}_{\text{inert}} \leftarrow \emptyset$ 
38: for each vision token  $v_i \in \mathcal{I}_v$  do
39:    $\mathcal{T}_{v_i} \leftarrow \text{LogitLensDecode}(M, v_i)$ 
40:    $r_{\text{hijack}}(v_i) \leftarrow \frac{1}{|\mathcal{T}_{v_i}|} \sum_{t \in \mathcal{T}_{v_i}} \mathbb{I}(t \in \mathcal{A}_{\text{hijack}})$  {Calculate Hijacking Ratio}
41:   if  $r_{\text{hijack}}(v_i) > \tau_r$  then
42:      $\mathcal{I}_{\text{inert}} \leftarrow \mathcal{I}_{\text{inert}} \cup \{v_i\}$ 
43:   end if
44: end for
45: return  $\mathcal{I}_{\text{inert}}$ 
46: end function
```

---

---

**Algorithm 2** Hijacking-Aware Visual Attention Enhancement (HVAE) Framework

---

**Require:** LVLm model  $M$ ; Calibration dataset  $\mathcal{D}$ ; HABI function ‘IdentifyInertTokens’ (from Alg. 1); Pre-computed Hijacking Anchors  $\mathcal{A}_{\text{hijack}}$ ,  $\tau_r$  (HABI threshold); Hyperparameters  $K$  (num heads),  $\alpha$  (strength).

**Ensure:** A modified inference process with mitigated hallucination.

```
1:
2: /* Stage 1: Principled Head Selection (Offline) */
3: function SelectCriticalHeads( $M, \mathcal{D}, K$ )
4:    $\mathcal{O}_{\text{real}} \leftarrow \text{CollectRealObjectTokens}(M, \mathcal{D})$  {Correspond to Section 3.2}
5:    $\overline{\text{NHAR}} \leftarrow \text{InitializeMatrix}(L, H, \text{zeros})$ 
6:   for each real object token  $y_k \in \mathcal{O}_{\text{real}}$  do
7:      $\mathcal{I}_v \leftarrow \text{GetCorrespondingVisualTokens}(y_k)$ 
8:      $\mathcal{I} \leftarrow \text{GetFullContext}(y_k)$ 
9:      $\mathcal{I}_{\text{inert}} \leftarrow \text{IdentifyInertTokens}(\mathcal{I}_v, M, \mathcal{A}_{\text{hijack}}, \tau_r)$  {Call HABI function}
10:     $A_{\text{all\_heads}} \leftarrow \text{GetAttentionForAllHeads}(M, y_k)$ 
11:    for each head  $(\ell, h)$  from  $(1, 1)$  to  $(L, H)$  do
12:      /* Calculate NHAR for this step */
13:       $A_{\ell, h} \leftarrow A_{\text{all\_heads}}[\ell, h]$ 
14:       $\text{Attn}_{\text{valid\_visual}} \leftarrow \sum_{v_i \in \mathcal{I}_v \setminus \mathcal{I}_{\text{inert}}} A_{\ell, h, i}$ 
15:       $\text{Attn}_{\text{total}} \leftarrow \sum_{v_i \in \mathcal{I}} A_{\ell, h, i}$ 
16:       $\text{NHAR\_score} \leftarrow \text{Attn}_{\text{valid\_visual}} / \text{Attn}_{\text{total}}$ 
17:       $\overline{\text{NHAR}}_{\ell, h} \leftarrow \overline{\text{NHAR}}_{\ell, h} + \text{NHAR\_score}$ 
18:    end for
19:  end for
20:   $\overline{\text{NHAR}} \leftarrow \overline{\text{NHAR}} / |\mathcal{O}_{\text{real}}|$  {Eq. 10}
21:   $\mathcal{H}_{\text{target}} \leftarrow \text{TopK}(\overline{\text{NHAR}}, K)$  {Select critical heads}
22:  return  $\mathcal{H}_{\text{target}}$ 
23: end function
24:
25: /* Stage 2: Collective Attention Reinforcement (Online) */
26: function HVAE_Attention_Forward( $A_{\text{original}}, \mathcal{H}_{\text{target}}, \alpha$ )
27: {Injects this logic into the model’s attention mechanism during inference.}
28:  $L, H \leftarrow \text{GetModelDimensions}()$ 
29:  $A_{\text{enhanced}} \leftarrow A_{\text{original}}$ 
30: for  $\ell = 1$  to  $L$  do
31:   /* Calculate Layer-wise Mean Attention (Eq. 11) */
32:    $\text{MeanAttn}_{\ell} \leftarrow \text{Zeros}(\text{size} = \text{num\_visual\_tokens})$ 
33:   for  $h' = 1$  to  $H$  do
34:      $\text{MeanAttn}_{\ell} \leftarrow \text{MeanAttn}_{\ell} + |A_{\text{original}}[\ell, h', \text{visual\_indices}]|$ 
35:   end for
36:    $\text{MeanAttn}_{\ell} \leftarrow \text{MeanAttn}_{\ell} / H$ 
37:   /* Apply reinforcement to target heads */
38:   for  $h = 1$  to  $H$  do
39:     if  $(\ell, h) \in \mathcal{H}_{\text{target}}$  then
40:        $A_{\text{enhanced}}[\ell, h, \text{visual\_indices}] \leftarrow A_{\text{original}}[\ell, h, \text{visual\_indices}] + \alpha \cdot \text{MeanAttn}_{\ell}$ 
41:     end if
42:   end for
43: end for
44: return  $A_{\text{enhanced}}$ 
45: end function
```

---