VISUAL SIGNAL INTEGRITY FOR DETECTING OBJECT HALLUCINATIONS IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) demonstrate remarkable content generation capabilities but are prone to object hallucinations, *i.e.*, generating contextually plausible yet visually ungrounded descriptions. This imposes critical reliability risks in multimodal understanding. Detecting such hallucinations remains challenging because they often arise from breakdowns in visual information preservation that are not captured by existing methods, which inspect only isolated points along the visual-to-language pathway. In this paper, we reveal that grounded object tokens exhibit higher semantic cohesion among attended visual tokens and stronger visual contributions in the residual stream than hallucinated ones. Inspired by these observations, we propose a training-free diagnostic framework, *Visual Signal Integrity* (VSI), to quantify the integrity of visual grounding during generation. Specifically, VSI introduces two complementary statistics: *Visual Semantic Cohesion* (VSC), which measures the geometric concentration of attended visual tokens to assess whether coherent visual evidence is formed, and *Visual Steering Ratio* (VSR), which measures the relative dominance of visual over textual contributions in the residual stream. By combining these two statistics, VSI provides a unified score for detecting breakdowns in visual information flow. Extensive experiments on multiple hallucination benchmarks and diverse LVLMs show that VSI delivers a reliable and interpretable solution for hallucination detection.

1 INTRODUCTION

The rapid advancement of Large Vision-Language Models (LVLMs) (Liu et al., 2024; Wang et al., 2024; Bai et al., 2025b) has led to their widespread deployment in applications ranging from autonomous scene understanding (Fu et al., 2023) to multimodal content creation (Wang et al., 2026; Zhou et al., 2025). Despite their remarkable generative capabilities, these models remain prone to object hallucinations (Rohrbach et al., 2018)—the generation of contextually plausible but visually ungrounded entities. This behavior imposes critical reliability risks, limiting the trustworthy deployment of LVLMs in real-world scenarios. Therefore, developing effective hallucination detection methods has become a pressing challenge to ensure the safe and dependable use of multimodal systems.

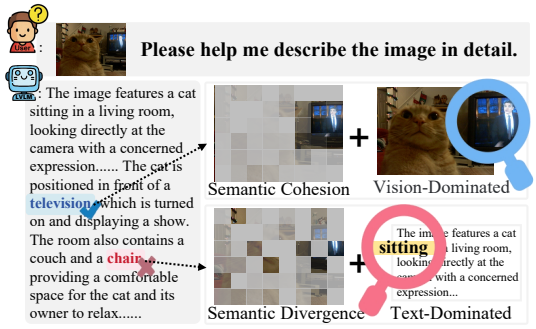


Figure 1: Visual information flow breakdown in LVLM hallucinations: failure to form cohesive visual semantics or being dominated by textual priors.

Existing LVLM hallucination detection methods primarily rely on output confidence (Malinin & Gales, 2021; Zhou et al., 2024; Jiang et al., 2025a), attention magnitude over visual tokens (Jiang et al., 2025b), or cross-modal representation alignment (Phukan et al., 2025; Park & Li, 2025). However, these approaches inspect isolated points along the visual-to-language pathway rather than tracking whether visual information is faithfully preserved from visual representation to decoding generation. Thus, they are unreliable when intact visual representations are overridden by strong language priors during decoding. We posit that hallucinations emerge when this preservation fails:

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

visual information either fails to form a concentrated visual concept or fails to maintain dominance over competing textual priors during decoding. This raises a central question:

Can we detect hallucinations by identifying the collapse of visual information flow?

Two challenges make this question non-trivial. 1) *Quantifying whether a visual concept has formed* is difficult: visual representations are distributed across many high-dimensional tokens, and a scalar confidence score cannot reveal whether they converge toward a coherent semantic direction. 2) *Disentangling modality-specific contributions* is similarly challenging: within the deep residual stream of a Transformer, visual and textual signals are deeply interwoven, making it hard to isolate which modality dominates at each decoding step.

In this paper, we propose **Visual Signal Integrity (VSI)**, a training-free diagnostic method built upon two key findings that address the above challenges. First, we observe that visual tokens tied to a correctly grounded object converge toward a shared semantic direction across layers, while hallucinated ones scatter across divergent directions (Observation 1). Inspired by this, we introduce *Visual Semantic Cohesion (VSC)*, which measures the attention-weighted similarity among attended visual tokens; this definition is equivalent to computing a single vector norm, reducing pairwise comparisons to linear-time aggregation (Proposition 1). Second, we observe that under induced language priors, visual contributions to the residual stream are systematically suppressed while textual contributions increase, causing the model to follow the prior rather than visual evidence (Observation 2). Inspired by this, we introduce *Visual Steering Ratio (VSR)*, which decomposes the residual update into visual and textual contributions and measures their relative magnitude, revealing which modality dominates decoding (Proposition 2). Together, VSC and VSR form VSI, which jointly evaluates whether visual information is coherently formed and preserved throughout generation.

Figure 1 illustrates how these two statistics operate on a concrete example: the grounded entity "television" is supported by visually cohesive tokens and vision-dominated decoding, while the hallucinated "chair" corresponds to scattered visual evidence and text-dominated decoding. By identifying where the visual information flow collapses, whether at concept formation or at the decoding competition stage, VSI provides not just a score, but a diagnostic view of the model's grounding process. Extensive experiments on MSCOCO (Lin et al., 2014), Objects365 (Shao et al., 2019), and POPE (Li et al., 2023b) across eight LVLMs show that VSI outperforms baselines on most model–dataset pairs while delivering interpretable diagnostics beyond a binary hallucination label. Our contributions are summarized as:

- We reframe hallucination in LVLMs as a failure of visual information preservation and propose Visual Signal Integrity (VSI), a training-free framework that jointly assesses whether visual evidence is coherently formed and preserved throughout generation. Unlike prior methods that inspect isolated points along the visual-to-language pathway, VSI tracks whether visual information is faithfully preserved and identifies where the flow collapses.
- We introduce two mechanistically grounded and interpretable statistics. Visual Semantic Cohesion (VSC) measures the geometric concentration of attended visual tokens, and Visual Steering Ratio (VSR) quantifies the relative dominance of visual over textual contributions in the residual stream. Both statistics are derived from empirical observations (Observations 1 and 2) and are theoretically grounded: VSC is shown to be equivalent to an efficient norm computation (Proposition 1), and VSR is built upon an exact token-wise residual decomposition (Proposition 2).
- Extensive experiments across eight LVLMs and three benchmarks (MSCOCO, Objects365, POPE) demonstrate that VSI outperforms baselines on most model–dataset pairs. Beyond a single hallucination score, VSI separates failures of visual concept formation from failures where language priors override intact visual evidence. This diagnostic granularity offers insights beyond binary labels, making VSI both a detector and an interpretability tool for multimodal analysis.

2 Related Work

Object Hallucination Detection. Object hallucination denotes the generation of objects that are absent from the input image but remain linguistically plausible. Detecting such errors is central to assessing the reliability of large vision-language models (LVLMs). Existing methods can be broadly categorized by the evidence they use for verification. Early evaluation protocols such as

CHAIR (Rohrbach et al., 2018) rely on ground-truth annotations, while subsequent approaches employ LVLMs judges (Janiak et al., 2025; Guan et al., 2024) or external detectors and scene graphs to verify generated claims. Although effective, these methods typically require multi-stage pipelines. Another line of work estimates hallucination likelihood from output probabilities, using metrics such as entropy or negative log-likelihood, as in LURE (Zhou et al., 2024). These methods are simple and applicable, but they conflate linguistic uncertainty with visual faithfulness, and may fail when hallucinated objects are assigned high confidence under strong language priors.

Recent studies therefore turn to internal signals for hallucination detection. Representation-based methods, including GLSim (Park & Li, 2025), ContextLens (Phukan et al., 2025), and Internal Confidence (IC) (Jiang et al., 2025a), probe intermediate visual or image-conditioned hidden states to assess cross-modal alignment. However, such alignment-level evidence does not explicitly reveal whether visual information is propagated to and dominates the final generation process. Attention-based methods provide a complementary perspective: SVAR (Jiang et al., 2025b) uses image-token attention as a grounding proxy, while OPERA (Huang et al., 2024) exploits self-attention patterns to identify hallucination-related anchor tokens for mitigation. Nevertheless, attention magnitude alone neither ensures the semantic consistency of attended visual content nor establishes that visual evidence ultimately governs the model's prediction.

3 MOTIVATIONS

Existing detection methods probe confidence, attention, or intermediate representations, but do not track whether visual evidence remains preserved and influential throughout generation. We therefore posit that hallucinations arise from breakdowns in visual information preservation. For visual information to ground a generated object, it needs to satisfy two conditions. 1) object-relevant visual representations form a coherent semantic direction, and 2) the visual evidence remains dominant over competing language priors during decoding. We examine these two conditions empirically below.

Observation 1. *Correctly grounded objects exhibit concentrated visual semantics in the hidden space; hallucinated ones exhibit scattered semantics across divergent directions.*

To test whether a visual concept has truly formed, we run a simple probe on 500 images annotated with ground-truth bounding boxes. Using LLaVA-1.5-7B, we extract hidden states of all visual tokens across all layers. For each image, we compute pairwise cosine similarities for two sets of tokens: those inside the object bounding box (object-relevant tokens) and an equal number randomly sampled from the whole image (global background tokens). More results are provided in Appendix D.3.

As shown in Figure 2, object-relevant tokens show higher similarity than random tokens across layers, indicating more concentrated visual representations. This motivates *Visual Semantic Cohesion* (VSC), which measures the geometric concentration of attended visual tokens (Section 4.1).

Observation 2. *Strong language priors suppress visual contributions in the residual stream, causing visual evidence to be abandoned even when it has been correctly formed.*

Even with intact visual representations, hallucinations can still arise during decoding. When the autoregressive language prior is strong enough, the model may choose to "believe its training experience" over what it sees. To verify this, we conduct an intervention experiment on the POPE benchmark (Li et al., 2023b). POPE is a binary yes/no question-answering task (*e.g.*, "Is there a chair in the image?"), which allows us to precisely manipulate the language prior. We create three conditions by prepending different prefixes to the original instruction:

> **Neutral:** Please help me answer this simple visual question: [original instruction]
> **Yes-Induce:** Since the image actually does contain this object: [original instruction]
> **No-Induce:** Since the image does not contain this object: [original instruction]

We track two quantities: (1) the frequency of "Yes" answers, to confirm that the language bias is successfully induced; and (2) the contribution norms from visual tokens ($\|\Delta \mathbf{h}_{\text{img}}\|_2$) and textual tokens ($\|\Delta \mathbf{h}_{\text{txt}}\|_2$) to the residual stream, which measure the actual influence of each modality during decoding. The exact decomposition is detailed in Section 4.2.

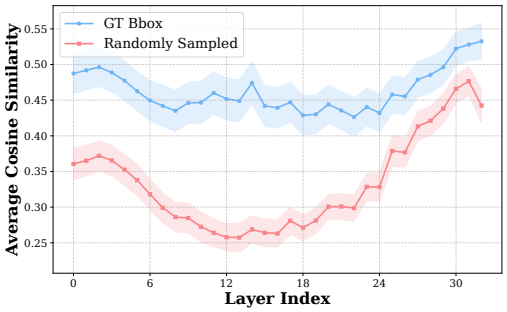


Figure 2: Motivation for S_{VSC}. Object-relevant visual tokens exhibit higher semantic similarity than randomly sampled ones, signaling that correctly grounded objects form concentrated visual representations.

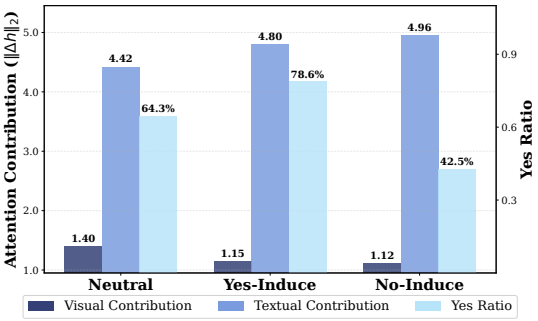


Figure 3: Motivation for S_{VSR}. Under both Yes and No induced prefixes, visual contributions are suppressed compared to the neutral prompt, while textual contributions increase and override visual evidence during decoding.

As shown in Figure 3, the results reveal a clear imbalance. Under the Yes-Induce prefix, the model's "Yes" frequency rises sharply from 64.3% to 78.6%, while the textual contribution norm increases to 4.80 and the visual contribution norm drops to 1.15. The No-Induce prefix produces the opposite pattern: "Yes" frequency falls to 42.5%, yet visual contributions remain similarly suppressed (1.12), confirming that textual priors, rather than visual evidence, dictate the model's decisions. This reveals how language priors dominate the residual stream: the textual signal grows loud enough to drown out the visual signal, causing the model to abandon visual facts. This gives us our second diagnostic clue: the mere presence of visual evidence is not enough; what matters is whether it prevails against the textual prior. It directly motivates *Visual Steering Ratio* (VSR, as in Section 4.2).

Together, these two observations provide us with a complete diagnostic logic: a *reliable* generation requires that the visual information flow satisfies two conditions, *i.e.*, the formation of concentrated visual representations (high VSC), and the ability to maintain dominance over competing textual priors during decoding (high VSR). A failure at either stage signals a likely hallucination. This leads directly to our Visual Signal Integrity (VSI, as in Section 4.3) in the next section, which combines VSC and VSR into a unified score for identifying where the visual information flow collapses.

4 Methodology

Method Overview. Fluent descriptions can mask unsupported objects (Rohrbach et al., 2018), while the interaction of visual evidence, textual context, and learned priors makes it difficult to locate where visual grounding fails, as evidenced by Observation 2. Following Section 3, we therefore examine whether visual information is coherently formed and preserved throughout generation. We propose **Visual Signal Integrity (VSI)**, a training-free framework for diagnosing hallucinations through visual information flow. VSI consists of two complementary statistics: **Visual Semantic Cohesion** (**VSC**, Section 4.1), which measures the coherence of attended visual tokens, and **Visual Steering Ratio** (**VSR**, Section 4.2), which measures the dominance of visual over textual residual contributions. VSC admits an equivalent linear-time formulation (Proposition 1), while VSR builds on an exact token-wise residual decomposition (Proposition 2). Their combination yields the unified VSI score (Section 4.3; Figure 4).

4.1 Visual Semantic Cohesion

We first measure whether the model forms a coherent visual signal for the generated object. When an object is clearly perceived, the visual tokens actively attended by the model are expected to converge toward a shared semantic direction. Conversely, if the visual evidence is ambiguous or the model fails to extract relevant object-level features, the attended representations become geometrically scattered.

Formally, let $\mathcal{V}$ denote the set of all visual tokens and q denote the currently generated token. For each visual token $i \in \mathcal{V}$, let $\mathbf{h}_i^l$ be its corresponding hidden state at layer l. To capture a comprehensive

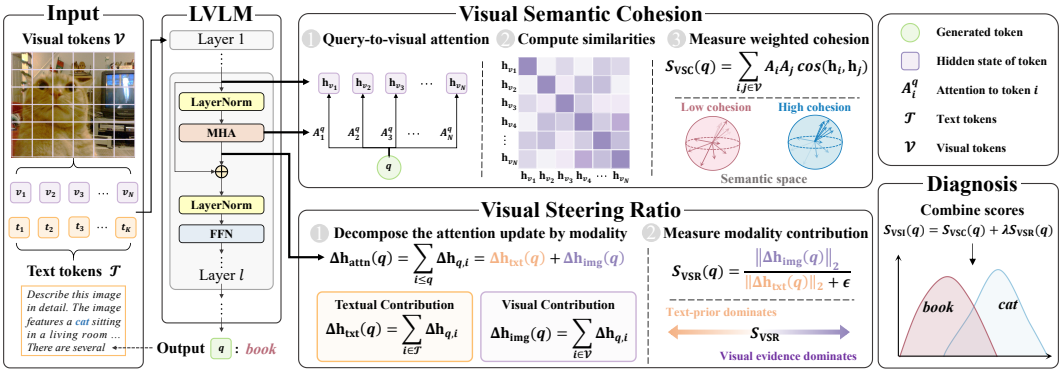


Figure 4: Overview of Visual Signal Integrity. For each generated object token, S_{VSC} measures the semantic cohesion of attended visual tokens, and S_{VSR} measures the dominance of visual over textual residual contributions. The combined S_{VSI} enables a fine-grained diagnosis of hallucination.

grounding signal, we aggregate the attention patterns across all H attention heads as follows:

$$A^l_{q,i} = \frac{1}{H}\sum_{h=1}^{H} \alpha^{l,h}_{q,i}, \quad \alpha^{l,h}_{q,i} = \text{softmax}\left(\frac{(\mathbf{h}^{l-1}_q \mathbf{W}^{l,h}_Q)(\mathbf{h}^{l-1}_i \mathbf{W}^{l,h}_K)^\top}{\sqrt{D_k}}\right), \tag{1}$$

where $\mathbf{W}^{l,h}_Q, \mathbf{W}^{l,h}_K \in \mathbb{R}^{D\times D_k}$ are the query and key projection matrices, respectively. $\alpha^{l,h}_{q,i}$ is the attention weight from the generated text token q to the visual token i, D is the hidden dimension of the model and D_k is the per-head dimension, typically $D_k = D/H$.

Following Observation 1 to measure geometric concentration, we define **Visual Semantic Cohesion**, S^l_{VSC}, as the attention-weighted sum of pairwise cosine similarities across all visual tokens:

$$S^l_{\text{VSC}}(q) := \sum_{i,j\in\mathcal{V}} A^l_{q,i} A^l_{q,j} \cos(\mathbf{h}^l_i, \mathbf{h}^l_j), \tag{2}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity between two vectors. By Proposition 1, this pairwise similarity can be rewritten as $S^l_{\text{VSC}}(q) = \left\|\sum_{i\in\mathcal{V}} A^l_{q,i}\hat{\mathbf{h}}^l_i\right\|_2^2$, which reduces the computational complexity from $\mathcal{O}(|\mathcal{V}|^2)$ to $\mathcal{O}(|\mathcal{V}|)$. Consequently, a large S^l_{VSC} indicates that the model not only attends strongly to the visual modality, but more importantly, that the attended tokens are highly clustered.

4.2 Visual Steering Ratio

Even when high visual semantic cohesion is achieved, it does not guarantee that the resulting evidence will prevail, as these signals may still be overridden by entrenched language priors. We therefore utilize *attention contributions* to measure the precise impact of input tokens on the generated output, revealing whether visual signals successfully steer the final decision.

Formally, let $\tilde{\mathbf{h}}^{l-1}_i = \text{LN}(\mathbf{h}^{l-1}_i)$ be the layer-normalized hidden state of source token i, where $\text{LN}(\cdot)$ denotes Layer Normalization. During the decoding of token q at layer l, each attention head first computes an attention-weighted value aggregation $\mathbf{o}^{l,h}_q$ over all preceding source tokens. The attention contribution is then obtained by projecting and summing the outputs of all heads:

$$\Delta\mathbf{h}^l_{\text{attn}}(q) = \sum_{h=1}^{H} \mathbf{o}^{l,h}_q \mathbf{W}^{l,h}_O. \tag{3}$$

Proposition 2 establishes that this attention-induced residual update admits an exact token-wise decomposition. By splitting the full context into the visual token set $\mathcal{V}$ and the preceding textual token set $\mathcal{T}$, the attention contributions can be decomposed as:

$$\Delta\mathbf{h}^l_{\text{attn}}(q) = \Delta\mathbf{h}^l_{\text{img}}(q) + \Delta\mathbf{h}^l_{\text{txt}}(q), \tag{4}$$

where $\Delta\mathbf{h}^l_{\text{img}}(q)$ and $\Delta\mathbf{h}^l_{\text{txt}}(q)$ denote the visual and textual contributions to the residual stream, obtained by summing the token-wise terms over $\mathcal{V}$ and $\mathcal{T}$, respectively.

Consistent with Observation 2, we adopt the magnitude of these *modality-specific contributions* as a proxy for their influence. We thus define the **Visual Steering Ratio (S_{VSR})** as:

$$S^l_{\text{VSR}}(q) := \frac{\|\Delta \mathbf{h}^l_{\text{img}}(q)\|_2}{\|\Delta \mathbf{h}^l_{\text{txt}}(q)\|_2 + \epsilon}. \tag{5}$$

Here, $\epsilon > 0$ prevents numerical instability. A high $S^l_{\text{VSR}}(q)$ signifies the predominance of *visual contributions* in the residual stream, effectively steering the generative process toward grounded evidence. In contrast, a low score reflects that the *visual contribution* is attenuated by language priors. By analyzing the trajectory of S_{VSR}, we can distinguish between failed visual transmission and prior-driven decoding.

4.3 Visual Signal Integrity

The two statistics above capture complementary aspects of visual signals. A reliable generation therefore requires both conditions to hold: the visual evidence should be semantically well formed, and it should also dominate over competing textual context when the token is produced.

We combine these two signals into a unified metric, **Visual Signal Integrity (S_{VSI})**, defined as

$$S^l_{\text{VSI}}(q) = S^l_{\text{VSC}}(q) + \lambda S^l_{\text{VSR}}(q), \tag{6}$$

where $\lambda = 5$ is a fixed scale-balancing coefficient. The ablation results provided in Appendix D.2. A high $S_{\text{VSI}}(q)$ indicates that visual information is both coherently formed and effectively preserved throughout generation. Conversely, a low score suggests a breakdown in the visual information flow.

4.4 Theoretical Analysis for Visual Signal Integrity

Our framework relies on two properties of transformer-based LVLMs. First, attention-weighted visual cohesion admits an equivalent norm-based formulation, enabling efficient computation of S^l_{VSC}. Second, the attention-induced residual update admits an exact source-wise decomposition, allowing separation of visual and textual contributions. Together, these properties support measuring both visual coherence and its influence during decoding.

Proposition 1 (Linear-time factorization of S_{VSC})**.** *Let $\mathcal{V}$ denote the set of visual tokens at layer l, and let $A^l_{q,i}$ be the attention weight assigned to visual token i by the current query token q. S^l_{VSC} admits the following equivalent formulation:*

$$S^l_{VSC}(q) = \left\| \sum_{i \in \mathcal{V}} A^l_{q,i} \hat{\mathbf{h}}^l_i \right\|_2^2, \tag{7}$$

where $\hat{\mathbf{h}}^l_i = \mathbf{h}^l_i / \|\mathbf{h}^l_i\|_2$ denotes the extracted semantic direction for visual token i at layer l.

Remark. Proposition 1 shows that S^l_{VSC} can be computed by first aggregating the attention-weighted semantic directions and then taking a single squared L_2-norm, reducing the computational complexity from $\mathcal{O}(|\mathcal{V}|^2)$ pairwise comparisons to $\mathcal{O}(|\mathcal{V}|)$ aggregation.

Proposition 2 (Token-wise contribution decomposition)**.** *Let $\tilde{\mathbf{h}}^{l-1}_i = \text{LN}(\mathbf{h}^{l-1}_i)$ denote the layer-normalized hidden state of source token i, and let $\alpha^{l,h}_{q,i}$ be the attention weight from query token q to source token i at head h. Then the attention contribution written to token q can be exactly decomposed into token-wise source contributions:*

$$\Delta \mathbf{h}^l_{\text{attn}}(q) = \sum_{i \le q} \Delta \mathbf{h}^l_{q,i}, \quad \textit{with} \quad \Delta \mathbf{h}^l_{q,i} = \sum_{h=1}^{H} \alpha^{l,h}_{q,i} \tilde{\mathbf{h}}^{l-1}_i W^{l,h}_V W^{l,h}_O, \tag{8}$$

where $\mathbf{W}^{l,h}_V$ and $\mathbf{W}^{l,h}_O$ are the value and output projection matrices for head h. $\Delta \mathbf{h}^l_{q,i}$ represents the residual contribution of source token i to the decoding state of token q across all attention heads.

Remark. Proposition 2 establishes the exact source-wise decomposition underlying our modality-level residual analysis. By aggregating the token-wise terms $\Delta \mathbf{h}^l_{q,i}$ over visual tokens $\mathcal{V}$ and textual tokens $\mathcal{T}$, we obtain $\Delta \mathbf{h}^l_{\text{img}}(q)$ and $\Delta \mathbf{h}^l_{\text{txt}}(q)$. These two contributions form the basis of S_{VSR}, which quantifies the relative dominance between visual evidence and textual priors during decoding.

Table 1: Performance comparison of hallucination detection. AUROC(%) and AUPR(%) scores are evaluated across four different LVLM architectures on two benchmark datasets.

Dataset	Method	LLaVA-1.5-7B		InternVL3-8B		Shikra		InstructBLIP	
		AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
MSCOCO	Entropy(Malinin & Gales, 2021)	69.23	79.47	63.49	88.33	68.95	78.04	54.42	79.01
	NLL(Zhou et al., 2024)	68.99	79.37	62.24	88.09	67.30	77.33	54.97	79.58
	Intern-Conf(Jiang et al., 2025a)	77.41	87.93	67.83	90.52	76.08	87.12	80.69	92.21
	SVAR(Jiang et al., 2025b)	73.11	87.64	74.67	93.81	66.59	81.12	74.68	90.19
	ContextLens(Phukan et al., 2025)	73.14	85.49	66.92	90.86	72.16	83.88	80.34	92.42
	GLSim(Park & Li, 2025)	**80.97**	**90.14**	68.83	91.55	68.91	81.68	56.97	82.42
	Ours	78.99	89.87	**78.33**	**95.18**	**77.21**	**87.78**	**83.33**	**93.64**
Objects365	Entropy(Malinin & Gales, 2021)	64.74	68.42	60.25	75.12	64.05	67.58	49.19	64.47
	NLL(Zhou et al., 2024)	65.07	68.68	59.38	74.85	62.69	66.85	55.37	67.19
	Intern-Conf(Jiang et al., 2025a)	**71.84**	75.22	60.41	76.45	66.14	71.34	69.70	77.07
	SVAR(Jiang et al., 2025b)	66.25	73.53	60.80	76.22	61.34	68.59	60.21	72.40
	ContextLens(Phukan et al., 2025)	64.58	70.77	59.01	75.86	63.64	69.41	70.21	79.27
	GLSim(Park & Li, 2025)	68.40	71.83	56.42	73.38	62.70	68.50	58.89	71.73
	Ours	70.18	**75.85**	**63.51**	**78.85**	**67.36**	**73.07**	**70.74**	**79.62**

Table 2: Performance comparison of hallucination detection. AUROC(%) and AUPR(%) scores are evaluated across four Qwen-family LVLMs on two benchmark datasets.

Dataset	Method	Qwen2-VL-7B		Qwen2.5-VL-7B		Qwen3-VL-4B		Qwen3-VL-8B	
		AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
MSCOCO	Entropy(Malinin & Gales, 2021)	63.69	82.99	65.75	85.10	61.75	78.09	60.76	77.02
	NLL(Zhou et al., 2024)	62.63	82.70	65.28	84.90	61.36	77.90	58.84	74.39
	Intern-Conf(Jiang et al., 2025a)	72.02	88.57	65.05	87.01	71.50	82.41	71.91	82.14
	SVAR(Jiang et al., 2025b)	78.88	92.68	76.16	91.90	75.35	87.45	74.61	86.62
	ContextLens(Phukan et al., 2025)	72.98	90.11	68.34	88.34	64.46	80.36	63.99	79.39
	GLSim(Park & Li, 2025)	72.75	90.74	70.99	89.59	69.36	82.40	61.25	78.78
	Ours	**79.88**	**93.06**	**78.19**	**92.62**	**77.94**	**88.74**	**76.68**	**87.95**
Objects365	Entropy(Malinin & Gales, 2021)	58.77	67.45	60.09	72.39	59.10	66.40	59.07	65.96
	NLL(Zhou et al., 2024)	58.46	67.37	59.21	72.15	59.15	66.59	57.69	64.30
	Intern-Conf(Jiang et al., 2025a)	**69.39**	**76.13**	62.06	75.27	68.59	71.64	67.32	70.71
	SVAR(Jiang et al., 2025b)	66.95	75.63	63.52	76.06	63.08	68.51	63.49	69.16
	ContextLens(Phukan et al., 2025)	63.94	71.79	59.51	73.70	53.90	61.24	55.86	62.08
	GLSim(Park & Li, 2025)	66.79	75.67	61.90	77.80	61.06	68.06	57.22	65.16
	Ours	67.78	75.78	**66.23**	**77.91**	**69.16**	**73.67**	**68.00**	**72.38**

5 EXPERIMENTS

Datasets. We evaluate our method on both object-level and reasoning-intensive settings. For object hallucination, we sample 5,000 random images each from MSCOCO (Lin et al., 2014) and Objects365 (Shao et al., 2019). We employ the CHAIR protocol (Rohrbach et al., 2018) to extract hallucination metrics; specifically for Objects365, we adapt the category list to its 365 classes (see Appendix C.6). Additionally, we evaluate on POPE GQA (Li et al., 2023b) using 3,000 samples from each of its three subsets (Popular, Random, and Adversarial).

Models. We conduct the evaluation on a diverse set of open-source LVLMs, including Qwen2-VL-7B (Wang et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B, Qwen3-VL-4B (Bai et al., 2025a), InternVL3-8B (Zhu et al., 2025), LLaVA-1.5-7B (Liu et al., 2024), Shikra (Chen et al., 2023), and InstructBLIP (Dai et al., 2023). More details are provided in Appendix C.3.

Baselines. We compare our approach against a comprehensive set of baselines, categorized as follows: (1) *Confidence*-based approaches—Negative Log-Likelihood (NLL) (Zhou et al., 2024), Token-level Entropy (Malinin & Gales, 2021) and Intern-Conf (Jiang et al., 2025a); (2) *Attention*-based approach—Summed Visual Attention Ratio (SVAR) (Jiang et al., 2025b); and (3) *Embedding similarity*-based approaches—Contextual Lens (Phukan et al., 2025) and GLSim (Park & Li, 2025). We evaluate all baselines on identical test sets using the default experimental configurations provided in their respective papers. Further details of baselines are discussed in Appendix C.1.

5.1 COMPARISONS WITH BASELINES

Results on Diverse LVLM Architectures. Table 1 reports the results on four LVLM architectures over MSCOCO and Objects365 with competitive object hallucination detection methods. VSI demonstrates strong and consistent performance across most model–dataset pairs. On MSCOCO,

Model	Method	AUROC	AUPR
LLaVA-1.5-7B	VSC	77.86	89.21
	VSR	77.68	89.25
	VSI	78.99	89.87
Qwen2.5-VL-7B	VSC	75.76	91.76
	VSR	74.13	91.39
	VSI	78.19	92.62

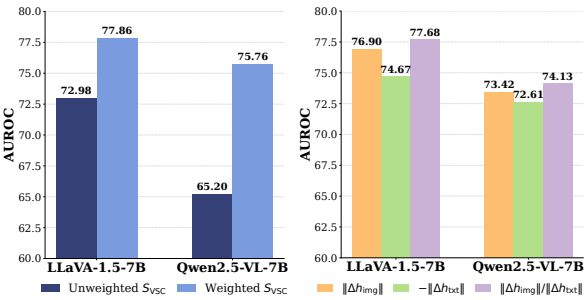


Figure 5: Ablation studies on MSCOCO. (i) Left: component ablation showing that VSC and VSR provide effective hallucination signals, while VSI yields stronger performance. (ii) Middle: comparison between attention-weighted VSC and its unweighted variant, validating the importance of query-dependent token weighting. (iii) Right: comparison among different formulations of visual steering, including raw visual contribution norm, negative textual contribution norm, and VSR.

VSI achieves the best performance on InternVL3-8B, Shikra, and InstructBLIP, with notable gains on InternVL3-8B (78.33% AUROC and 95.18% AUPR). Compared with recent methods, VSI improves AUROC by 3.66% over SVAR and by 9.50% over ContextLens on InternVL3-8B, while also yielding a 1.37% AUPR gain over SVAR. On Objects365, VSI maintains strong performance across models, achieving leading results on InternVL3-8B, Shikra and InstructBLIP. On InternVL3-8B, VSI improves AUROC by 2.71% over SVAR and by 4.50% over ContextLens, while also outperforming confidence-based methods such as Entropy and NLL. Unlike SVAR, which relies on attention magnitude, and GLSim, which focuses on representation similarity, VSI jointly models visual semantic cohesion and residual-stream steering, leading to more reliable hallucination detection.

Results on Qwen-family LVLMs. Table 2 reports the results on four Qwen-family LVLMs over MSCOCO and Objects365. On MSCOCO, VSI achieves the best performance across all four models in both AUROC and AUPR. Specifically, on Qwen2.5-VL-7B, VSI improves AUROC by 2.03% over SVAR and by 7.20% over GLSim, while also yielding a 0.72% AUPR gain over SVAR. On Qwen3-VL-4B, VSI further improves AUROC by 2.59% over SVAR and by 6.44% over Intern-Conf, demonstrating its effectiveness on newer Qwen variants. On Objects365, VSI maintains strong performance across models, achieving leading results on Qwen2.5-VL-7B, Qwen3-VL-4B, and Qwen3-VL-8B. Specifically, on Qwen3-VL-4B, VSI improves AUROC by 5.57% over SVAR and by 0.57% over Intern-Conf, while also yielding a 2.03% AUPR gain over Intern-Conf. These results show that VSI remains effective across both MSCOCO and the more challenging Objects365 setting.

5.2 Ablation Studies

Component Ablation. We evaluate VSC, VSR, and their combination VSI on MSCOCO for object hallucination detection. As shown in the left panel of Figure 5, both VSC and VSR provide effective signals, while VSI achieves the best overall performance. On LLaVA-1.5-7B, the proposed VSI improves AUROC from 77.86% / 77.68% to 78.99%. On Qwen2.5-VL-7B, VSI increases AUROC from 75.76% / 74.13% to 78.19%. These gains show that VSC and VSR capture complementary yet distinct hallucination cues from visual concept cohesion and decoding-time visual steering.

Effect of Attention Weighting in VSC. We compare the proposed attention-weighted VSC with an unweighted variant that uniformly aggregates the top 50% attended visual tokens. As shown in the middle panel of Figure 5, attention weighting improves AUROC from 72.98% to 77.86% on LLaVA-1.5-7B and from 65.20% to 75.76% on Qwen2.5-VL-7B. This indicates that query-dependent token importance is essential for accurately estimating object-specific visual cohesion.

Effect of VSR Formulation. We compare VSR with raw visual and negative textual contribution norms. As shown in the right panel of Figure 5, the raw visual norm is informative, achieving 76.90% and 73.42% AUROC on LLaVA-1.5-7B and Qwen2.5-VL-7B. The VSR improves performance to 77.68% and 74.13% AUROC, outperforming both raw-norm variants. This confirms that hallucination detection benefits from measuring the relative dominance of visual evidence over textual priors.

5.3 More Discussion

Generalization to probing-style QA. To further examine whether the proposed VSI generalizes beyond caption-based object hallucination detection, we evaluate VSI on POPE GQA, a probing-style binary QA benchmark for object-existence hallucinations. POPE contains three subsets constructed with different negative-sampling strategies: Random (Ran.), Popular (Pop.), and Adversarial (Adv.), which respectively evaluate general hallucination behavior, frequency bias, and robustness to hard negatives. Table 3 reports the averaged results across all evaluated LVLMs. VSI achieves the best performance across all AUROC and AUPR metrics. Specifically, VSI improves AUROC over Entropy by 1.03%, 2.13%, and 2.71% on the Ran., Pop., and Adv. subsets, respectively, while also yielding consistent AUPR gains over NLL. In this probing-style setting, object hallucination detection methods such as GLSim and SVAR show clear performance degradation, suggesting that object-level hallucination cues from captioning may not transfer directly to question-answering-based detection. In contrast, VSI remains consistently effective, indicating that VSC and VSR capture a more general visual reliability signal across evaluation protocols.

Table 3: AUROC(%) and AUPR(%) on the POPE GQA benchmark. We report results across three subsets.

Method	AUROC ↑			AUPR ↑		
	Ran.	**Pop.**	**Adv.**	**Ran.**	**Pop.**	**Adv.**
Entropy	<u>67.71</u>	<u>64.26</u>	<u>60.75</u>	<u>90.83</u>	85.52	81.73
NLL	67.09	63.90	60.58	90.47	<u>86.25</u>	<u>82.36</u>
Intern-Conf	44.22	47.11	48.71	82.69	77.99	76.45
SVAR	49.56	48.96	49.03	83.38	78.54	75.89
ContextLens	61.60	59.06	56.68	88.75	83.44	80.10
GLSim	54.86	53.22	49.87	86.84	81.35	77.46
Ours	**68.74**	**66.39**	**63.46**	**91.25**	**87.86**	**84.68**

Beyond object-existence hallucinations. While POPE still focuses on whether an object exists, we further test whether VSI extends to more fine-grained visual grounding errors. Specifically, we evaluate the Attribute and Relation subsets of MMHal-Bench (Sun et al., 2024) and the Color subset of MME (Fu et al., 2023) on LLaVA-1.5-7B and InstructBLIP. The Attribute subset tests whether visual properties are assigned to the correct entity, Relation evaluates whether interactions or spatial relations between entities are correctly grounded, and Color tests localized visual attributes. Table 4 shows that VSI consistently outperforms confidence- and representation-based baselines across both models. On LLaVA, VSI achieves 81.36/80.00/67.58 AUROC(%) on attribute/relation/color, improving over the strongest baseline by 0.75/0.44/2.60 points; on InstructBLIP, it reaches 85.56/78.89/75.38, exceeding GLSim by 3.78/4.67/2.14 points. These results show that the visual-integrity signal extends beyond object existence to incorrect properties and inter-object relations.

Table 4: Fine-grained AUROC (%) on MMHalBench Attribute/Relation and MME Color.

Method	LLaVA-1.5-7B			InstructBLIP		
	Attr.	Rel.	Color	Attr.	Rel.	Color
Entropy	75.76	79.56	63.78	46.67	50.74	51.03
NLL	79.82	78.15	63.51	68.89	49.63	37.76
Intern-Conf	75.00	42.67	38.95	37.78	56.67	32.15
GLSim	80.61	78.56	64.98	81.78	74.22	73.24
VSC	69.32	78.44	59.70	80.00	76.85	67.92
VSR	76.52	69.78	**68.03**	84.89	68.15	70.04
VSI	**81.36**	**80.00**	67.58	**85.56**	**78.89**	**75.38**

The component results reveal category-dependent behavior: VSR is stronger on attribute and color hallucinations, while VSC performs better on relations, suggesting that localized properties rely more on decoding-time visual steering whereas relations benefit from coherent multi-region representations. Their combination achieves the strongest component-level result in five of six settings, supporting their complementarity. However, high visual integrity does not guarantee semantic correctness; VSI measures whether visual evidence is coherently formed and used, rather than whether it is correct.

6 Conclusion

In this paper, we reframe hallucination in LVLMs as a failure of visual information preservation and propose Visual Signal Integrity (VSI), a training-free diagnostic framework that combines Visual Semantic Cohesion (VSC) and Visual Steering Ratio (VSR) to jointly assess whether visual evidence is coherently formed and faithfully preserved throughout generation. Extensive experiments across various LVLMs and benchmarks demonstrate that VSI outperforms baselines on most model–dataset pairs while providing interpretable diagnostic information beyond a binary hallucination label. Beyond object hallucination, our results on probing-style benchmarks suggest that VSI captures a more general notion of visual grounding, highlighting its potential for broader hallucination detection tasks and offering new insights into the internal mechanisms of multimodal generation.

AI USE STATEMENT

Generative AI tools were used only to assist with language polishing and presentation. Specifically, they were used to improve grammar, readability, and phrasing, and to assist with LaTeX formatting and the organization of tables. They were not used to formulate the research problem, design the proposed method, conduct experiments, generate or alter experimental data, determine evaluation protocols, interpret the results, or produce the scientific claims and conclusions of this work. The authors reviewed and verified all AI-assisted edits, numerical values, equations, citations, tables, and final manuscript content, and take full responsibility for the accuracy and integrity of the paper.

ETHICS STATEMENT

This work studies the detection and interpretation of hallucinations in large vision-language models. It does not involve human-subject research, crowdsourcing, personally identifiable information, private user data, or the collection of a new dataset. All experiments use established research benchmarks and publicly available or appropriately licensed model checkpoints. We use these resources for research evaluation in accordance with their intended terms and cite their original sources.

The intended purpose of VSI is to improve the transparency and reliability of multimodal systems by identifying when generated content lacks adequate visual grounding. Nevertheless, a hallucination detector should not be interpreted as a guarantee of factual correctness, particularly in safety-critical settings. Detector performance may vary across models, domains, demographic contexts, and image distributions, and errors may create either unwarranted trust or unnecessary rejection of valid model outputs. We therefore recommend that VSI be used as a diagnostic aid together with task-specific evaluation and appropriate human oversight. We have considered the potential benefits, limitations, and indirect risks of the proposed approach in Appendix F.

REPRODUCIBILITY STATEMENT

We provide the mathematical definitions, theoretical derivations, evaluation protocol, dataset statistics, model configurations, selected layers, hyperparameters, prompts, and implementation details required to reproduce the reported experiments throughout the paper and appendix. VSI is training-free; reproducing it requires only inference with the evaluated LVLMs and extraction of the hidden states, attention weights, and projection matrices specified in Section 4 and Appendix C.6. The pseudocode in Appendix C.6 describes the complete scoring procedure, while the reported dataset splits, random-seed protocol, and evaluation metrics specify how the final results are computed.

The complete source code will be released publicly. The release will include scripts for extracting model-internal signals, computing VSC, VSR, and VSI, running the evaluated baselines, preparing benchmark inputs, reproducing the ablations and additional experiments, and generating the reported tables and figures. We will also provide configuration files containing the model-specific layer choices, the shared coefficient λ, random seeds, prompts, and evaluation commands, together with environment and hardware instructions. Existing datasets and model weights will not be redistributed; instead, the documentation will identify their official sources, versions, and licenses and provide preprocessing instructions. These materials are intended to enable independent verification of the main findings and straightforward extension to additional open-weight LVLMs.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. In *arXiv preprint arXiv:2308.12966*, 2023. URL https://api.semanticscholar.org/CorpusID:261101015.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=s63dtq0mwA.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6688–6706, 2025.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13861–13871, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

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

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34716–34733, 2025.

Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25004–25014, 2025b.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025.

Shahar Katz and Yonatan Belinkov. Visit: Visualizing and interpreting the semantic information flow of transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14094–14113, 2023.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27228–27238, 2024.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7069, 2020. doi: 10.18653/v1/2020.emnlp-main.574.

Benno Krojer, Shravan Nayak, Oscar Mañas, Vaibhav Adlakha, Desmond Elliott, Siva Reddy, and Marius Mosbach. Latentlens: Revealing highly interpretable visual tokens in llms. *arXiv preprint arXiv:2602.00462*, 2026.

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

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023a.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31645–31653, 2026.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulic, and Anders Søgaard. Lost in embeddings: Information loss in vision-language models. *arXiv preprint*, 2025a.

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

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*, 2025b.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. *arXiv preprint arXiv:2510.08510*, 2025.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4183–4198, 2025.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3378–3400, 2023.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chanJGoa7f.

Seongheon Park and Sharon Li. GLSim: Detecting object hallucinations in LVLMs via global-local similarity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ZO8LyCizx9.

Anirudh Phukan, Divyansh Divyansh, Harshit Kumar Morj, Vaishnavi Vaishnavi, Apoorv Saxena, and Koustava Goswami. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. In *North American Chapter of the Association for Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:279232593.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439, 2019.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann Lecun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *International Conference on Machine Learning*, pp. 55854–55875. PMLR, 2025.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, et al. Multimodal learning with next-token prediction for large multimodal models. *Nature*, pp. 1–7, 2026.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 13040–13051, 2024.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gfFVATffPd.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11991–12011, 2024.

Eric B Zhou, Dokyun Lee, and Bin Gu. Who expands the human creative frontier with generative ai: Hive minds or masterminds? *Science Advances*, 11(36):eadu5800, 2025.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1tZbq88f27.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

CONTENTS

A Theoretical Analysis **16**

A.1 Proof of Linear-time Factorization of Visual Semantic Cohesion 16

A.2 Proof of Residual Contribution Decomposition 16

B More Related Work **18**

B.1 Large Vision-Language Models 18

B.2 Object Hallucination in LVLMs 18

B.3 Hallucination Benchmarks 18

B.4 Mechanistic Interpretability and Multimodal Information Flow 18

C More Details on Experiment Settings **20**

C.1 More Details on Baselines 20

C.2 More Details on Datasets 21

C.3 More Details on Models 21

C.4 More Details on Evaluation Metrics 21

C.5 Pseudo Code of VSI 22

C.6 Implementation Details 22

C.7 Statistics of Evaluated Object Mentions 23

D More Experimental Results **24**

D.1 Main Results Across Random Seeds 24

D.2 Sensitivity to the Weighting Coefficient 24

D.3 More Results on Observations 25

D.4 Controlled Non-assertive Intervention on Textual Priors 26

D.5 Separating Attention Magnitude from Directional Coherence 27

D.6 Cross-model Proxy Evaluation 28

D.7 Ablation on Token Selection for Multi-token Object Names 28

D.8 Layer-wise Stability and Depth-dependent Behavior 29

D.9 Computational Overhead 29

E Limitations and Future Directions **30**

E.1 Limitations 30

E.2 Future Directions 30

F Broader Impacts **30**

G Visualizations **31**

G.1 Diagnostic Quadrant Analysis of Visual Signal Integrity 31

G.2 KDE Visualization Across Models 31

G.3 Visualization of Benchmark Cases 33

A Theoretical Analysis

A.1 Proof of Linear-time Factorization of Visual Semantic Cohesion

Proof of Proposition 1. Recall that the current query token q's Visual Semantic Cohesion score at layer l is defined as the attention-weighted pairwise cosine similarity among visual tokens:

$$S^l_{\text{VSC}}(q) = \sum_{i,j\in\mathcal{V}} A^l_{q,i} A^l_{q,j} \cos(\mathbf{h}^l_i, \mathbf{h}^l_j). \tag{9}$$

Let

$$\hat{\mathbf{h}}^l_i = \frac{\mathbf{h}^l_i}{\|\mathbf{h}^l_i\|_2}$$

denote the normalized semantic direction of visual token i. Since

$$\cos(\mathbf{h}^l_i, \mathbf{h}^l_j) = (\hat{\mathbf{h}}^l_i)^\top \hat{\mathbf{h}}^l_j,$$

we can rewrite S^l_{VSC} as

$$S^l_{\text{VSC}}(q) = \sum_{i,j\in\mathcal{V}} A^l_{q,i} A^l_{q,j} (\hat{\mathbf{h}}^l_i)^\top \hat{\mathbf{h}}^l_j. \tag{10}$$

By regrouping the scalar weights with the normalized vectors, we obtain

$$S^l_{\text{VSC}}(q) = \sum_{i,j\in\mathcal{V}} \left(A^l_{q,i}\hat{\mathbf{h}}^l_i\right)^\top \left(A^l_{q,j}\hat{\mathbf{h}}^l_j\right). \tag{11}$$

Using the bilinearity of the inner product, the double summation can be factorized as

$$S^l_{\text{VSC}}(q) = \left(\sum_{i\in\mathcal{V}} A^l_{q,i}\hat{\mathbf{h}}^l_i\right)^\top \left(\sum_{j\in\mathcal{V}} A^l_{q,j}\hat{\mathbf{h}}^l_j\right). \tag{12}$$

Since the two summed vectors are identical, this is exactly the squared L_2-norm of the aggregated attention-weighted semantic direction:

$$S^l_{\text{VSC}}(q) = \left\|\sum_{i\in\mathcal{V}} A^l_{q,i}\hat{\mathbf{h}}^l_i\right\|^2_2. \tag{13}$$

Therefore, the pairwise cohesion score admits an equivalent linear-time formulation. □

A.2 Proof of Residual Contribution Decomposition

Proof of Proposition 2. We follow the row-vector convention used in modern Transformer implementations. Let $\tilde{\mathbf{h}}^{l-1}_i \in \mathbb{R}^D$ be the layer-normalized hidden state of source token i at layer $l-1$.

For each attention head $h \in \{1, \dots, H\}$, the value vector of source token i is

$$\mathbf{v}^{l,h}_i = \tilde{\mathbf{h}}^{l-1}_i \mathbf{W}^{l,h}_V, \tag{14}$$

where $\mathbf{W}^{l,h}_V \in \mathbb{R}^{D\times D_k}$. The output of head h for query token q is the attention-weighted aggregation of value vectors:

$$\mathbf{o}^{l,h}_q = \sum_{i\le q} \alpha^{l,h}_{q,i} \mathbf{v}^{l,h}_i = \sum_{i\le q} \alpha^{l,h}_{q,i} \tilde{\mathbf{h}}^{l-1}_i \mathbf{W}^{l,h}_V. \tag{15}$$

The outputs of all heads are concatenated as

$$\mathbf{O}^l_q = [\mathbf{o}^{l,1}_q, \dots, \mathbf{o}^{l,H}_q] \in \mathbb{R}^D.$$

The output projection matrix $\mathbf{W}^l_O \in \mathbb{R}^{D\times D}$ can be partitioned according to attention heads:

$$\mathbf{W}^l_O = \begin{bmatrix} \mathbf{W}^{l,1}_O \\ \vdots \\ \mathbf{W}^{l,H}_O \end{bmatrix}, \qquad \mathbf{W}^{l,h}_O \in \mathbb{R}^{D_k\times D}. \tag{16}$$

Therefore, by block matrix multiplication, the attention-induced residual update written to token q is

$$\Delta \mathbf{h}^l_{\text{attn}}(q) = \mathbf{O}^l_q \mathbf{W}^l_O = \sum_{h=1}^{H} \mathbf{o}^{l,h}_q \mathbf{W}^{l,h}_O. \tag{17}$$

Substituting the expression of $\mathbf{o}^{l,h}_q$ gives

$$\Delta \mathbf{h}^l_{\text{attn}}(q) = \sum_{h=1}^{H} \left(\sum_{i \le q} \alpha^{l,h}_{q,i} \tilde{\mathbf{h}}^{l-1}_i \mathbf{W}^{l,h}_V \right) \mathbf{W}^{l,h}_O. \tag{18}$$

By the linearity of matrix multiplication and finite summation, we can exchange the order of summation:

$$\Delta \mathbf{h}^l_{\text{attn}}(q) = \sum_{i \le q} \sum_{h=1}^{H} \alpha^{l,h}_{q,i} \tilde{\mathbf{h}}^{l-1}_i \mathbf{W}^{l,h}_V \mathbf{W}^{l,h}_O. \tag{19}$$

Define the source-token residual contribution as

$$\Delta \mathbf{h}^l_{q,i} := \sum_{h=1}^{H} \alpha^{l,h}_{q,i} \tilde{\mathbf{h}}^{l-1}_i \mathbf{W}^{l,h}_V \mathbf{W}^{l,h}_O. \tag{20}$$

Then the total attention-induced residual update can be written as

$$\Delta \mathbf{h}^l_{\text{attn}}(q) = \sum_{i \le q} \Delta \mathbf{h}^l_{q,i}. \tag{21}$$

This establishes the exact token-wise residual contribution decomposition.

Finally, since the source tokens can be partitioned into visual tokens $\mathcal{V}$ and textual tokens $\mathcal{T}$, the same decomposition yields the modality-level updates:

$$\Delta \mathbf{h}^l_{\text{img}}(q) = \sum_{i \in \mathcal{V}} \Delta \mathbf{h}^l_{q,i}, \qquad \Delta \mathbf{h}^l_{\text{txt}}(q) = \sum_{i \in \mathcal{T}} \Delta \mathbf{h}^l_{q,i}. \tag{22}$$

Thus,

$$\Delta \mathbf{h}^l_{\text{attn}}(q) = \Delta \mathbf{h}^l_{\text{img}}(q) + \Delta \mathbf{h}^l_{\text{txt}}(q), \tag{23}$$

which provides the residual decomposition used by the Visual Steering Ratio. □

B More Related Work

B.1 Large Vision-Language Models

Recent LVLMs are typically built upon three components: a visual encoder (e.g., CLIP (Radford et al., 2021), EVA (Fang et al., 2023)) that encodes the image, a modality connector that projects visual features into the language embedding space, and a pretrained LLM backbone that jointly processes visual and textual tokens to produce responses. LLaVA-1.5 (Liu et al., 2024) uses an MLP projector to map the visual encoder output into 576 image tokens, Shikra (Chen et al., 2023) adopts a single linear layer for alignment, and MiniGPT-4 (Zhu et al., 2024) uses a learnable querying transformer that reduces image input to just 32 tokens. More recent models such as Qwen-VL (Bai et al., 2023), Qwen2-VL (Wang et al., 2024), Qwen2.5-VL (Bai et al., 2025b), InternVL3 (Zhu et al., 2025), LLaVA-NeXT-Interleave (Li et al., 2024), and mPLUG-Owl2 (Ye et al., 2024) further scale this paradigm with improved visual encoders, more flexible multimodal interleaving, and larger LLM backbones.

B.2 Object Hallucination in LVLMs

Object hallucination refers to the tendency of LVLMs to generate descriptions of objects that are absent from the image (Rohrbach et al., 2018; Li et al., 2023b). This issue has been attributed to statistical biases in training data (Zhou et al., 2024), strong language model priors (Leng et al., 2024) that cause models to favor contextually plausible but visually absent objects, and visual information loss during encoding (Li et al., 2025a) that leaves the model under-constrained by the actual image content. Recent work has also studied mitigation or diagnosis through alignment and inference-time intervention, including factually augmented RLHF for multimodal models (Sun et al., 2024) and intervention-based truthfulness elicitation at decoding time (Li et al., 2023a). On the detection side, uncertainty-based criteria remain a common baseline, including recent LVLM-specific uncertainty formulations (Zhang et al., 2024). Standard evaluation protocols include MSCOCO- and Objects365-based caption analysis, which measure how often nonexistent objects are mentioned; POPE (Li et al., 2023b), which frames object probing as binary yes/no questions; MME (Fu et al., 2023), which uses multiple-choice perception tasks; and HallusionBench (Guan et al., 2024), which targets compositional and counter-intuitive cases requiring strong visual grounding.

B.3 Hallucination Benchmarks

Existing hallucination benchmarks evaluate different forms of visual grounding failure. CHAIR (Rohrbach et al., 2018) measures object hallucination in free-form captions by comparing generated object mentions with ground-truth annotations, and is commonly used with MSCOCO (Lin et al., 2014). POPE (Li et al., 2023b) reformulates object hallucination as binary object-existence probing and constructs random, popular, and adversarial negative samples to evaluate robustness against language priors. Objects365 (Shao et al., 2019) provides a larger object vocabulary than MSCOCO and enables more fine-grained object-level evaluation. Broader diagnostic suites such as MME (Fu et al., 2023), HallusionBench (Guan et al., 2024), AMBER (Wang et al., 2023), and THRONE (Kaul et al., 2024) further extend evaluation to perception, reasoning, attribute, relation, or free-form hallucination settings. These benchmarks are complementary: some emphasize open-ended caption faithfulness, while others stress closed-form object existence or visually counter-intuitive cases. We therefore evaluate VSI across both caption-based and probing-based settings to test whether its internal grounding signal generalizes across evaluation protocols.

B.4 Mechanistic Interpretability and Multimodal Information Flow

Mechanistic interpretability provides tools for analyzing how information is represented and propagated in Transformers. Residual-stream analyses view computation as additive updates from attention heads and MLPs (Elhage et al., 2021; Vaswani et al., 2017), while logit-lens, tuned-lens, and attention-flow methods trace how intermediate representations evolve and propagate across layers (Belrose et al., 2023; Abnar & Zuidema, 2020). Prior work also shows that attention weights alone do not determine token influence, because effective contributions depend on value vectors and output projections (Kobayashi et al., 2020; Mohebbi et al., 2023; Yuksekgonul et al., 2024). In multimodal

models, recent studies examine how visual information is transferred, attenuated, and organized inside LVLMs (Basu et al., 2024; Luo et al., 2025; Katz & Belinkov, 2023; Krojer et al., 2026; Neo et al., 2025; Li et al., 2026; Kang et al., 2025). Building on these insights, VSI turns information-flow analysis into a hallucination diagnostic: VSC measures the semantic cohesion of attended visual tokens, and VSR measures whether visual contributions dominate textual ones in the residual stream.

C More Details on Experiment Settings

C.1 More Details on Baselines

GLSim. In GLSim, both local and global similarity scores are computed for each object. Object-relevant regions are identified via the visual logit lens:

$$\mathcal{I}(o) = \mathrm{TopK}_{v_i \in \mathbf{v}} \left(\{ \mathrm{softmax}(\mathrm{VLL}_l(v_i))[o] \} \right), \tag{24}$$

where

$$\mathrm{VLL}_l(v_i) = h_l(v_i)\, W_U, \quad W_U \in \mathbb{R}^{d \times |\mathcal{V}|}. \tag{25}$$

The local similarity score is defined as:

$$s_{\mathrm{local}}(o, x) = \frac{1}{K} \sum_{v_i \in \mathcal{I}(o)} \mathrm{sim}\left(h_l\left(v_i\right), h_{l'}(o) \right). \tag{26}$$

The global similarity score is defined as :

$$s_{\mathrm{global}}(o, x) = \mathrm{sim}\left(h_l(\mathbf{v}, \mathbf{t}), h_{l'}(o) \right). \tag{27}$$

The final score is defined as follows:

$$s_{\mathrm{GLSim}}(o, x) = w \cdot s_{\mathrm{global}}(o, x) + (1 - w) \cdot s_{\mathrm{local}}(o, x). \tag{28}$$

Negative Log-Likelihood (NLL). Following the original method, the hallucination score is derived from the autoregressive generation probability of the object token. For the object o be generated at position j, the conditional probability is defined as $p(o \mid \mathbf{y}_{<j}, \mathbf{v})$. The NLL-based score is:

$$s_{\mathrm{NLL}}(o, x) = -\log p(o \mid \mathbf{y}_{<j}, \mathbf{v}). \tag{29}$$

Entropy. Entropy measures the uncertainty of the token distribution at position j. It is defined as:

$$s_{\text{Entropy}}(o, x) = -\sum_{y \in \mathcal{V}} p(y \mid \mathbf{y}_{<j}, \mathbf{v}) \log p(y \mid \mathbf{y}_{<j}, \mathbf{v}), \tag{30}$$

where $\mathcal{V}$ is the vocabulary space, we use $s'_{\text{Entropy}} = -s_{\text{Entropy}}$.

Internal Confidence (IC). Internal Confidence leverages the visual logit lens to probe object likelihood from visual representations. Let $\mathrm{VLL}_l(v_i)$ denote the projected logits at layer l for visual token v_i. The score is defined as the maximum probability across all layers and visual tokens:

$$s_{\mathrm{IC}}(o, x) = \max_{l \in [L]} \max_{i \in [N]} \mathrm{VLL}_l(v_i)[o], \tag{31}$$

where L is the number of layers and N is the number of image tokens.

Summed Visual Attention Ratio (SVAR). SVAR quantifies how much attention an object token attends to image tokens. For a given layer ℓ and attention head h, the Visual Attention Ratio (VAR) is:

$$\mathrm{VAR}^{(\ell,h)}(o) \triangleq \sum_{i=1}^{N} A^{(\ell,h)}(o, v_i), \tag{32}$$

where $A^{(\ell,h)}(o, v_i)$ denotes the attention weight from object token o to visual token v_i. SVAR aggregates over heads and layers:

$$s_{\mathrm{SVAR}}(o, x) = \frac{1}{H} \sum_{\ell=\ell_1}^{\ell_2} \sum_{h=1}^{H} \mathrm{VAR}^{(\ell,h)}(o), \tag{33}$$

where H is the number of heads, and $[\ell_1, \ell_2]$ defines the selected layer range.

Contextual Lens. Contextual Lens is an embedding similarity-based method. Originally designed for sentence-level detection, we adapt it to object-level scoring. Let $h_{l_T}(o)$ denote the embedding of object token o at text layer l_T, and $h_{l_I}(v_i)$ denote the embedding of visual token v_i at image layer l_I. The score is computed as:

$$s_{\mathrm{CL}}(o, x) = \max_{i \in [N]} \mathrm{sim}\left(h_{l_T}(o), h_{l_I}(v_i) \right), \tag{34}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity.

C.2 More Details on Datasets

In this section, we provide additional details regarding the datasets used for evaluating object hallucinations in our method:

MSCOCO (Lin et al., 2014): The MSCOCO dataset contains approximately 80K training images and 40K validation images, covering 80 object categories.

Objects365 (Shao et al., 2019): The Objects365 dataset consists of around 2 million images with 365 object categories and over 30 million bounding box annotations.

POPE (Li et al., 2023b): The POPE benchmark contains approximately 9K binary (yes/no) question–answer pairs constructed on GQA images. It is further divided into three subsets based on negative sampling strategies, including *Random*, *Popular*, and *Adversarial*, which respectively evaluate general hallucination behavior, frequency bias, and robustness under hard negative samples.

To standardize the evaluation process, we apply consistent instruction templates across all models for these datasets. The exact prompting formats used for the POPE binary object-existence probing benchmark and the free-form captioning benchmarks (MSCOCO and Objects365) are detailed below:

Inducing prompt for POPE

Original Prompt: Is there a/an [object] in the image?
Neutral: Please help me answer this simple visual question: [Original Prompt]
Yes-Induce: Since the image actually does contain this object: [Original Prompt]
No-Induce: Since the image does not contain this object: [Original Prompt]

Prompt for CHAIR&Objects365

Please help me describe the image in detail.

C.3 More Details on Models

To demonstrate the generalizability of our hallucination detection method, we evaluate it across a comprehensive suite of LVLMs, covering diverse architectures and model scales:

LLaVA-1.5-7B (Liu et al., 2024): A widely used baseline model on visual instruction data, which connects a CLIP-based visual encoder with the Vicuna language model via an MLP-based adapter.

Qwen-VL Series (Bai et al., 2025a;b; Wang et al., 2024): We conduct experiments on the evolution of the Qwen-VL family, including Qwen2-VL-7B, Qwen2.5-VL-7B, Qwen3-VL-4B, and Qwen3-VL-8B. These models feature dynamic resolution and advanced multi-modal understanding, serving as highly representative paradigms for handling diverse visual aspect ratios and complex reasoning tasks.

InternVL3 (Zhu et al., 2025): As a high-capacity vision-language foundation model, the InternVL3-8B variant is evaluated to assess our zero-shot detection performance on architectures that possess exceptionally strong fine-grained visual-linguistic alignment capabilities.

InstructBLIP (Dai et al., 2023): This model utilizes a Q-Former architecture to bridge a frozen image encoder and a language model. It serves to verify our detector's robustness on models employing a bottleneck-based information flow, as opposed to direct linear or MLP projections.

Shikra (Chen et al., 2023): Specifically optimized for referential dialogue, Shikra is natively trained to handle spatial coordinates (bounding boxes). It thereby provides a robust baseline for evaluating the precise spatial alignment between text mentions and specific image regions.

C.4 More Details on Evaluation Metrics

For MSCOCO and Objects365, we evaluate hallucination detection at the level of generated object mentions. Each mention is assigned a scalar detection score, where a larger score indicates that the

token is more likely to be grounded rather than hallucinated. We then compare these scores against the binary ground-truth label of each object mention and report both AUROC and AUPR.

Specifically, given a set of object mentions $\{i\}_{i=1}^{N}$, each mention has a binary label $y_i \in \{0, 1\}$, where $y_i = 1$ denotes a grounded object mention and $y_i = 0$ denotes a hallucinated one, together with a detection score s_i. For any threshold τ, mentions with $s_i \geq \tau$ are predicted as grounded.

AUROC is computed as the area under the receiver operating characteristic curve obtained by sweeping τ over all possible score values and plotting the true positive rate against the false positive rate:

$$\mathrm{TPR}(\tau) = \frac{\sum_i \mathbb{I}[s_i \geq \tau \wedge y_i = 1]}{\sum_i \mathbb{I}[y_i = 1]}, \quad \mathrm{FPR}(\tau) = \frac{\sum_i \mathbb{I}[s_i \geq \tau \wedge y_i = 0]}{\sum_i \mathbb{I}[y_i = 0]}.$$

$$\mathbf{AUROC} = \int_0^1 \mathrm{TPR}(\tau) d\mathrm{FPR}(\tau).$$

AUPR is computed as the area under the precision-recall curve under the same threshold sweep, where precision and recall are defined as:

$$\mathrm{Precision}(\tau) = \frac{\sum_i \mathbb{I}[s_i \geq \tau \wedge y_i = 1]}{\sum_i \mathbb{I}[s_i \geq \tau]}, \quad r := \mathrm{Recall}(\tau) = \frac{\sum_i \mathbb{I}[s_i \geq \tau \wedge y_i = 1]}{\sum_i \mathbb{I}[y_i = 1]}.$$

$$\mathbf{AUPR} = \int_0^1 \mathrm{Precision}(r) dr.$$

Following standard practice, for all methods we orient the score direction so that higher values consistently correspond to grounded object mentions before computing these ranking-based metrics. For POPE, the same computation is applied to the queried object score on each image-question pair.

C.5 Pseudo Code of VSI

Algorithm 1 Computing Visual Signal Integrity

Input: Image I; prompt P; LVLM $\mathcal{M}$; generated object token q; visual token set $\mathcal{V}$; textual token set $\mathcal{T}$; selected layer l; scale coefficient λ; stability constant ϵ.
Run $\mathcal{M}(I, P)$ and obtain hidden states $\{h_i^l\}$, layer-normalized states $\{\tilde{h}_i^{l-1}\}$, attention weights $\{\alpha_{q,i}^{l,h}\}$, and projection matrices $\{W_V^{l,h}, W_O^{l,h}\}_{h=1}^{H}$.
for $i \in \mathcal{V}$ **do**
    $A_{q,i}^l \leftarrow \frac{1}{H} \sum_{h=1}^{H} \alpha_{q,i}^{l,h}$;
    $\hat{h}_i^l \leftarrow h_i^l / \|h_i^l\|_2$;
end for
$S_{\mathrm{VSC}}^l(q) \leftarrow \left\| \sum_{i \in \mathcal{V}} A_{q,i}^l \hat{h}_i^l \right\|_2^2$;
for $i \in \mathcal{V} \cup \mathcal{T}$ **do**
    $\Delta h_{q,i}^l \leftarrow \sum_{h=1}^{H} \alpha_{q,i}^{l,h} \tilde{h}_i^{l-1} W_V^{l,h} W_O^{l,h}$;
end for
$\Delta h_{\mathrm{img}}^l(q) \leftarrow \sum_{i \in \mathcal{V}} \Delta h_{q,i}^l$;
$\Delta h_{\mathrm{txt}}^l(q) \leftarrow \sum_{i \in \mathcal{T}} \Delta h_{q,i}^l$;
$S_{\mathrm{VSR}}^l(q) \leftarrow \frac{\|\Delta h_{\mathrm{img}}^l(q)\|_2}{\|\Delta h_{\mathrm{txt}}^l(q)\|_2 + \epsilon}$;
$S_{\mathrm{VSI}}^l(q) \leftarrow S_{\mathrm{VSC}}^l(q) + \lambda S_{\mathrm{VSR}}^l(q)$;
Output: Visual Signal Integrity score $S_{\mathrm{VSI}}^l(q)$.

C.6 Implementation Details

We implement our method with greedy decoding and a maximum generation length of 512 tokens. All hyperparameters are selected on a separate validation set. For multi-token object names, we compute the scores using the first token and consider only the first occurrence of each object for hallucination detection. The total number of generated object mentions is reported in Table 6. Unless otherwise specified, all results are averaged over three random seeds.

All experiments are conducted using Python 3.11.3 and PyTorch 2.7.1 on a single NVIDIA A800 GPU with 80GB of memory. For older LVLMs (e.g., LLaVA, InstructBLIP, and Shikra), we use Transformers 4.36.0; for newer LVLMs (e.g., Qwen2-VL, Qwen2.5-VL, Qwen3-VL, and InternVL3), we use Transformers 4.57.3 for inference. For baselines, we use the official GLSim codebase and reimplement the remaining methods by strictly following the algorithms described in their papers.

We evaluate on three standard hallucination benchmarks. **MSCOCO** (Lin et al., 2014) and **Objects365** (Shao et al., 2019) follow the evaluation settings of GLSim, with 5,000 randomly sampled images for each dataset. For MSCOCO, we use the CHAIR-style object-matching protocol based on ground-truth annotations and synonym dictionaries. For Objects365, we extend the same matching logic to the 365-category vocabulary to support large-vocabulary hallucination evaluation. **POPE** (Li et al., 2023b) is used as a binary hallucination benchmark. We evaluate on its three subsets, *Random*, *Popular*, and *Adversarial*, sampling 3,000 images from each subset and comparing our detector's predictions against the ground-truth existence of the queried object.

The key hyperparameters, including the layer index l to extract internal representations and the scaling coefficient λ for computing the final VSI score, are selected based on a separate validation set for each architecture, as detailed in Table 5. Additionally, the stability constant ϵ in the VSR formulation is consistently set to 1×10^{-7} across all experiments to prevent numerical instability.

Table 5: Hyperparameters used for different LVLMs in our main evaluations.

Model	Layer index l	Coefficient λ
InternVL3-8B	13	5.0
LLaVA-1.5-7B	10	5.0
Qwen2-VL-7B	11	5.0
Qwen2.5-VL-7B	11	5.0
Qwen3-VL-4B	13	5.0
Qwen3-VL-8B	14	5.0
Shikra	7	5.0
InstructBLIP	9	5.0

C.7 Statistics of Evaluated Object Mentions

Table 6 reports the numbers of grounded ("Real") and hallucinated ("Hal") object mentions extracted from model generations on the MSCOCO and Objects365 captioning benchmarks after applying the dataset-specific object-matching protocol. These counts reflect the effective sample size for object-level hallucination detection and show that hallucination ratios vary substantially across datasets and model families.

Table 6: Statistics of generated object mentions used for object hallucination evaluation. "Real" and "Hal" denote grounded and hallucinated object mentions, respectively. The hallucination ratio is computed as $\mathrm{Hal}/(\mathrm{Real} + \mathrm{Hal})$.

	MSCOCO				Objects365			
Model	Real	Hal	Total	Hal ratio	Real	Hal	Total	Hal ratio
InternVL3-8B	8,339	1,448	9,787	14.8%	11,813	5,000	16,813	29.7%
LLaVA-1.5-7B	10,924	4,696	15,620	30.1%	10,965	7,895	18,860	41.9%
Qwen2-VL-7B	9,074	2,539	11,613	21.9%	13,245	7,750	20,995	36.9%
Qwen2.5-VL-7B	8,597	2,195	10,792	20.3%	12,987	6,151	19,138	32.1%
Qwen3-VL-4B	9,138	3,931	13,069	30.1%	14,658	9,819	24,477	40.1%
Qwen3-VL-8B	9,153	4,047	13,200	30.7%	14,560	9,940	24,500	40.6%
Shikra	10,414	5,091	15,505	32.8%	10,315	7,624	17,939	42.5%
InstructBLIP	8,236	2,726	10,962	24.9%	9,945	5,710	15,655	36.5%

D MORE EXPERIMENTAL RESULTS

D.1 MAIN RESULTS ACROSS RANDOM SEEDS

To complement the main-paper results, we additionally report the full hallucination-detection performance with variability across three random seeds. In this subsection, we reproduce the two main comparison tables and augment each AUROC/AUPR entry with its standard deviation. This presentation makes the stability of the proposed method more explicit and helps distinguish consistent improvements from gains that may be sensitive to random initialization or sampling effects.

Unless otherwise specified, each result is reported in the form of mean $\pm$ standard deviation over three runs with different random seeds. We keep the same evaluation protocol, model settings, and benchmark splits as in the main paper so that these tables are directly comparable to the main results.

Table 7: Main hallucination-detection results on four LVLMs. Each entry reports mean $\pm$ standard deviation over three random seeds for AUROC(%) and AUPR(%) on MSCOCO and Objects365.

Dataset	Method	LLaVA-1.5-7B		InternVL3-8B		Shikra		InstructBLIP	
		AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
MSCOCO	Entropy(Malinin & Gales, 2021)	69.23 ± 1.09	79.47 ± 0.58	63.49 ± 1.57	88.33 ± 0.65	68.95 ± 0.79	78.04 ± 0.52	54.42 ± 1.05	79.01 ± 0.68
	NLL(Zhou et al., 2024)	68.99 ± 1.04	79.37 ± 0.69	62.24 ± 1.57	88.09 ± 0.66	67.30 ± 0.85	77.33 ± 0.51	54.97 ± 0.76	79.58 ± 0.64
	Intern-Conf(Jiang et al., 2025a)	77.41 ± 0.97	87.93 ± 0.79	67.83 ± 2.09	90.52 ± 0.77	76.08 ± 0.73	87.12 ± 0.52	80.69 ± 0.58	92.21 ± 0.25
	SVAR(Jiang et al., 2025b)	73.11 ± 1.32	87.64 ± 0.68	74.67 ± 1.84	93.81 ± 0.79	66.59 ± 0.92	81.12 ± 0.86	74.68 ± 1.09	90.19 ± 0.56
	ContextLens(Phukan et al., 2025)	73.14 ± 1.02	85.49 ± 0.74	66.92 ± 0.90	90.86 ± 0.60	72.16 ± 0.51	83.88 ± 0.31	80.34 ± 0.50	92.42 ± 0.29
	GLSim(Park & Li, 2025)	**80.97 ± 0.45**	**90.14 ± 0.39**	68.83 ± 1.66	91.55 ± 0.61	68.91 ± 0.59	81.68 ± 0.60	56.97 ± 1.02	82.42 ± 0.63
	Ours	78.99 ± 0.65	89.87 ± 0.36	**78.33 ± 1.46**	**95.18 ± 0.50**	**77.21 ± 0.89**	**87.78 ± 0.38**	**83.33 ± 0.80**	**93.64 ± 0.36**
Objects365	Entropy(Malinin & Gales, 2021)	64.74 ± 0.53	68.42 ± 0.84	60.25 ± 0.81	75.12 ± 0.76	64.05 ± 0.57	67.58 ± 0.88	49.19 ± 0.85	64.47 ± 0.82
	NLL(Zhou et al., 2024)	65.07 ± 0.50	68.68 ± 0.90	59.38 ± 0.88	74.85 ± 0.79	62.69 ± 0.58	66.85 ± 0.95	55.37 ± 0.99	67.19 ± 0.67
	Intern-Conf(Jiang et al., 2025a)	**71.84 ± 0.91**	75.22 ± 0.99	60.41 ± 1.41	76.45 ± 1.09	66.14 ± 0.36	71.34 ± 0.57	69.70 ± 0.77	77.07 ± 0.57
	SVAR(Jiang et al., 2025b)	66.25 ± 0.64	73.53 ± 0.58	60.80 ± 0.40	76.22 ± 0.48	61.34 ± 0.41	68.59 ± 0.41	60.21 ± 1.28	72.40 ± 0.87
	ContextLens(Phukan et al., 2025)	64.58 ± 0.95	70.77 ± 0.94	59.01 ± 1.31	75.86 ± 0.78	63.64 ± 0.63	69.41 ± 0.54	70.21 ± 0.61	79.27 ± 0.67
	GLSim(Park & Li, 2025)	68.40 ± 0.35	71.83 ± 1.25	56.42 ± 1.17	73.38 ± 0.63	62.70 ± 0.56	68.50 ± 0.57	58.89 ± 0.80	71.73 ± 0.76
	Ours	70.18 ± 0.46	**75.85 ± 0.33**	**63.51 ± 0.53**	**78.85 ± 0.23**	**67.36 ± 0.73**	**73.07 ± 0.85**	**70.74 ± 1.03**	**79.62 ± 0.68**

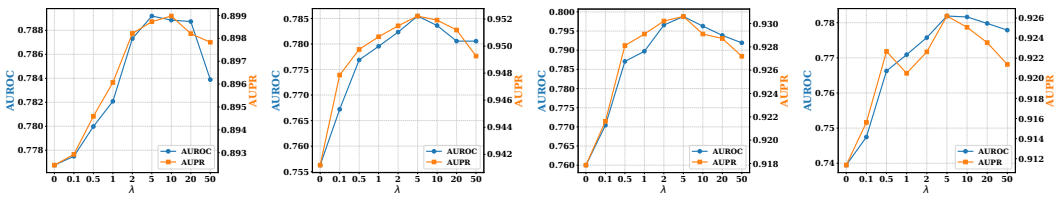
Table 8: Main hallucination-detection results on Qwen-family LVLMs. Each entry reports mean $\pm$ standard deviation over three random seeds for AUROC(%) and AUPR(%) on MSCOCO and Objects365.

Dataset	Method	Qwen2-VL-7B		Qwen2.5-VL-7B		Qwen3-VL-4B		Qwen3-VL-8B	
		AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑	AUROC ↑	AUPR ↑
MSCOCO	Entropy(Malinin & Gales, 2021)	63.69 ± 1.51	82.99 ± 1.24	65.75 ± 1.75	85.10 ± 0.82	61.75 ± 0.46	78.09 ± 0.49	60.76 ± 0.56	77.02 ± 0.61
	NLL(Zhou et al., 2024)	62.63 ± 1.67	82.70 ± 1.24	65.28 ± 1.89	84.90 ± 0.84	61.36 ± 0.16	77.90 ± 0.37	58.84 ± 0.48	74.39 ± 0.29
	Intern-Conf(Jiang et al., 2025a)	72.02 ± 0.70	88.57 ± 0.55	65.05 ± 0.68	87.01 ± 0.53	71.50 ± 0.97	82.41 ± 0.85	71.91 ± 0.92	82.14 ± 0.66
	SVAR(Jiang et al., 2025b)	78.88 ± 1.06	92.68 ± 0.51	76.16 ± 1.06	91.90 ± 0.49	75.35 ± 0.24	87.45 ± 0.19	74.61 ± 0.76	86.62 ± 0.72
	ContextLens(Phukan et al., 2025)	72.98 ± 1.51	90.11 ± 0.69	68.34 ± 1.10	88.34 ± 0.74	64.46 ± 1.32	80.36 ± 0.85	63.99 ± 0.59	79.39 ± 0.75
	GLSim(Park & Li, 2025)	72.75 ± 1.31	90.74 ± 0.43	70.99 ± 0.96	89.59 ± 0.60	69.36 ± 0.73	82.40 ± 0.93	61.25 ± 0.98	78.78 ± 1.33
	Ours	**79.88 ± 1.16**	**93.06 ± 0.42**	**78.19 ± 1.19**	**92.62 ± 0.59**	**77.94 ± 0.51**	**88.74 ± 0.55**	**76.68 ± 1.02**	**87.95 ± 0.61**
Objects365	Entropy(Malinin & Gales, 2021)	58.77 ± 0.89	67.45 ± 0.78	60.09 ± 0.99	72.39 ± 0.66	59.10 ± 0.71	66.40 ± 0.71	59.07 ± 1.25	65.96 ± 1.13
	NLL(Zhou et al., 2024)	58.46 ± 0.99	67.37 ± 0.85	59.21 ± 0.92	72.15 ± 0.66	59.15 ± 0.73	66.59 ± 0.62	57.69 ± 0.89	64.30 ± 0.60
	Intern-Conf(Jiang et al., 2025a)	**69.39 ± 1.18**	**76.13 ± 1.11**	62.06 ± 0.64	75.27 ± 0.19	68.59 ± 0.82	71.64 ± 0.99	67.32 ± 0.48	70.71 ± 0.51
	SVAR(Jiang et al., 2025b)	66.95 ± 0.21	75.63 ± 0.42	63.52 ± 0.77	76.06 ± 0.69	63.08 ± 0.51	68.51 ± 0.44	63.49 ± 0.50	69.16 ± 0.45
	ContextLens(Phukan et al., 2025)	63.94 ± 0.82	71.79 ± 0.56	59.51 ± 1.23	73.70 ± 1.00	53.90 ± 0.58	61.24 ± 0.76	55.86 ± 0.64	62.08 ± 1.11
	GLSim(Park & Li, 2025)	66.79 ± 1.01	75.67 ± 0.77	61.90 ± 0.78	77.80 ± 0.70	61.06 ± 0.79	68.06 ± 0.55	57.22 ± 0.50	65.16 ± 0.95
	Ours	67.78 ± 0.42	75.78 ± 0.49	**66.23 ± 0.72**	**77.91 ± 1.16**	**69.16 ± 0.74**	**73.67 ± 0.48**	**68.00 ± 0.64**	**72.38 ± 0.85**

D.2 SENSITIVITY TO THE WEIGHTING COEFFICIENT

We additionally examine how the relative weight between the two components affects detection performance. Figure 6 sweeps the coefficient λ in the additive combination of VSC and VSR on the CHAIR-style MSCOCO setting across four representative LVLMs. Several consistent patterns emerge. First, assigning a non-zero weight to VSR always improves substantially over the degenerate case λ=0, confirming that decoding-time visual steering contributes complementary information beyond semantic cohesion alone. Second, performance typically increases rapidly when moving from very small values of λ to moderate ones, and then remains relatively stable over a broad range around the optimum. In particular, InternVL3, Qwen2-VL-7B, and Qwen2.5-VL-7B all peak or plateau around $\lambda \in [5, 10]$, while LLaVA-1.5-7B reaches its best performance in a slightly larger range around $\lambda \in [5, 20]$.

These trends suggest that the combined score is not overly sensitive to the exact coefficient once both components are assigned comparable influence. This is desirable in practice: the gain from VSR is robust, while the broad plateau around the optimum indicates that VSI does not depend on fragile

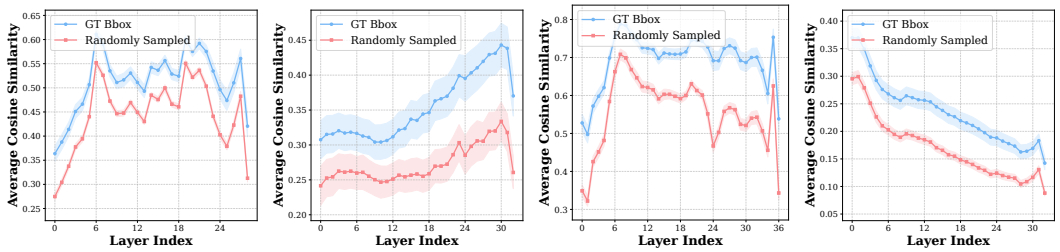


LLaVA-1.5-7B InternVL3-8B Qwen2-VL-7B Qwen2.5-VL-7B

Figure 6: Sensitivity of the combined hallucination score to the weighting coefficient λ on the CHAIR-style MSCOCO setting. Across different LVLMs, performance improves markedly when VSR is given a non-zero weight, then remains broadly stable across a moderate range of λ. This indicates that the benefits of combining VSC and VSR are robust and do not rely on highly precise tuning of the coefficient.

hyperparameter tuning. At the same time, the monotonic improvement from λ=0 to moderate values further supports our central claim that hallucination detection benefits from jointly modeling both perception-level concept formation and decoding-time residual steering.

D.3 More Results on Observations

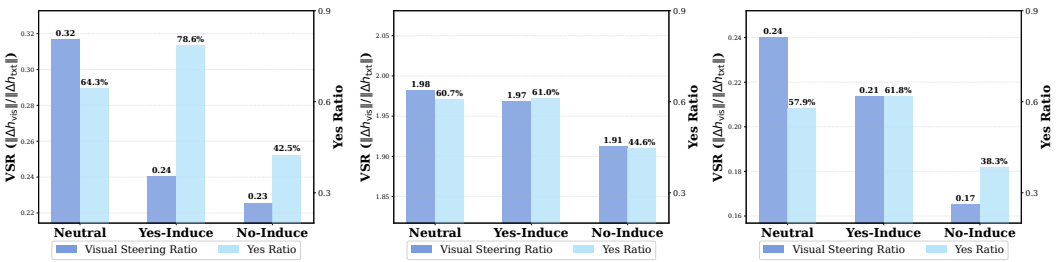


InternVL3-8B InstructBLIP Qwen3-VL-4B Shikra

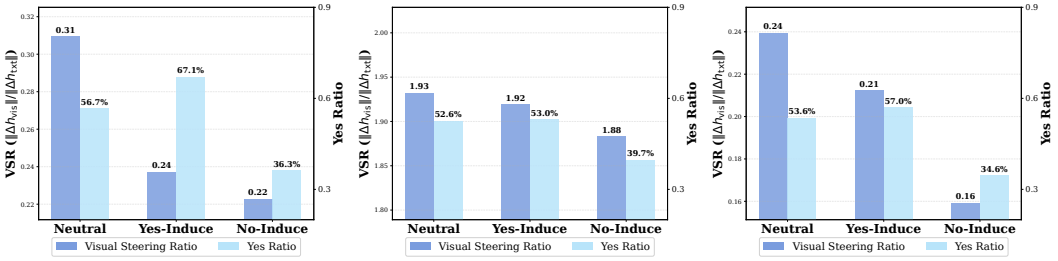
Figure 7: Additional observation results on visual semantic cohesion across LVLMs. Each panel shows the layer-wise pairwise cosine similarity of object-relevant visual tokens and randomly sampled visual tokens. Across InternVL3, InstructBLIP, Qwen3-VL, and Shikra, target-related tokens consistently exhibit stronger semantic cohesion than random visual tokens, supporting that correctly grounded visual concepts are associated with concentrated visual representations.

To further validate Observation 1, we provide additional layer-wise semantic cohesion visualizations across diverse LVLMs in Figure 7. The results consistently show that target-related visual tokens have higher pairwise cosine similarity than randomly sampled visual tokens across most layers. This trend holds for both LLaMA-based and Qwen-based architectures, indicating that the emergence of coherent visual semantics is a common internal property of grounded perception rather than a model-specific artifact. These findings further support our use of visual semantic cohesion as a diagnostic signal for hallucination detection.

We further supplement Observation 2 with norm-based visualizations across models and POPE splits in Figure 8. The results show that prompt-induced textual priors substantially change the model's response tendency, as reflected by the Yes ratio. Meanwhile, visual contributions are consistently weakened relative to textual contributions under biased prompts, especially in more challenging split settings. This confirms that hallucinations can arise not only from weak visual representations, but also from decoding-time dominance of textual priors in the residual stream. Together, these additional observations demonstrate that the two failure modes captured by VSI are stable across different LVLM architectures and hallucination regimes.



LLaVA-1.5-7B, Adversarial | InstructBLIP, Adversarial | Qwen3-VL-4B, Adversarial



LLaVA-1.5-7B, Popular | InstructBLIP, Popular | Qwen3-VL-4B, Popular

Figure 8: Additional observation results on visual steering under different POPE splits. Each panel compares VSR under Neutral, Yes-Induce, and No-Induce prompts, together with the corresponding Yes ratio. Across LLaVA-1.5-7B, InstructBLIP, and Qwen3-VL-4B on Adversarial and Popular splits, induced textual priors increase the tendency to answer "Yes" or "No" while suppressing the relative dominance of visual contributions, supporting the role of residual-stream steering in hallucination.

D.4 Controlled Non-assertive Intervention on Textual Priors

The Yes-Induce and No-Induce prompts used in the original motivation experiment explicitly state the desired answer. Consequently, their effects may reflect both textual-prior competition and direct instruction following. To separate these factors, we conduct a controlled intervention in which none of the prompts reveals whether the queried object is present. The experiment uses LLaVA-1.5-7B and 500 ground-truth negative examples from each of the Popular and Adversarial splits of POPE. Because the correct answer is "No" for every example, accuracy is the proportion of responses that correctly reject the absent queried object.

We compare three prompt conditions. **Original** uses the unmodified POPE question, such as "Is there a chair in the image?" **Image-guided** adds a neutral grounding instruction—"Please answer based on the image"—before the original question. **Misleading prior** prepends a fluent, non-assertive caption generated by Qwen-VL-Max that mentions a semantically related object but does not state the answer to the queried question. For example, the context may state "The scene shows a table" before asking whether a chair is present. Thus, the misleading context introduces a plausible association without directly asserting that the target object exists.

Prompt wording and length can change the absolute scale of residual contributions. We therefore report the visual contribution norm $\|\Delta\mathbf{h}_{\text{img}}\|_2$, textual contribution norm $\|\Delta\mathbf{h}_{\text{txt}}\|_2$, and VSR relative to the Original condition within each split. The absolute VSR of the Original condition is 0.31 on Popular and 0.37 on Adversarial. All contribution statistics are averaged over the 500 examples in the corresponding split.

Table 9 shows opposite effects for image-guided and misleading contexts. On Popular, the misleading prior reduces accuracy from 65.0% to 43.0%, suppresses the visual contribution to $\times 0.54$, and lowers VSR to $\times 0.48$. The same pattern holds on Adversarial, where accuracy decreases from 37.0% to 24.0%, the visual contribution falls to $\times 0.58$, and VSR falls to $\times 0.54$. The textual contribution increases only moderately in both cases, indicating that the VSR reduction is driven primarily by weakened visual steering rather than solely by an enlarged textual denominator.

Table 9: Controlled non-assertive intervention on POPE using LLaVA-1.5-7B. Contribution norms and VSR are normalized by the corresponding Original-condition mean within each split. The misleading captions introduce contextual bias without explicitly revealing the correct answer.

Split	Condition	Accuracy	Relative $\|\Delta\mathbf{h}_{\text{img}}\|_2$	Relative $\|\Delta\mathbf{h}_{\text{txt}}\|_2$	Relative VSR
Popular	Original	65.0%	$\times 1.00$	$\times 1.00$	$\times 1.00$
	Image-guided	78.0%	$\times 1.88$	$\times 1.01$	$\times 1.84$
	Misleading prior	43.0%	$\times 0.54$	$\times 1.11$	$\times 0.48$
Adversarial	Original	37.0%	$\times 1.00$	$\times 1.00$	$\times 1.00$
	Image-guided	56.0%	$\times 1.84$	$\times 0.99$	$\times 1.86$
	Misleading prior	24.0%	$\times 0.58$	$\times 1.08$	$\times 0.54$

Additional prompt text alone cannot explain this behavior. The Image-guided condition also lengthens the prompt, yet it improves accuracy to 78.0% and 56.0% on Popular and Adversarial, respectively. It nearly doubles the visual contribution while leaving the textual contribution approximately unchanged, producing corresponding VSR increases to $\times 1.84$ and $\times 1.86$. Because neither modified condition states the answer and the two similarly lengthened prompts produce opposite effects, the results substantially reduce the direct answer-copying and prompt-length confounds. They support a more precise interpretation of Observation 2: semantically misleading textual context can suppress visual steering and shift predictions away from image evidence, whereas an explicit request to ground the answer in the image strengthens visual influence.

D.5 Separating Attention Magnitude from Directional Coherence

VSC is intentionally an attention-weighted cohesion statistic. Expanding its definition shows that it depends jointly on the total attention assigned to visual tokens and the directional agreement among the attended hidden states. To determine whether its empirical gain can be attributed to directional coherence alone, we evaluate two magnitude-invariant variants on LLaVA-1.5-7B using the CHAIR-style MSCOCO setting.

The **Top-k variant** selects the k=20 visual tokens receiving the highest attention from the generated object token and computes their unweighted mean pairwise cosine similarity. The **Top-p variant** first normalizes attention within the visual-token set, selects the smallest set of tokens whose cumulative attention mass reaches p=0.5, and computes the same unweighted mean pairwise similarity. In both variants, attention is used only for token selection: the resulting score is invariant to the total amount of attention assigned to the image. We compare these scores with SVAR, which captures visual attention magnitude without explicitly measuring semantic agreement, and with the complete VSC statistic.

Table 10: Ablation of visual attention magnitude and directional coherence on LLaVA-1.5-7B using MSCOCO. Top-k and Top-p remove total visual attention magnitude after selecting the relevant visual tokens.

Method	AUROC	AUPR
SVAR	73.11	87.64
Top-k coherence (k=20)	71.25	83.79
Top-p coherence (p=0.5)	72.98	84.39
VSC	**77.86**	**89.21**

The magnitude-invariant variants remain discriminative, confirming that the directional organization of attended visual tokens contains hallucination-relevant information. However, neither variant exceeds SVAR, and both are clearly weaker than the complete VSC score. VSC improves over SVAR by 4.75 AUROC points and 1.57 AUPR points, while also outperforming the stronger magnitude-invariant Top-p variant by 4.88 AUROC points and 4.82 AUPR points. These results clarify that the gain should not be attributed to coherence alone. Rather, VSC benefits from coupling *evidence strength* with *semantic consistency*: visual attention magnitude indicates how much image evidence is used, whereas directional cohesion indicates whether that evidence consistently supports a shared visual concept. Removing either factor produces a less complete grounding signal.

D.6 Cross-model Proxy Evaluation

Direct VSI auditing requires access to the evaluated generator's hidden states, attention weights, and projection matrices. We therefore study a complementary setting in which an open-weight model serves as a proxy evaluator for responses produced by another LVLM. The proxy receives the same image together with the generated response, re-encodes the pair, and computes VSI from its own internal representations. This setup does not reconstruct the generator's hidden computation; instead, it asks whether an independent multimodal model finds coherent and sufficiently strong visual support for the generated object mention.

We use Qwen3-VL-4B as the proxy and evaluate responses generated by Qwen2-VL-7B and InternVL3-8B. For a controlled comparison, all detection methods in Table 11 are computed using the proxy model's outputs or internal states for the same image–response pair. Thus, differences reflect the diagnostic score rather than access to different evaluators.

Table 11: Cross-model hallucination detection using Qwen3-VL-4B as an open-weight proxy evaluator. The generator supplies the response, while all detection scores are obtained from the proxy's evaluation of the image–response pair.

Generator	Detection method	AUROC	AUPR
Qwen2-VL-7B	Entropy	53.21	79.25
	NLL	52.24	77.75
	Intern-Conf	73.47	87.97
	GLSim	69.49	82.88
	VSI	**77.61**	**89.97**
InternVL3-8B	Entropy	55.22	87.88
	NLL	52.33	86.31
	Intern-Conf	67.39	91.33
	GLSim	68.03	85.33
	VSI	**78.32**	**92.96**

Proxy-derived VSI achieves the best AUROC and AUPR for both generators. For Qwen2-VL-7B outputs, VSI reaches 77.61 AUROC and 89.97 AUPR, exceeding the strongest baseline by 4.14 and 2.00 points, respectively. For InternVL3-8B outputs, it reaches 78.32 AUROC and 92.96 AUPR, improving over the strongest corresponding baselines by 10.29 and 1.63 points. The result provides preliminary evidence that an open-weight evaluator's visual-integrity signals can transfer across generator architectures.

This proxy setting should be distinguished from direct white-box diagnosis. Direct VSI can identify how visual information is organized and propagated inside the original generator, whereas proxy VSI only assesses whether the generated claim appears visually supported according to the evaluator. Its reliability may depend on the proxy's own recognition ability, biases, and architectural compatibility. Nevertheless, the experiment suggests a practical route for assessing outputs from systems whose internal activations are unavailable, while retaining a more mechanistic grounding signal than output confidence alone.

D.7 Ablation on Token Selection for Multi-token Object Names

As reported in Table 12, the first-token strategy achieves the best performance on LLaVA-1.5-7B and remains competitive on Qwen2-VL-7B, where the last-token strategy performs slightly better. Given its simplicity and consistently strong performance, we adopt the first-token strategy by default.

Table 12: Ablation on token selection strategies for multi-token object names on MSCOCO. We compare computing the VSI score using the first token, the last token, and the average over all tokens.

Model	First	Last	Average
LLaVA-1.5-7B	78.99	78.73	78.74
Qwen2-VL-7B	79.88	79.97	79.24

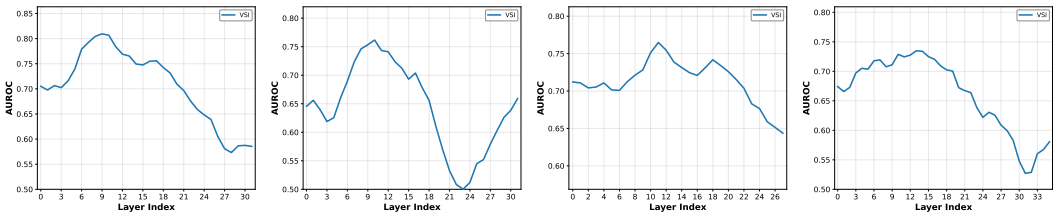
D.8 Layer-wise Stability and Depth-dependent Behavior

We further analyze how VSI varies with the layer used to extract internal visual signals. As shown in Figure 9, the strongest performance generally appears in the early-to-middle or middle layers rather than in the final layers. InstructBLIP peaks around layers 8–10, LLaVA-1.5-7B around layers 9–11, Qwen2.5-VL-7B around layer 11 with a secondary high-performing region near layer 18, and Qwen3-VL-4B shows a broader plateau around layers 10–15. Although the exact peak differs across architectures, neighboring layers around these high-performing regions yield similar AUROC, suggesting that VSI is not dependent on a single fragile layer choice. This non-final-layer behavior is consistent with recent analyses showing that hallucination-related internal signals can be stronger in contextual or intermediate representations than in final-layer readouts (Phukan et al., 2025; Li et al., 2025b; Skean et al., 2025).

This trend is consistent with the visual information-flow perspective. In shallow layers, visual tokens have not yet been sufficiently integrated with the generated object token, so their representations may still reflect low-level or modality-specific features. Prior work has similarly observed a depth-wise transition from low-level visual details to more semantic and task-conditioned visual representations, and has shown that shallow or middle visual representations can outperform deeper representations on visually grounded tasks such as counting, localization, and spatial reasoning (Ganz et al., 2024; Chen et al., 2025). In intermediate layers, visual evidence becomes more semantic and query-aware, allowing VSI to better capture both coherent visual concept formation and effective visual steering.

In deeper layers, however, the generated token is increasingly dominated by autoregressive textual context and learned language priors, which can dilute or override visual contributions in the residual stream. This explains why later layers often show a clear performance drop, especially for LLaVA-1.5-7B and Qwen3-VL-4B. Such a late-layer degradation is consistent with evidence that LVLM generation can increasingly rely on textual context even for image-related tokens, and that hallucinated autoregressive context can snowball despite the presence of correct visual evidence (Min et al., 2025; Zhong et al., 2024).

Overall, the layer-wise results indicate that VSI is most effective at the stage where visual concepts have been sufficiently formed but have not yet been substantially overwritten by textual priors. This supports our interpretation that hallucination detection benefits from measuring visual information flow at intermediate depths, where grounded and hallucinated object mentions are most separable.



InstructBLIP LLaVA-1.5-7B Qwen2.5-VL-7B Qwen3-VL-4B

Figure 9: Layer-selection stability across four LVLMs. The results show that the performance of the proposed score remains broadly stable across nearby layer choices, indicating that VSI is not overly sensitive to the exact layer range used in practice.

D.9 Computational Overhead

Table 13: Estimated *scoring-stage* computational overhead of different hallucination detection scores, reported as proxy FLOP counts per evaluated object mention under cached model outputs. Lower is better. Intern-Conf is abbreviated as IC.

Method	Entropy	NLL	Intern-Conf (IC)	GLSim	SVAR	ContextLens	Ours (S_{VSC})	Ours (S_{VSR})
FLOPs	456.19K	1	152.06K	6.29M	262.14K	67.11M	3.47M	3.15M

To better understand the practical cost of different hallucination detectors, we further compare their *additional scoring-stage* overhead per evaluated object mention. This experiment complements the detection-performance results by revealing the efficiency–effectiveness trade-off across methods after a generation forward pass has already produced the logits, hidden states, and attention maps needed by the detector. Accordingly, Table 13 should be interpreted as a relative proxy for post-hoc scoring cost rather than as exact end-to-end generation FLOPs. For concreteness, the proxy FLOP estimation is instantiated using a Qwen2-VL-7B-style configuration with L=32 Transformer layers, hidden size D=4096, number of attention heads H=32, vocabulary size V=152,064, N_v=256 visual tokens, and an estimated N_t=100 text tokens. Here, L, D, and H follow the standard 7B-scale backbone configuration; V reflects the larger Qwen2-VL vocabulary that includes multilingual tokens; N_v corresponds to a typical visual-token count induced by a 224×224 image after patchification and visual encoding; and N_t is an estimate of the prompt length used during evaluation. For methods involving token selection, we set the selection budget to K=16. As shown in Table 13, NLL is by far the cheapest baseline, followed by Intern-Conf, SVAR, and Entropy, whereas analysis-heavy white-box methods such as GLSim and ContextLens incur substantially larger overhead. Our two proposed statistics occupy a favorable middle ground: both are markedly cheaper than ContextLens and GLSim, with VSR slightly cheaper than VSC. These results suggest that VSI improves interpretability and detection capability without introducing prohibitive additional scoring cost.

E Limitations and Future Directions

E.1 Limitations

Although VSI shows strong empirical performance and offers a more interpretable view of hallucination formation, the current study still has several limitations. First, our evaluation mainly focuses on object-level hallucination detection in image captioning and closely related prompting settings. It remains to be further verified how well the proposed signals transfer to broader multimodal tasks such as long-form visual question answering, multi-image reasoning, or video understanding. Second, while VSI is training-free, it is still a white-box analysis method that requires access to internal activations, attention patterns, and residual updates. This may limit its applicability to closed-source or API-only LVLMs. Finally, our method is primarily designed as a diagnostic tool rather than a direct mitigation strategy. A high-quality diagnosis can help reveal where hallucination arises, but it does not by itself prevent the model from generating hallucinated content.

E.2 Future Directions

There are several promising directions for future work. One natural extension is to generalize VSI beyond object hallucinations to richer forms of multimodal errors, such as attribute hallucinations, relation hallucinations, and reasoning failures that unfold over longer generation horizons. Another direction is to investigate whether the proposed mechanistic signals can be used not only for post hoc detection, but also for online intervention, for example by triggering decoding-time correction, adaptive refusal, or targeted re-attention when visual grounding appears weak. It would also be valuable to study lighter-weight approximations of VSI that retain its interpretability while reducing the need for full internal access, which could broaden applicability to more practical deployment settings. Finally, we hope the framework can serve as a starting point for a more general science of multimodal model diagnosis, where internal information flow is used not only to detect failures, but also to better understand how reliable visual grounding emerges in large vision-language models.

F Broader Impacts

This work studies hallucination detection in large vision-language models and is intended to improve the reliability and interpretability of multimodal systems. A potential positive impact is that better diagnosis of hallucination can support safer deployment in settings where visual grounding matters, such as assistive technologies, information retrieval, scientific analysis, and other applications that require users to trust model-generated descriptions. In addition, the mechanistic perspective adopted in this paper may help researchers better understand when a model is relying on visual evidence and when it is instead being overly driven by language priors.

At the same time, the work also has limitations and possible indirect risks. Because VSI is a white-box diagnostic framework, it may be mainly accessible to researchers and developers with internal model access, which could widen the gap between highly transparent systems and black-box commercial systems. More broadly, improved diagnosis does not automatically eliminate hallucinations; if users over-interpret detector scores as a guarantee of correctness, this could create a false sense of security. There is also a general dual-use consideration: tools that analyze internal model behavior may eventually be used not only to detect failures, but also to better understand how to preserve or manipulate model outputs in unintended ways. We therefore view VSI primarily as a reliability-oriented research tool and believe it should be used together with careful evaluation, transparent reporting, and appropriate human oversight in high-stakes applications.

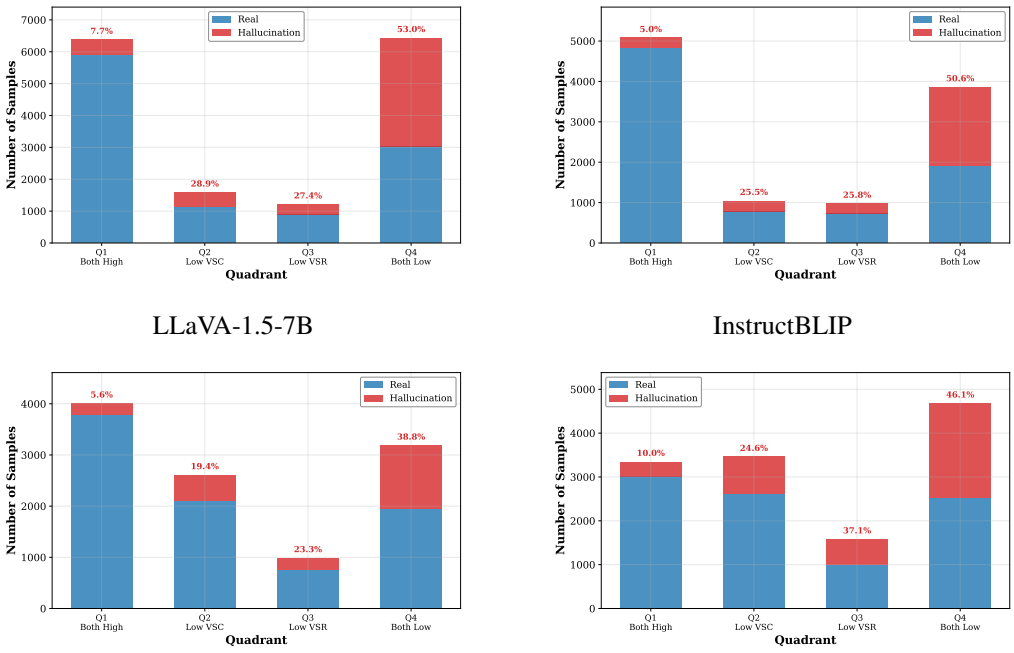
G VISUALIZATIONS

G.1 DIAGNOSTIC QUADRANT ANALYSIS OF VISUAL SIGNAL INTEGRITY

Figure 10 visualizes the diagnostic space induced by the two components of VSI on the MSCOCO benchmark. Each generated object mention is assigned to one of four quadrants according to whether its Visual Semantic Cohesion (VSC) and Visual Steering Ratio (VSR) are high or low: Q1 represents both high cohesion and strong visual steering, Q2 corresponds to low VSC, Q3 corresponds to low VSR, and Q4 indicates that both signals are weak. Across different LVLMs, grounded object mentions are concentrated more heavily in Q1, suggesting that reliable generations are typically supported by both coherent visual evidence and vision-dominated residual steering. In contrast, hallucinated mentions appear more frequently in the failure-related quadrants, especially Q4, where the model neither forms a cohesive visual concept nor preserves sufficient visual influence during decoding. The intermediate quadrants further reveal two distinct hallucination modes: Q2 reflects perception-level degradation, where attended visual tokens are semantically scattered, while Q3 reflects decoding-time prior hijacking, where coherent visual evidence is overridden by textual contributions. These results support the interpretation that VSI provides not only a scalar hallucination score, but also a fine-grained diagnostic view of where visual information flow breaks down.

G.2 KDE VISUALIZATION ACROSS MODELS

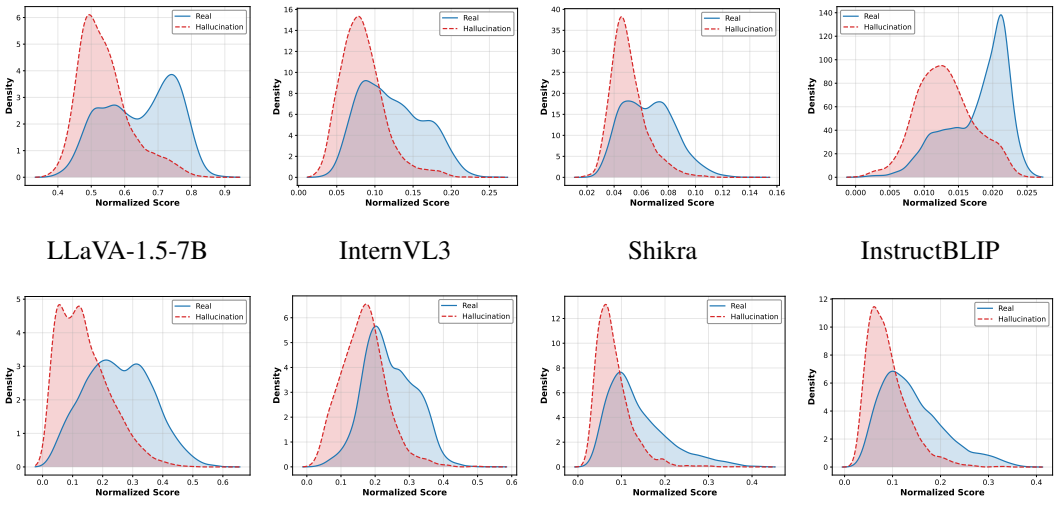
To further examine how the proposed mechanistic signals separate grounded and hallucinated object mentions across model families, we visualize the kernel density estimation (KDE) distributions for the combined VSI score on the CHAIR-style MSCOCO setting. Each subplot corresponds to one LVLM and compares the score distributions of grounded and hallucinated object mentions. A clearer separation between the two density curves indicates that the detector provides a more discriminative ranking signal for object-level hallucination detection. Overall, these visualizations show that the proposed score exhibits a consistent tendency to assign larger values to grounded mentions than to hallucinated ones, while the degree of separation still varies across model architectures and hallucination regimes.



LLaVA-1.5-7B InstructBLIP

Qwen2.5-VL-7B Qwen3-VL-4B

Figure 10: Visualization of the four diagnostic quadrants induced by VSI on MSCOCO across different LVLMs. Each generated object mention is assigned according to whether its Visual Semantic Cohesion (VSC) and Visual Steering Ratio (VSR) are high or low: Q1 indicates both high VSC and high VSR, Q2 indicates low VSC but high VSR, Q3 indicates high VSC but low VSR, and Q4 indicates that both signals are weak. The resulting distributions highlight how grounded mentions concentrate more heavily in Q1, whereas hallucinated mentions appear more frequently in the failure-related quadrants.



LLaVA-1.5-7B InternVL3 Shikra InstructBLIP

Qwen2-VL-7B Qwen2.5-VL-7B Qwen3-VL-4B Qwen3-VL-8B

Figure 11: Kernel density estimation (KDE) visualizations of the combined VSI score for grounded and hallucinated object mentions across different LVLMs on the CHAIR-style MSCOCO setting. Better separation between the two curves indicates stronger discriminative ability.

G.3 Visualization of Benchmark Cases

To complement the quantitative results, we provide qualitative visualizations of the proposed signals across different LVLMs. The goal is to examine whether VSC and VSR exhibit consistent diagnostic behaviors under hallucinated versus grounded predictions across model architectures.

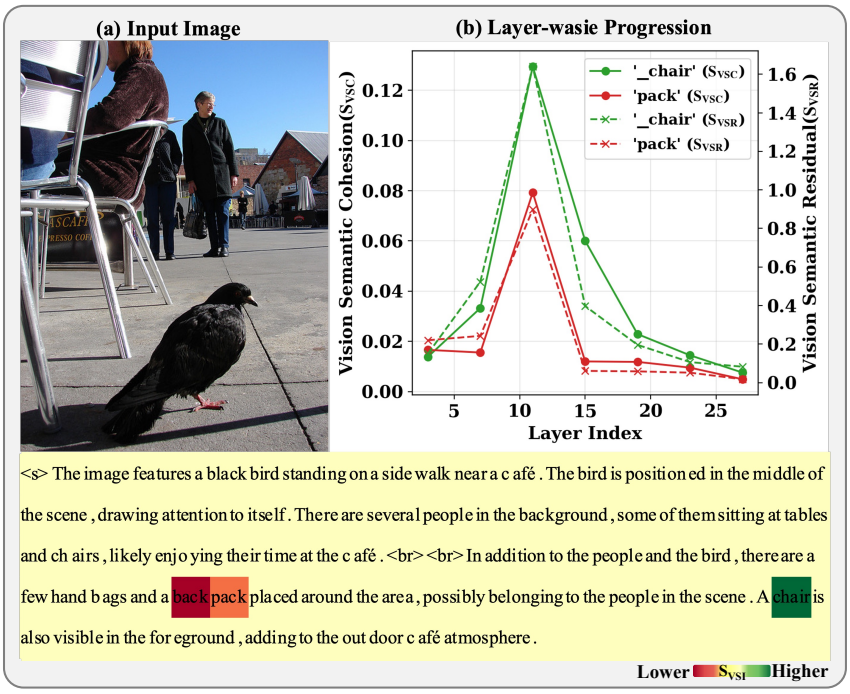


Figure 12: Qualitative CHAIR cases for **LLaVA-1.5-7B**.

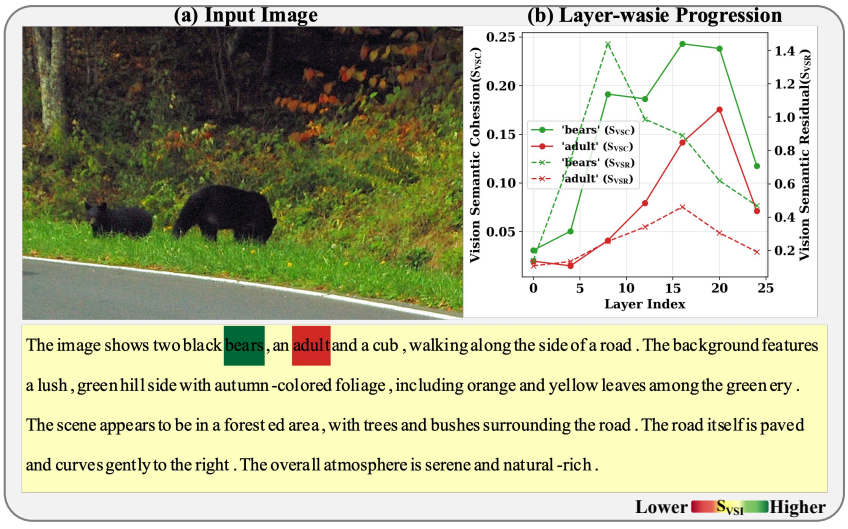


Figure 13: Qualitative CHAIR cases for **InternVL3**.

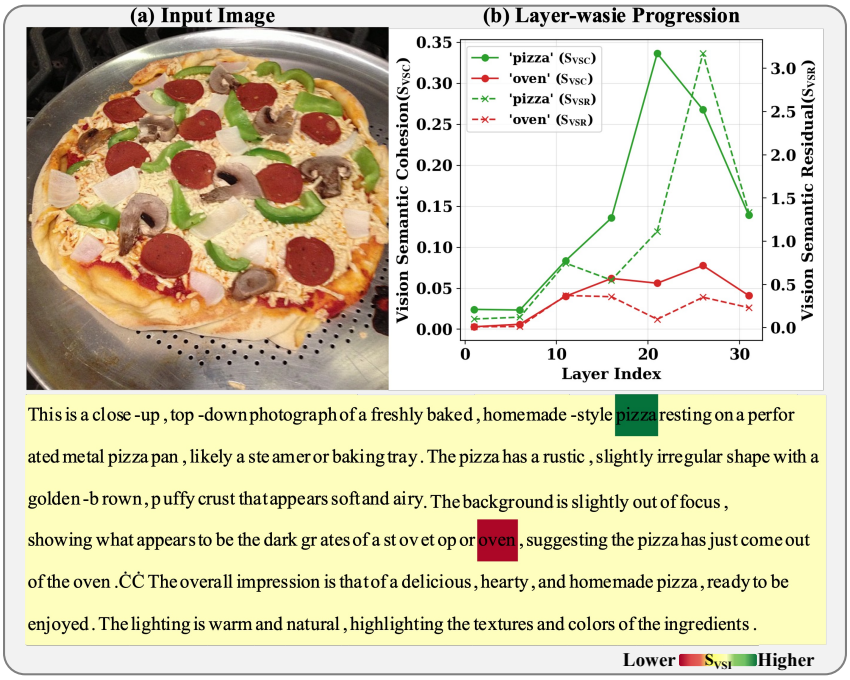


Figure 14: Qualitative CHAIR cases for **Qwen3-VL**.

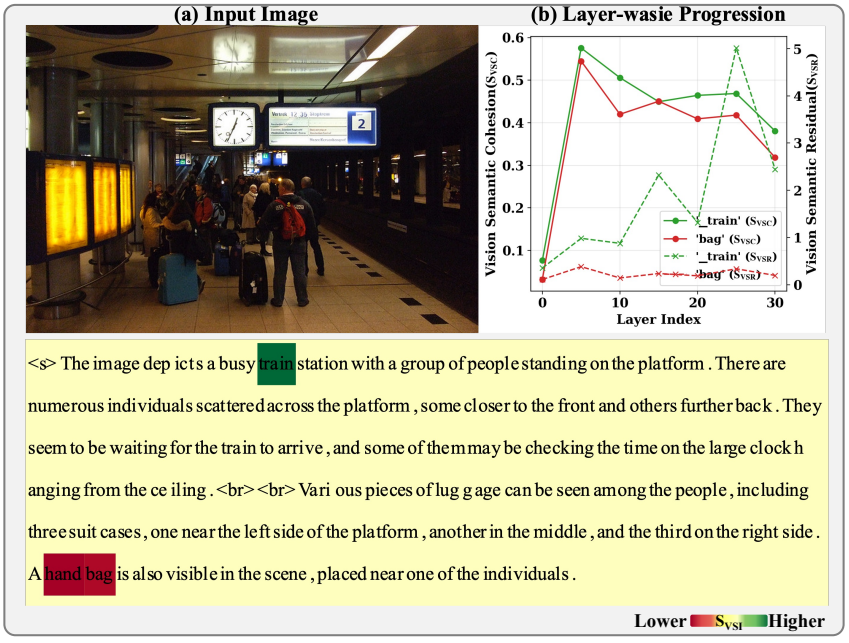


Figure 15: Qualitative CHAIR cases for **Shikra**.