WHAT MATTERS, WHAT MISLEADS: VISUAL ATTRIBUTION AND HALLUCINATION MITIGATION IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) can produce sophisticated multimodal predictions, yet it remains unclear which visual evidence actually drives those predictions. Revealing this evidence is fundamental to understanding LVLM behavior and enabling targeted interventions on the visual information that contributes to hallucinations. However, existing attribution methods based on attention, gradients, or perturbations can overlook computational pathways, provide unreliable estimates, or require costly and context-dependent interventions, motivating a direct approach to measuring the contribution of localized visual-token subsets. In this work, we introduce VIRAH, a training-free framework that retains a local neighborhood of visual tokens at a chosen vision-encoder depth, removes other visual tokens, and propagates the retained representations through the remaining model to measure the evidence supported by selected regions. The resulting output distribution serves as a self-probe of the visual evidence supported by selected regions. We evaluate VIRAH for visual attribution, where the probability of an affirmative answer to a target concept query provides the regional attribution score. We further demonstrate that this self-probing capability can be directly leveraged to develop a training-free hallucination mitigation method, using regional object-identification entropy to identify unreliable visual tokens for removal during captioning. Across various LVLMs (Qwen, InternVL, and LLaVA) and multiple benchmarks, VIRAH consistently outperforms existing attribution methods. Specifically, on Qwen2.5-VL-7B, it improves Insertion$-$Deletion by 46.08% and AUPR_1 by 11.94% over the strongest attribution baseline on MS-COCO. For hallucination mitigation on MS-COCO, this simple training-free intervention achieves performance competitive with or surpassing recent methods, reducing CHAIR_i by up to 1.90% and CHAIR_s by up to 2.20% while improving Recall by up to 4.04%.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) integrate visual and textual information to support visual question answering, image description, and multimodal reasoning Li et al. (2023a); Dai et al. (2023); Liu et al. (2024). As these models are increasingly used for complex multimodal tasks, understanding which image regions contribute to their predictions is crucial for interpreting their decisions and assessing whether their outputs are grounded in relevant visual evidence. Visual attribution provides a natural means to identify such evidence. This capability is particularly important given that LVLMs often generate content unsupported by the input image Rohrbach et al. (2018); Li et al. (2023b); Wang et al. (2023), as identifying the visual evidence underlying predictions can enable targeted interventions to mitigate hallucinations while preserving valid visual information.

Existing visual attribution methods estimate regional importance from attention weights, gradients, or input perturbations. Attention-based scores do not capture all computational paths through a Transformer, while gradient-based methods provide local sensitivity estimates and can be affected by imbalanced gradient flow Chefer et al. (2021); Zhang et al. (2025); Mehri et al. (2025). Perturbation methods instead mask, replace, or reveal image regions and measure the resulting change in the model output Petsiuk et al. (2018); Chen et al. (2026). They often require repeated encoding of modified images, and their scores can depend on the baseline or visual context created by the perturbation. Recent LVLM-specific methods also estimate patch contributions through cooperative-game formulations Zhu et al. (2025). These limitations lead to three questions: whether local

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-token subsets can support output-level attribution, how the intervention depth changes the evidence assigned to a region, and whether the same interface can guide hallucination mitigation.

Transformer-based LVLMs provide a natural interface for such interventions. Prior intervention-based studies have used controlled changes to internal activations or visual tokens to test their functional role in model predictions Vig et al. (2020); Meng et al. (2022); Izadi et al. (2026); Neo et al. (2025). We introduce VIRAH, **V**isual **I**ntervention for **R**egional **A**ttribution and **H**allucination mitigation, which retains a local neighborhood of visual tokens at a chosen vision-encoder depth, removes the remaining tokens, and propagates the retained representations through the rest of the model. The resulting decoder distribution measures what the selected visual evidence supports. For visual attribution, we ask whether a target concept is present and use the affirmative-answer probability as the regional attribution score. The local neighborhoods can be evaluated in batches, avoiding a separate image-encoding pass for every perturbed region.

The same intervention also allows us to examine the object-level information supported by localized visual representations. We use the LVLM as a semantic probe by presenting each retained neighborhood with an object-recognition question and measuring the predictive entropy of its response. A high-entropy response indicates that the selected neighborhood provides weak or unstable support for an object prediction. Prior work uses predictive uncertainty for visual explanation or unreliable-generation detection Patro et al. (2019); Farquhar et al. (2024), while training-free mitigation methods modify visual inputs, attention, token scores, or hidden activations Leng et al. (2024); Huang et al. (2024); Chen et al. (2024); Jung et al. (2025); Li et al. (2025a); Wang et al. (2026a). However, these methods do not use localized, layer-controlled token retention as a common interface for attribution and mitigation. We spatialize the uncertainty signal by clustering the regional entropy scores and selecting the highest-entropy cluster as potentially unreliable evidence. In a separate full-image captioning pass, we knock out the corresponding visual tokens at a chosen encoder depth. This training-free intervention reduces object hallucination without external detectors, extra supervision, or changes to the decoding objective, and improves over the evaluated hallucination-mitigation methods.

To study how visual context affects these measurements, VIRAH supports intervention at different vision-encoder depths. At the input level, a retained neighborhood is encoded without information from the excluded patches. At later depths, all visual tokens interact before the intervention, so the retained representations include context accumulated from the full image. Removing tokens at a later depth does not erase information that they have already transferred to the surviving tokens. It instead measures the evidence available in the retained representations at that stage of processing. Input-level intervention already provides strong attribution performance, while intervention on contextualized embeddings yields further gains, especially in segmentation-based localization. This result is consistent with prior evidence that visual information changes and is redistributed across LVLM layers Zhang et al. (2025); Neo et al. (2025).

We evaluate VIRAH across several LVLM families and examine both object- and relation-conditioned attribution. We assess attribution using faithfulness and localization metrics, and evaluate hallucination mitigation using CHAIR and Recall. Across these settings, VIRAH outperforms the evaluated baselines, supporting visual-token intervention as a common mechanism for explaining and controlling visual evidence in LVLMs.

In conclusion, our contributions are summarized as follows:

- **Layer-controlled visual-token intervention.** We introduce VIRAH, a framework that retains spatially grounded visual-token subsets at a chosen vision-encoder depth and measures the evidence they support through the model's output distribution. This design supports the analysis of both local and contextualized visual representations.
- **Target-conditioned visual attribution.** We formulate regional attribution as the probability of an affirmative response under retained local visual evidence. The formulation supports batched evaluation and produces more faithful and spatially precise attribution maps across the evaluated LVLMs and benchmarks.
- **Uncertainty-guided hallucination mitigation.** We derive regional predictive uncertainty by probing localized visual-token subsets and use this signal to select tokens for knockout during full-image generation. This training-free procedure reduces object hallucination while preserving valid visual content, without external detectors or additional supervision.

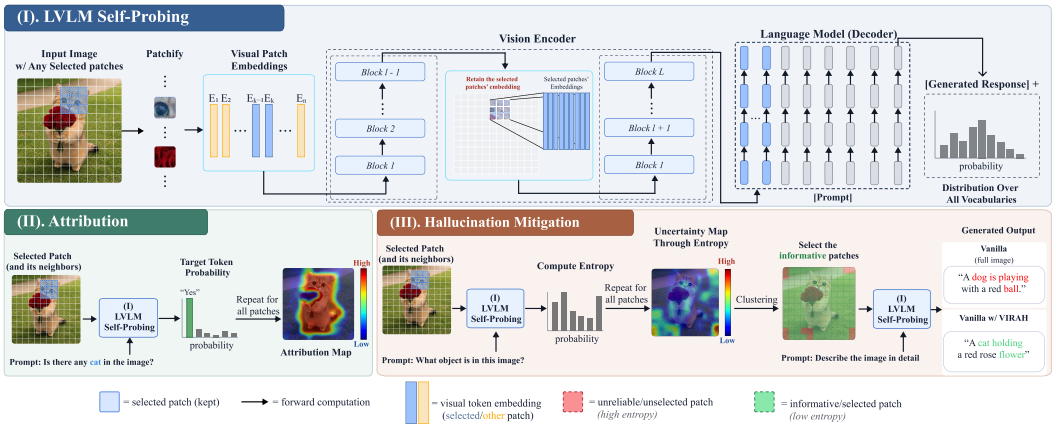

Figure 1: Overview of VIRAH. (I) For a selected patch subset, VIRAH retains the corresponding embeddings at a chosen vision-encoder layer and propagates them through the remaining vision blocks and language decoder to obtain a vocabulary distribution. (II) Each patch is selected in turn, and its target-token probability defines the attribution score, yielding a target-specific attribution map. (III) Predictive entropy estimates patch uncertainty; high-entropy patches are filtered as unreliable, while low-entropy patches are retained and fed back into (I) as the selected subset for final generation, yielding a response with reduced visual hallucination.

2 Related Work

2.1 Attribution

Attribution methods for vision transformers and vision-language models can be broadly divided into four lines: activation-based methods that construct spatial relevance maps from internal activations (Li et al., 2025b); gradient-based methods that back-propagate a target score through the vision backbone, following the class-activation-mapping paradigm (Zhang et al., 2025); perturbation-based methods that optimize masks or evaluate externally segmented regions under modified visual inputs (Xing et al., 2025; Chen et al., 2026); and causal methods, which are rooted in mechanistic interpretability for unimodal language models and formulate attribution as a controlled intervention on internal activations through activation patching or causal mediation analysis (Vig et al., 2020; Meng et al., 2022; Wang et al., 2022). This approach has also recently been used for ViT attribution by inserting source-image patch activations into a neutral target context at intermediate layers (Izadi et al., 2026).

VIRAH follows the causal line but differs in how the intervention is performed. EAGLE (Chen et al., 2026) relies on externally generated superpixels and repeated per-region queries, introducing segmentation-dependent bias, potential region-generation errors, objective–metric alignment with insertion–deletion evaluation, and high computational cost, whereas VIRAH avoids these limitations by operating directly on native visual tokens without an external region generator or metric-specific optimization. Gradient-based methods (Zhang et al., 2025) estimate local sensitivity and can be affected by gradient saturation or vanishing gradients; VIRAH instead measures the effect of an explicit intervention on the model output. Activation- and mask-optimization methods (Li et al., 2025b; Xing et al., 2025) derive attribution from activation maps or optimized masks, while VIRAH uses a controlled visual-token intervention. CAAP (Izadi et al., 2026) patches source-image activations into a neutral target context in a ViT classifier, whereas VIRAH requires no source-target pair and propagates retained local-token neighborhoods through the projector and language decoder. The same intervention also supports hallucination mitigation, which CAAP does not address. Classical causal-mediation methods (Vig et al., 2020; Meng et al., 2022; Wang et al., 2022) typically compare reference and corrupted forward passes. VIRAH instead performs spatial, patch-level attribution without a reference image or second clean run, making it lighter while retaining the intervention-based formulation.

2.2 Hallucination

Existing hallucination work can be divided into three strategies. Decoding-time methods intervene without modifying internal representations: contrastive decoding compares outputs from the original

and a visually degraded input (Leng et al., 2024), while attention recalibration reweights attention during decoding (Huang et al., 2024). A second group detects hallucination from internal signals, including logit-lens and saliency analysis, mutual information, instruction embeddings, token-to-patch grounding, layer-level causal attribution, and caption-trajectory and latent-identifier analysis. Some of these methods couple the signal with a lightweight, post hoc masking or reweighting step (Jung et al., 2025; Wang et al., 2026a; Zhang et al., 2026a; Hasani et al., 2026), while others remain purely diagnostic (Hoang-Xuan et al., 2026; Lai et al., 2026; Nguyen et al., 2026; Afriat et al., 2026). A third, more recent group intervenes causally on internal computation to mitigate hallucination directly, by suppressing attention heads responsible for hallucinated text (Yang et al., 2025; Ji et al., 2026; Cheng et al., 2026), steering hidden activations away from hallucination-inducing directions (Su et al., 2025; Wang et al., 2026b; Dang et al., 2026), intervening at a single decoding layer or prefill step (Wan et al., 2025; Zhang et al., 2026b), or jointly across the pathways that produce hallucination (Qian et al., 2026).

VIRAH is most closely related to this third group, but does not require components external to its attribution mechanism. Decoding-time methods (Leng et al., 2024; Huang et al., 2024) act post hoc without localizing unreliable evidence; purely diagnostic methods (Hoang-Xuan et al., 2026; Lai et al., 2026; Nguyen et al., 2026; Afriat et al., 2026) leave mitigation to a separate procedure; and diagnosis-coupled methods (Jung et al., 2025; Wang et al., 2026a; Zhang et al., 2026a; Hasani et al., 2026) use a masking rule external to the mechanism that produces the signal. Causal-intervention methods (Yang et al., 2025; Ji et al., 2026; Cheng et al., 2026; Su et al., 2025; Wang et al., 2026b; Dang et al., 2026; Wan et al., 2025; Zhang et al., 2026b; Qian et al., 2026) still rely on machinery estimated or trained separately from the forward pass, such as learned head classifiers or gating parameters, calibration sets or contrastive reference pairs, and auxiliary entropy or pathway heuristics. VIRAH instead derives the hallucination signal from the same patch-level knockout used for attribution. Diagnosis and mitigation therefore use the same mechanism, with no auxiliary steering vectors, separate causal-effect model, or external contrastive data, at negligible cost beyond attribution itself.

3 METHOD

3.1 OVERVIEW

Transformer-based LVLMs provide the flexibility to process arbitrary subsets of visual tokens, enabling direct evaluation of any selected set of image patches. We exploit this property by selecting, for each image location, a local subset of patches, such as a center patch and its surrounding patches. After several vision-encoder layers, only the selected tokens that have already been contextualized by the full image are propagated through the remaining vision blocks and the language decoder, yielding a probability distribution over the output vocabulary (Figure 1, I).

We use this LVLM self-probing for two purposes. First, for visual attribution, the diagnostic prompt asks whether a target object is present in the image. For each selected local region, we compute the probability that the model answers "yes" and assign this probability to the center patch of that region. Repeating this self-probing procedure across image locations yields a spatial attribution map indicating which regions provide evidence for the target object (Figure 1, II).

Second, for hallucination analysis, the diagnostic prompt asks the model to identify the object visible from the selected local region. Rather than measuring the probability of a fixed answer, we measure the entropy of the model's response distribution. High entropy indicates that the selected region does not support a stable object-level prediction. We cluster these regional uncertainty scores within each image and select the cluster with the high entropy of the model's response distribution as a set of candidate unreliable regions. Based on this uncertainty signal, we use the detected high-entropy regions for hallucination mitigation. During caption generation, we perform a separate full-image forward pass and apply visual-token knockout to the patch tokens assigned to this cluster at the chosen encoder depth. The caption is then generated from this uncertainty-guided intervention (Figure 1, III).

3.2 LVLM SELF-PROBING

The common operation underlying both attribution and hallucination analysis is LVLM self-probing. Given an image and a selected subset of visual tokens, we ask the model a diagnostic question and use the resulting output distribution as a measurement of what the selected visual evidence supports.

Let x be an input image and $\mathcal{P} = \{p_1, \ldots, p_N\}$ denote its spatial visual patches. For each center patch p_i, we define a local visual region $\mathcal{R}_i$ consisting of p_i and a set of neighboring patches determined by a predefined spatial neighborhood rule. Boundary regions are clipped to the image grid.

The vision encoder maps the image to a sequence of visual representations. At a chosen encoder depth, we retain the selected visual tokens in $\mathcal{R}_i$ and propagate the retained tokens through the remaining model components. This produces a regional visual representation V_i for the selected local region at the final layer of the vision encoder.

Given a diagnostic prompt Q, the language decoder produces a distribution over the next token:

$$\pi_i(v \mid Q) = p(v \mid V_i, Q), \tag{1}$$

where v is a vocabulary token. We use this distribution in two ways. For visual attribution, we measure the probability assigned to a target answer. For hallucination analysis, we use predictive uncertainty to identify ambiguous regions.

Unlike input-perturbation methods that repeatedly encode separately modified full images, VIRAH reuses the visual representations computed up to the intervention depth. Each regional evaluation then processes only the selected tokens through the subsequent vision blocks and language decoder. These shortened sequences can be batched and evaluated in parallel, reducing redundant computation and enabling efficient evaluation of multiple regions within an image.

3.3 Target-Token Attribution

For attribution, the diagnostic prompt asks whether a target object c is present: "Is there a {object_label} in the image or not? Answer with exactly one word: yes or no."

Let $Q_{\text{attr}}(c)$ denote this prompt. For each local region $\mathcal{R}_i$, we compute the distribution $\pi_i(v \mid Q_{\text{attr}}(c))$ and use the probability assigned to single-token variants of "yes" as the attribution score:

$$A(p_i) = \sum_{v \in \mathcal{Y}_{\text{yes}}} \pi_i(v \mid Q_{\text{attr}}(c)), \tag{2}$$

where $\mathcal{Y}_{\text{yes}}$ is the set of accepted single-token forms of "yes".

The score $A(p_i)$ is assigned to the center patch p_i, although the model receives the full local region $\mathcal{R}_i$. Evaluating all center patches produces a vector of scores

$$A = [A(p_1), \ldots, A(p_N)] \in [0, 1]^N, \tag{3}$$

which is arranged on the image grid to form the attribution map. Larger values indicate stronger regional evidence for the target object's presence.

3.4 Regional Uncertainty and Hallucination Mitigation

We use regional predictive uncertainty to identify visual evidence that may contribute to hallucinated object descriptions. The underlying hypothesis is that regions providing ambiguous evidence for object recognition may encourage the decoder to rely on linguistic priors. We therefore estimate uncertainty for each local region and use these estimates to select visual tokens for knockout during caption generation.

For each local region $\mathcal{R}_i$, we obtain its visual representation V_i through the regional intervention described above and prompt the model with "What object exists in the image?" Let Q_{obj} denote this prompt. At generation step s, the model predicts

$$\pi_i^{(s)}(v) = p(v \mid V_i, Q_{\text{obj}}, \hat{y}_{i,<s}), \tag{4}$$

where $\hat{y}_{i,<s}$ denotes the preceding generated tokens. We measure predictive uncertainty using entropy:

$$H_i^{(s)} = -\sum_{v \in \mathcal{V}} \pi_i^{(s)}(v) \log_2 \pi_i^{(s)}(v), \tag{5}$$

where $\mathcal{V}$ is the model vocabulary. The score assigned to the center patch p_i is

$$U(p_i) = \frac{1}{T_i} \sum_{s=1}^{T_i} H_i^{(s)}, \tag{6}$$

where T_i is the number of evaluated generation steps. Larger values indicate less concentrated predictive distributions when the model identifies an object from the selected region.

We apply K-means to these scalar uncertainty scores independently within each image. Let $g_i \in \{1, \ldots, K\}$ denote the cluster assignment of patch p_i, and let μ_k denote the mean uncertainty of cluster k. We select

$$k^\star = \arg \max_{k \in \{1,\ldots,K\}} \mu_k, \qquad \mathcal{U} = \{i \mid g_i = k^\star\}. \tag{7}$$

The patches indexed by $\mathcal{U}$ are treated as potentially unreliable. This selection identifies uncertain visual evidence.

For caption generation, we reuse the cached final-layer encoder representations. Since knockout is applied after the final vision block, no second vision-encoder pass is required. Let $\mathcal{J}_\mathcal{U}$ denote the encoder-token indices corresponding to the unreliable patches. We remove these tokens from the sequence:

$$H_{\text{mit}}^{L_{\text{vis}}} = \left[h_j^{L_{\text{vis}}}(x) \right]_{j \notin \mathcal{J}_\mathcal{U}}. \tag{8}$$

The resulting sequence is projected into the language-model embedding space, and the decoder generates a caption:

$$V_{\text{mit}} = P\left(H_{\text{mit}}^{L_{\text{vis}}}\right), \qquad \hat{s}_{\text{mit}} = D(V_{\text{mit}}, Q_{\text{cap}}), \tag{9}$$

where Q_{cap} is the caption prompt and D denotes autoregressive decoding. Thus, uncertainty estimation and knockout use the same cached vision-encoder output; after knockout, the selected sequence is projected and passed to the decoder for caption generation.

4 Experiments

We investigate how effectively VIRAH can provide reliable visual attribution and support hallucination mitigation in LVLMs. For visual attribution, we compare VIRAH with existing attribution methods and evaluate whether the identified regions are both faithful and spatially aligned with the corresponding visual concepts. For hallucination mitigation, we examine whether the regional uncertainty estimated by VIRAH can be used to identify unreliable visual evidence and reduce unsupported object generation while preserving valid visual information.

4.1 Attribution Comparison

We first evaluate whether VIRAH accurately identifies the visual evidence responsible for an LVLM prediction. We compare its attribution maps against existing attribution approaches under identical model, dataset, and evaluation settings.

Evaluation metrics. We evaluate attribution from two perspectives: *faithfulness* and *localization*. Faithfulness measures whether highly attributed regions influence the model's confidence in the target token, while localization measures spatial agreement between the attribution map and the target's annotated region.

For faithfulness, we use Insertion and Deletion (Petsiuk et al., 2018). Regions are ranked by attribution score. Insertion progressively restores highly ranked regions to a baseline image; a faithful attribution yields a high Insertion AUC by rapidly recovering the target-token probability. Deletion progressively removes highly ranked regions from the original image; a faithful attribution yields a low Deletion AUC by rapidly reducing this probability. We also report Insertion$-$Deletion, where higher values indicate better faithfulness.

For localization, we use Pointing Game, AUPR_1, and AUPR_0 (Margolin et al., 2014). Pointing Game tests whether the center of the highest-ranked region lies inside the ground-truth target region. AUPR_1 and AUPR_0 measure the agreement of attribution scores with foreground and background pixels, respectively. Higher values indicate better localization for all three metrics.

Experimental setup. We conduct our attribution experiments on four LVLMs from different model families: Qwen2.5-VL-7B (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025), and LLaVA1.5-7B (Liu et al., 2024). Our main evaluations are performed on MS-COCO (Lin et al., 2015) and MMVP (Tong et al., 2024).

Table 1: Faithfulness results on MS-COCO (Lin et al., 2015) for Qwen2.5-VL-7B (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026), and InternVL3.5-8B (Wang et al., 2025). Bold indicates the best in each column.

	Qwen2.5-VL-7B (Bai et al., 2025)			Qwen3.5-9B (Qwen Team, 2026)			InternVL3.5-8B (Wang et al., 2025)		
Method	Del↓	Ins↑	Ins−Del↑	Del↓	Ins↑	Ins−Del↑	Del↓	Ins↑	Ins−Del↑
LLaVA-CAM (Zhang et al., 2025)	$45.20_{\pm 2.35}$	$45.68_{\pm 2.34}$	$0.48_{\pm 1.10}$	$49.47_{\pm 2.30}$	$44.17_{\pm 2.29}$	$-5.30_{\pm 0.94}$	$43.51_{\pm 2.39}$	$49.56_{\pm 2.40}$	$6.05_{\pm 1.23}$
iGOS++ (Xing et al., 2025)	$38.77_{\pm 2.36}$	$47.39_{\pm 2.37}$	$8.63_{\pm 1.08}$	$42.27_{\pm 2.32}$	$48.86_{\pm 2.30}$	$6.59_{\pm 0.97}$	$36.54_{\pm 5.53}$	$39.23_{\pm 5.74}$	$2.69_{\pm 2.29}$
TAM (Li et al., 2025b)	$38.01_{\pm 2.40}$	$51.07_{\pm 2.34}$	$13.06_{\pm 1.36}$	$40.59_{\pm 2.35}$	$50.62_{\pm 2.24}$	$10.03_{\pm 1.17}$	$38.14_{\pm 3.39}$	$55.09_{\pm 3.19}$	$16.95_{\pm 2.12}$
EAGLE (Chen et al., 2026)	$23.51_{\pm 2.06}$	$66.88_{\pm 2.02}$	$43.37_{\pm 1.80}$	$30.31_{\pm 2.49}$	$70.50_{\pm 1.90}$	$40.18_{\pm 1.87}$	$21.53_{\pm 5.31}$	$75.10_{\pm 4.30}$	$53.57_{\pm 4.86}$
VIRAH (Ours)	$\mathbf{4.76}_{\pm 2.01}$	$\mathbf{94.21}_{\pm 2.56}$	$\mathbf{89.45}_{\pm 1.96}$	$\mathbf{17.71}_{\pm 2.43}$	$\mathbf{92.65}_{\pm 1.19}$	$\mathbf{74.94}_{\pm 2.82}$	$\mathbf{20.34}_{\pm 4.49}$	$\mathbf{96.08}_{\pm 4.71}$	$\mathbf{75.74}_{\pm 3.24}$

Table 2: Localization results on MS-COCO (Lin et al., 2015) for Qwen2.5-VL-7B (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026), and InternVL3.5-8B (Wang et al., 2025). Bold indicates the best in each column.

	Qwen2.5-VL-7B (Bai et al., 2025)			Qwen3.5-9B (Qwen Team, 2026)			InternVL3.5-8B (Wang et al., 2025)		
Method	AUPR_1 ↑	AUPR_0 ↑	PG↑	AUPR_1 ↑	AUPR_0 ↑	PG↑	AUPR_1 ↑	AUPR_0 ↑	PG↑
LLaVA-CAM (Zhang et al., 2025)	$15.90_{\pm 1.13}$	$88.49_{\pm 0.88}$	$14.40_{\pm 2.22}$	$12.34_{\pm 0.97}$	$86.79_{\pm 0.91}$	$8.00_{\pm 1.72}$	$18.84_{\pm 1.15}$	$89.76_{\pm 0.90}$	$20.80_{\pm 2.57}$
iGOS++ (Xing et al., 2025)	$16.55_{\pm 1.03}$	$89.97_{\pm 0.80}$	$13.20_{\pm 2.15}$	$16.38_{\pm 1.05}$	$89.84_{\pm 0.80}$	$8.00_{\pm 1.72}$	$16.05_{\pm 3.04}$	$88.57_{\pm 2.37}$	$7.89_{\pm 4.43}$
TAM (Li et al., 2025b)	$31.80_{\pm 1.68}$	$93.55_{\pm 0.73}$	$33.20_{\pm 2.98}$	$28.78_{\pm 1.62}$	$94.05_{\pm 0.67}$	$30.00_{\pm 2.90}$	$34.41_{\pm 1.78}$	$94.51_{\pm 0.67}$	$38.80_{\pm 3.09}$
EAGLE (Chen et al., 2026)	$23.56_{\pm 1.20}$	$87.25_{\pm 1.00}$	$41.20_{\pm 3.12}$	$25.03_{\pm 1.37}$	$87.95_{\pm 1.10}$	$46.48_{\pm 3.43}$	$21.75_{\pm 2.73}$	$86.17_{\pm 2.83}$	$47.37_{\pm 8.21}$
VIRAH (Ours)	$\mathbf{43.74}_{\pm 1.12}$	$\mathbf{97.56}_{\pm 0.94}$	$\mathbf{57.43}_{\pm 2.76}$	$\mathbf{31.95}_{\pm 1.26}$	$\mathbf{94.32}_{\pm 0.89}$	$\mathbf{50.66}_{\pm 3.26}$	$\mathbf{36.81}_{\pm 2.18}$	$\mathbf{97.15}_{\pm 3.04}$	$\mathbf{60.00}_{\pm 5.11}$

We compare VIRAH against various LVLM attribution methods: LLaVA-CAM (Zhang et al., 2025), iGOS++ (Xing et al., 2025), Token Activation Map (TAM) (Li et al., 2025b), and EAGLE (Chen et al., 2026). These baselines represent activation-based, gradient-based, and perturbation-based attribution strategies.

In addition to object-centered attribution, we evaluate VIRAH on relation-conditioned predictions using the Visual Genome Relations benchmark (Krishna et al., 2017). This benchmark requires distinguishing the correct subject–relation–object ordering in an image. This evaluation examines whether VIRAH can identify the visual evidence supporting relational predictions. The corresponding results are presented in Appendix B.2.

Quantitative results. Tables 1 and 2 summarize the attribution results for Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026), and InternVL3.5-8B (Wang et al., 2025). Across model backbones and datasets, VIRAH demonstrates stronger faithfulness under both the Insertion and Deletion evaluations. In particular, regions identified by VIRAH recover the target-token probability more rapidly during Insertion, whereas removing these regions produces a larger reduction in prediction confidence during Deletion.

For example, Qwen2.5-VL-7B-Instruct, VIRAH improves over the strongest competing attribution method by reducing the Deletion AUC by 18.75% and increasing the Insertion AUC by 27.33%. Consequently, it improves the Insertion−Deletion score by 46.08%. These results indicate that the attribution maps produced by VIRAH more faithfully identify the visual evidence that influences the model's prediction.

VIRAH also achieves strong performance across the localization metrics, including the Pointing Game, AUPR_1, and AUPR_0. These results indicate that the regions identified by VIRAH align closely with the corresponding visual concepts while assigning relatively low importance to irrelevant background regions. The consistency of these results across different LVLM architectures further suggests the robustness and generalizability of VIRAH across diverse models. Additional quantitative results on further models and benchmarks are provided in Appendix B.1.

Qualitative results. Figure 2 (a) presents qualitative comparisons between VIRAH and existing LVLM attribution methods on the Qwen2.5-VL-7B-Instruct (Bai et al., 2025) model, including LLaVA-CAM, iGOS++, TAM, and EAGLE methods. We visualize attribution maps for different target concepts to examine how effectively each method identifies the visual regions associated with the queried concepts.

The examples show that existing approaches often assign high importance to broad image regions or activate areas unrelated to the target concept. In contrast, VIRAH produces more spatially concentrated attribution maps that better align with the corresponding visual evidence. For example, when attributing the presence of specific objects, VIRAH consistently highlights the object regions

Table 3: Hallucination results on MS-COCO (Lin et al., 2015) evaluated using CHAIR (Rohrbach et al., 2018) and Recall across Qwen2.5-VL-7B (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025), and LLaVA1.5-7B (Liu et al., 2024). Bold indicates the best result in each column.

	Qwen2.5-VL-7B (Bai et al., 2025)			Qwen3.5-9B (Qwen Team, 2026)			InternVL3.5-8B (Wang et al., 2025)			LLaVA1.5-7B (Liu et al., 2024)		
Method	$CHAIR_i \downarrow$	$CHAIR_s \downarrow$	Recall↑	$CHAIR_i \downarrow$	$CHAIR_s \downarrow$	Recall↑	$CHAIR_i \downarrow$	$CHAIR_s \downarrow$	Recall↑	$CHAIR_i \downarrow$	$CHAIR_s \downarrow$	Recall↑
SADT (Wang et al., 2026a)	5.10	19.30	55.20	**6.70**	26.00	58.13	8.68	**27.33**	49.33	11.20	42.00	78.50
Saliency-Guided (Zhang et al., 2026a)	8.92	24.00	52.20	10.83	33.33	49.43	12.12	42.00	65.20	16.51	56.00	74.00
OPERA (Huang et al., 2024)	6.75	28.67	70.36	8.50	**22.00**	45.89	7.74	33.33	69.98	13.18	51.33	79.54
VCD (Leng et al., 2024)	9.13	34.00	69.79	10.98	40.67	66.73	9.95	40.67	71.13	13.40	52.00	78.78
SPARC (Jung et al., 2025)	7.00	28.67	69.98	10.04	42.00	**67.69**	9.73	41.33	69.60	10.38	47.33	77.63
LTS-FS (Nullu) (Dang et al., 2026)	6.55	28.67	63.67	9.05	32.00	62.33	10.33	34.00	58.89	12.77	51.33	77.63
LTS-FS (VTI) (Dang et al., 2026)	6.02	22.67	64.82	9.31	36.00	62.14	7.99	37.33	71.70	10.51	40.67	77.73
PTI (Zhang et al., 2026b)	6.20	24.67	69.41	11.50	41.33	58.51	9.90	34.67	59.46	12.50	46.67	79.16
CausalLens (Ji et al., 2026)	8.79	34.00	69.79	10.75	44.00	66.92	8.42	35.33	70.94	12.33	46.00	78.59
Grounding-Ids (Hasani et al., 2026)	5.26	20.00	54.88	8.27	28.67	52.39	7.57	29.33	59.85	11.44	42.00	72.85
VIRAH (Ours)	**3.20**	**17.10**	**74.40**	8.11	30.41	67.34	**6.33**	31.26	**73.67**	**10.11**	**39.52**	**80.47**

rather than surrounding background areas. These results demonstrate that VIRAH produces more semantically aligned attribution maps compared with attention-, gradient-, and perturbation-based approaches. Additional qualitative results are provided in Appendix B.2

4.2 Hallucination Mitigation

We evaluate the effectiveness of VIRAH in mitigating object hallucinations across different LVLMs and benchmarks. We compare VIRAH with recent hallucination-mitigation approaches and analyze its ability to reduce unsupported object generation while preserving correct visual content.

Evaluation metrics. We use the CHAIR metrics (Rohrbach et al., 2018) to quantify object hallucinations. $CHAIR_i$ measures the proportion of hallucinated object instances, while $CHAIR_s$ measures the proportion of responses containing at least one hallucinated object. Lower values indicate fewer hallucinations. We additionally report Recall to measure how well the model preserves valid object information. Higher Recall indicates better performance.

Experimental setup. We evaluate hallucination mitigation on Qwen2.5-VL-7B (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025), and LLaVA1.5-7B (Liu et al., 2024) models. We compare VIRAH against recent hallucination-mitigation methods: SADT (Wang et al., 2026a), Saliency-Guided (Zhang et al., 2026a), OPERA (Huang et al., 2024), Visual Contrastive Decoding (VCD) (Leng et al., 2024), SPARC (Jung et al., 2025), LTS-FS (both Nullu and VTI) (Dang et al., 2026), Prefill-Time Intervention (PTI) (Zhang et al., 2026b), CausalLens (Ji et al., 2026), and Grounding-Ids (Hasani et al., 2026). We conduct experiments on MS-COCO (Lin et al., 2015) and POPE (Li et al., 2023b) using the same evaluation protocol for all methods.

Quantitative results. Table 3 summarizes the main hallucination-mitigation results on MS-COCO (Lin et al., 2015). VIRAH reduces both $CHAIR_i$ and $CHAIR_s$, showing that it decreases hallucinated object mentions at both the instance and response levels. At the same time, it improves Recall relative to existing methods, indicating that hallucination reduction is not achieved by simply generating fewer object mentions or overly conservative responses. Additional results on further benchmarks are reported in Appendix B.1.

Qualitative results. Figure 2 (b) presents qualitative comparisons of hallucination mitigation methods on image caption generation using the Qwen2.5-VL-7B(Bai et al., 2025) model. We compare VIRAH with the five best-performing hallucination mitigation methods reported in Table 3 under the same generation configuration.

Existing approaches produce descriptions containing details that are not directly supported by the image, including inaccurate object attributes or unverified scene descriptions. In contrast, VIRAH generates captions that remain more consistent with the visible content. In the presented example, VIRAH describes the main objects and scene elements while avoiding additional details that are not clearly supported by the image. These qualitative results demonstrate improved visual grounding in generated captions compared with the existing approaches.



Figure 2: Qualitative comparison on Qwen2.5-VL-7B-Instruct (Bai et al., 2025). (a) Visual attribution maps from LLaVA-CAM (Zhang et al., 2025), iGOS++ (Xing et al., 2025), TAM (Li et al., 2025b), EAGLE (Chen et al., 2026), and VIRAH. VIRAH yields more localized target-relevant regions. (b) Hallucination mitigation results, where VIRAH reduces unsupported descriptions while preserving relevant details. All methods use the same generation setting with a maximum length of 256 tokens.

5 Discussion and Limitations

VIRAH treats the LVLM as a probe of its own visual representations. By retaining a local neighborhood of visual tokens at a chosen encoder depth and obtaining the decoder's output distribution over the vocabulary, VIRAH measures what the selected visual evidence supports. The same operation serves two purposes: the probability of an affirmative response produces attribution maps, while the entropy of an object-identification response identifies ambiguous regions whose removal reduces hallucination. Both signals are obtained from the model's forward computation alone, without external region generators, detectors, reference images, or learned steering directions.

Empirically, VIRAH improves attribution faithfulness and localization across Qwen, InternVL, and LLaVA, with stronger Insertion−Deletion and localization performance on AUPR_1, AUPR_0, and Pointing Game. It also reduces object hallucinations across benchmarks by lowering CHAIR while improving Recall. Overall, localized visual-token intervention provides an effective mechanism for both visual evidence analysis and more reliable LVLM generation.

Despite these advantages, VIRAH has some limitations. First, VIRAH operates at the visual-token level rather than the pixel level, which limits the granularity of the attribution maps. While token-level analysis aligns with the internal representation of LVLMs, it may miss fine-grained spatial details that pixel-level or patch-refinement approaches can capture. Second, uncertainty-based token removal may occasionally suppress visual evidence in ambiguous scenes where the model is uncertain but the information remains relevant.

AI USE STATEMENT

Generative AI tools were used during the preparation of this work to assist with language editing and polishing, as well as code development and debugging. AI-assisted code was reviewed, tested, and modified by the authors. All experiments and evaluations reported in this paper were designed and conducted by the authors, and all AI-assisted manuscript content was carefully reviewed and verified before inclusion in the paper.

Generative AI tools were not used to generate synthetic datasets, develop the theoretical framework or conceptual foundations of the work, formulate mathematical claims, design or modify the research methodology or experiments, interpret experimental results, perform qualitative or thematic analysis, clean or reformat datasets, or generate scientific findings.

REFERENCES

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023a. URL https://arxiv.org/abs/2301.12597.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

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*, pages 26296–26306, 2024.

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*, pages 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

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*, pages 292–305. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation, 2023. URL https://arxiv.org/abs/2311.07397.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Information flow in LVLMs across reasoning tasks. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2289–2299, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.115. URL https://aclanthology.org/2025.naacl-long.115/.

Faridoun Mehri, Mahdieh Soleymani Baghshah, and Mohammad Taher Pilehvar. LibraGrad: Balancing gradient flow for universally better vision transformer attributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 67–78, 2025.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference*, 2018. URL https://arxiv.org/abs/1806.07421.

Ruoyu Chen, Xiaoqing Guo, Kangwei Liu, Siyuan Liang, Shiming Liu, Qunli Zhang, Laiyuan Wang, Hua Zhang, and Xiaochun Cao. Where mllms attend and what they rely on: Explaining autoregressive token generation, 2026. URL https://arxiv.org/abs/2509.22496.

Yao Zhu, Yunjian Zhang, Zizhe Wang, Xiu Yan, Peng Sun, and Xiangyang Ji. Patchwise cooperative game-based interpretability method for large vision-language models. *Transactions of the Association for Computational Linguistics*, 13:744–759, 2025. doi: 10.1162/tacl_a_00756. URL https://aclanthology.org/2025.tacl-1.34/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Amirmohammad Izadi, Mohammadali Banayeeanzade, Alireza Mirrokni, Hosein Hasani, Mobin Bagherian, Faridoun Mehri, and Mahdieh Soleymani Baghshah. Causal attribution via activation patching, 2026. URL https://arxiv.org/abs/2603.13652.

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.

Badri N. Patro, Mayank Lunayach, Shivansh Patel, and Vinay P. Namboodiri. U-CAM: Visual explanation using uncertainty based class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7444–7453, 2019.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Understanding and 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*, pages 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

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*, pages 13418–13427, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7824–7846. PMLR, 2024.

Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. Visual attention never fades: Selective progressive attention ReCalibration for detailed image captioning in multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 28527–28551. PMLR, 2025. URL https://proceedings.mlr.press/v267/jung25c.html.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 35799–35819. PMLR, 2025a.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate lvlm object hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25315–25325, June 2026a.

Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal llms, 2025b. URL https://arxiv.org/abs/2506.23270.

Xiaoying Xing, Chia-Wen Kuo, Li Fuxin, Yulei Niu, Fan Chen, Ming Li, Ying Wu, Longyin Wen, and Sijie Zhu. Where do large vision-language models look at when answering questions?, 2025. URL https://arxiv.org/abs/2503.13891.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv.org/abs/2211.00593.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops, 2026a. URL https://arxiv.org/abs/2601.20279.

Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. Uncovering grounding ids: How external cues shape multimodal binding, 2026. URL https://arxiv.org/abs/2509.24072.

Nhat Hoang-Xuan, Minh Vu, My T. Thai, and Manish Bhattarai. Pas: Prelim attention score for detecting object hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Runhe Lai, Xinhua Lu, Yanqi Wu, Jinlun Ye, Weijiang Yu, and Ruixuan Wang. Instruction lens score: Your instruction contributes a powerful object hallucination detector for multimodal large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=kCJ9kaebgd.

Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Yutong Xie, Nguyen Cam-Tu, Minh Khoi Nguyen, Dang Huy Pham Nguyen, Anton van den Hengel, Johan W. Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Beyond the global scores: Fine-grained token grounding as a robust detector of lvlm hallucinations, 2026. URL https://arxiv.org/abs/2604.04863.

Gabriel Afriat, Ryan Lucas, Xiang Meng, Yufang Hou, Yada Zhu, and Rahul Mazumder. Seeing what's wrong: A trajectory-guided approach to caption error detection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=73FGjnKu1P.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Bjq4W7P2Us.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. Causallens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4199–4209, June 2026.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating hallucinations in large vision-language models via causal route gating, 2026. URL https://arxiv.org/abs/2605.24024.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.634. URL https://aclanthology.org/2025.acl-long.634/.

Tianbo Wang, Yuqing Ma, Kewei Liao, Zhange Zhang, Simin Li, Jinyang Guo, and Xianglong Liu. After: Mitigating the object hallucination of lvlm via adaptive factual-guided activation editing, 2026b. URL https://arxiv.org/abs/2601.01957.

Tiantian Dang, Chao Bi, Shufan Shen, Jinzhe Liu, Qingming Huang, and Shuhui Wang. Locate-then-sparsify: Attribution guided sparse strategy for visual hallucination mitigation, 2026. URL https://arxiv.org/abs/2603.16284.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. *arXiv preprint arXiv:2507.00898*, 2025.

Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. Prefill-time intervention for mitigating hallucination in large vision-language models, 2026b. URL https://arxiv.org/abs/2604.25642.

Jiaye Qian, Ge Zheng, Yuchen Zhu, and Sibei Yang. Intervene-all-paths: Unified mitigation of lvlm hallucinations across alignment formats, 2026. URL https://arxiv.org/abs/2511.17254.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

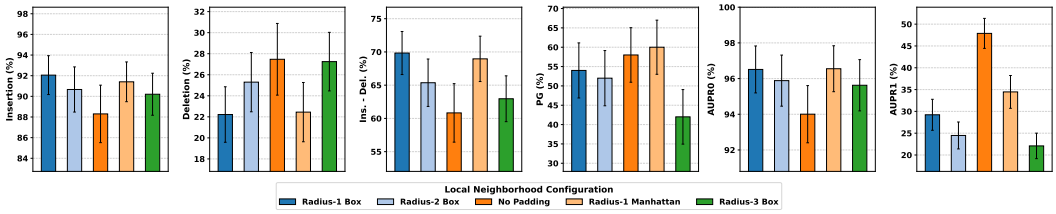

Figure 3: Ablation of the local neighborhood configuration for attribution.

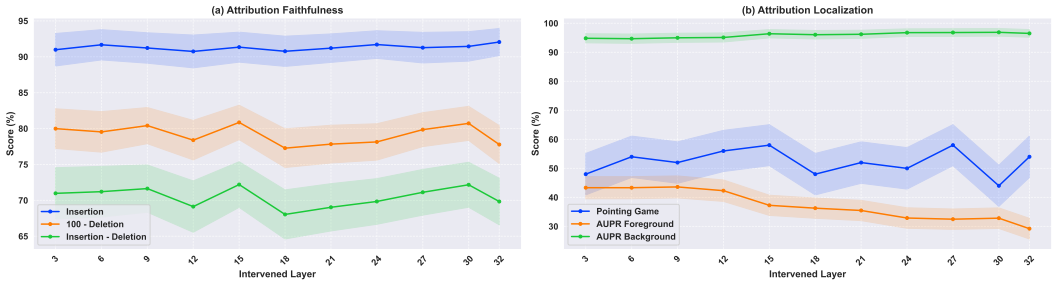

Figure 4: Effect of intervention depth on attribution faithfulness and localization.

A Ablation Studies

We ablate three design choices of VIRAH on MS-COCO: (i) the local neighborhood configuration, (ii) the intervened vision-encoder layer, and (iii) the number of uncertainty clusters used for hallucination mitigation.

Local neighborhood configuration. We vary the spatial neighborhood retained around each center patch. *No Padding* retains only the center patch, *Radius-r Box* retains a $(2r+1) \times (2r+1)$ neighborhood, and *Radius-1 Manhattan* retains the center patch together with its horizontal and vertical neighbors. Figure 3 shows that both Radius-1 Box and Radius-1 Manhattan achieve strong faithfulness, obtaining higher Insertion and Insertion$-$Deletion scores and lower Deletion scores compared with No Padding. Radius-1 Box provides the strongest overall faithfulness, while No Padding achieves better foreground localization on Pointing Game and AUPR_1. Larger neighborhoods do not provide consistent additional improvements. We therefore use Radius-1 Box by default as it offers a favorable balance between faithfulness and localization.

Choice of intervened layer. We vary the vision-encoder depth at which tokens outside the selected region are removed. Earlier intervention limits full-image interaction, while later intervention preserves more globally contextualized representations. Figure 4 shows relatively stable faithfulness across layers, but greater variation in localization, particularly for Pointing Game and AUPR_1, while AUPR_0 remains consistently high. For hallucination mitigation, Figure 5(b) shows that intervention depth affects the trade-off between hallucination reduction and Recall.

Number of clusters. We vary the number of K-means clusters used to group regional uncertainty scores, removing the cluster with the highest mean entropy as potentially unreliable. Figure 5(a) shows that three clusters provide the best trade-off, reducing CHAIR_s and CHAIR_i while preserving Recall. We therefore use three clusters in the main experiments.

B Additional Experiments

In this section, we provide additional quantitative and qualitative experiments to further evaluate the generalizability and robustness of VIRAH across different model backbones, datasets, and visual concepts.

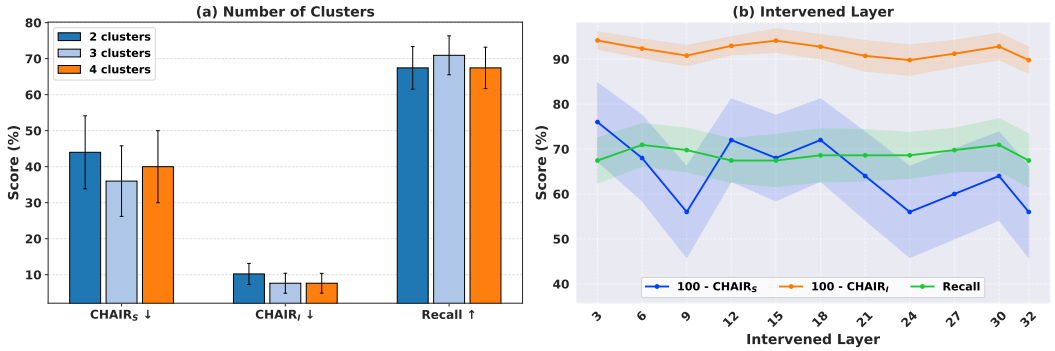

Figure 5: Hallucination-mitigation ablations over (a) uncertainty clusters and (b) intervention depth.

Table 4: Faithfulness results on MMVP (Tong et al., 2024) for Qwen2.5-VL-7B (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025), and LLaVA1.5-7B (Liu et al., 2024). Bold indicates the best in each column.

	Qwen2.5-VL-7B (Bai et al., 2025)			Qwen3.5-9B (Qwen Team, 2026)			InternVL3.5-8B (Wang et al., 2025)			LLaVA1.5-7B (Liu et al., 2024)		
Method	Del↓	Ins↑	Ins−Del↑	Del↓	Ins↑	Ins−Del↑	Del↓	Ins↑	Ins−Del↑	Del↓	Ins↑	Ins−Del↑
LLaVA-CAM (Zhang et al., 2025)	$73.37_{\pm 0.79}$	$73.13_{\pm 0.77}$	$-0.24_{\pm 0.34}$	$80.26_{\pm 0.45}$	$79.20_{\pm 0.46}$	$-1.06_{\pm 0.33}$	$79.39_{\pm 1.23}$	$80.31_{\pm 1.06}$	$0.92_{\pm 0.59}$	$77.01_{\pm 1.46}$	$77.19_{\pm 1.40}$	$0.18_{\pm 0.22}$
iGOS++ (Xing et al., 2025)	$71.02_{\pm 0.75}$	$72.60_{\pm 0.77}$	$1.58_{\pm 0.28}$	$77.69_{\pm 0.49}$	$78.95_{\pm 0.51}$	$1.26_{\pm 0.19}$	$77.56_{\pm 1.25}$	$78.88_{\pm 1.19}$	$1.32_{\pm 0.37}$	$76.98_{\pm 1.48}$	$77.03_{\pm 1.41}$	$0.05_{\pm 0.16}$
TAM (Li et al., 2025b)	$72.83_{\pm 0.70}$	$73.94_{\pm 0.83}$	$1.11_{\pm 0.39}$	$79.76_{\pm 0.42}$	$79.79_{\pm 0.48}$	$0.03_{\pm 0.32}$	$79.40_{\pm 1.11}$	$81.65_{\pm 1.03}$	$2.25_{\pm 0.59}$	$76.34_{\pm 1.46}$	$77.61_{\pm 1.39}$	$1.27_{\pm 0.31}$
EAGLE (Chen et al., 2026)	$69.05_{\pm 0.65}$	$75.82_{\pm 0.88}$	$6.77_{\pm 0.58}$	$75.74_{\pm 0.48}$	**82.86**$_{\pm 0.41}$	$7.12_{\pm 0.29}$	$72.47_{\pm 1.29}$	$86.23_{\pm 0.92}$	$13.76_{\pm 0.88}$	$74.39_{\pm 1.46}$	$79.21_{\pm 1.39}$	$4.82_{\pm 0.39}$
VIRAH (Ours)	**38.21**$_{\pm 0.81}$	**82.15**$_{\pm 0.79}$	**43.94**$_{\pm 0.46}$	**37.70**$_{\pm 0.21}$	$78.03_{\pm 0.23}$	**40.33**$_{\pm 0.17}$	**39.13**$_{\pm 0.96}$	**87.21**$_{\pm 2.04}$	**48.08**$_{\pm 0.54}$	**56.22**$_{\pm 1.59}$	**81.51**$_{\pm 1.31}$	**25.29**$_{\pm 0.28}$

B.1 Quantitative Results

Attribution To further assess the generalizability of VIRAH, we conduct additional attribution experiments on the MMVP (Tong et al., 2024) benchmark using the four LVLMs described in our experimental setup. Table 6 presents the results for LLaVA1.5-7B (Liu et al., 2024) on the MS-COCO (Lin et al., 2015) dataset, while Table 4 reports results on the MMVP benchmark under the same experimental settings and evaluation protocol as the main experiments. Across both datasets and model backbones, VIRAH consistently improves faithfulness and localization metrics, demonstrating its effectiveness beyond the specific models and datasets evaluated in the main experiments.

Moreover, we evaluate VIRAH on spatial relationships using the Visual Genome Relations benchmark (Krishna et al., 2017). Unlike object-centered attribution, where the target evidence typically corresponds to a single visual region, explaining a spatial relationship requires identifying the visual evidence associated with multiple objects and their relative configuration. For example, predicting a relation such as *"person riding horse"* requires capturing not only the person and the horse, but also the spatial interaction between them. This setting therefore provides a more challenging evaluation of whether VIRAH can explain relational predictions rather than only individual object concepts. The results in Table 5 show that VIRAH can effectively localize the visual evidence supporting spatial relationships, demonstrating its applicability beyond object-level attribution.

Hallucination Mitigation We further evaluate the hallucination mitigation capability of VIRAH on the POPE benchmark (Li et al., 2023b). These experiments complement the main MS-COCO evaluation by examining whether the proposed mitigation strategy generalizes across different model architectures and extends to a discriminative object-hallucination setting. The results are provided in Table 7. The results show that VIRAH consistently reduces hallucination while maintaining strong object-recognition performance across the evaluated models and benchmarks.

B.2 Qualitative Results

Attribution Examples We present additional qualitative comparisons to examine the behavior of our attribution method across different model backbones and object categories. As shown in Figures 6–18, our method produces more precise and spatially coherent attribution maps than the baseline approaches. In particular, the attributed regions align more closely with the relevant objects, exhibit sharper boundaries, and contain fewer irrelevant activations. These examples further

Table 5: Faithfulness results on Visual Genome (Krishna et al., 2017) for Qwen2.5-VL-7B (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025), and LLaVA1.5-7B (Liu et al., 2024). Bold indicates the best in each column.

	Qwen2.5-VL-7B (Bai et al., 2025)			Qwen3.5-9B (Qwen Team, 2026)			InternVL3.5-8B (Wang et al., 2025)			LLaVA1.5-7B (Liu et al., 2024)		
Method	Del↓	Ins↑	Ins−Del↑	Del↓	Ins↑	Ins−Del↑	Del↓	Ins↑	Ins−Del↑	Del↓	Ins↑	Ins−Del↑
LLaVA-CAM (Zhang et al., 2025)	$24.55_{\pm 1.72}$	$25.26_{\pm 1.73}$	$0.71_{\pm 0.51}$	$26.79_{\pm 1.95}$	$26.97_{\pm 1.92}$	$0.18_{\pm 0.26}$	$27.48_{\pm 2.08}$	$27.70_{\pm 2.09}$	$0.22_{\pm 0.22}$	$37.15_{\pm 1.87}$	$36.38_{\pm 1.85}$	$-0.77_{\pm 0.33}$
iGOS++ (Xing et al., 2025)	$22.89_{\pm 3.10}$	$24.79_{\pm 3.47}$	$1.90_{\pm 0.77}$	$23.73_{\pm 3.50}$	$23.77_{\pm 3.56}$	$0.04_{\pm 0.27}$	$24.77_{\pm 3.82}$	$24.65_{\pm 3.84}$	$-0.12_{\pm 0.31}$	$32.80_{\pm 3.53}$	$32.99_{\pm 3.55}$	$0.19_{\pm 0.19}$
TAM (Li et al., 2025b)	$24.81_{\pm 1.75}$	$25.39_{\pm 1.75}$	$0.58_{\pm 0.47}$	$25.97_{\pm 1.94}$	$27.43_{\pm 1.91}$	$1.46_{\pm 0.35}$	$27.28_{\pm 2.09}$	$28.11_{\pm 2.09}$	$0.83_{\pm 0.27}$	$36.94_{\pm 1.88}$	$36.77_{\pm 1.85}$	$-0.17_{\pm 0.30}$
EAGLE (Chen et al., 2026)	$17.41_{\pm 2.69}$	$39.54_{\pm 4.01}$	$22.13_{\pm 2.74}$	$\mathbf{19.96}_{\pm 3.27}$	$34.09_{\pm 4.00}$	$14.13_{\pm 1.93}$	$\mathbf{21.37}_{\pm 3.52}$	$35.89_{\pm 4.19}$	$14.52_{\pm 1.76}$	$\mathbf{29.21}_{\pm 3.22}$	$39.67_{\pm 4.19}$	$10.46_{\pm 1.33}$
VIRAH (Ours)	$\mathbf{13.65}_{\pm 1.59}$	$\mathbf{85.79}_{\pm 1.32}$	$\mathbf{72.13}_{\pm 1.85}$	$37.74_{\pm 1.19}$	$\mathbf{73.12}_{\pm 0.1}$	$\mathbf{35.38}_{\pm 1.59}$	$34.18_{\pm 0.2}$	$\mathbf{95.21}_{\pm 1.33}$	$\mathbf{61.03}_{\pm 2.42}$	$61.68_{\pm 0.08}$	$\mathbf{84.83}_{\pm 0.06}$	$\mathbf{23.15}_{\pm 0.08}$

Table 6: Faithfulness and localization results on MS-COCO (Lin et al., 2015) for LLaVA1.5-7B (Liu et al., 2024). Bold indicates the best in each column.

	Faithfulness			Localization		
Method	Del↓	Ins↑	Ins−Del↑	AUPR_1 ↑	AUPR_0 ↑	PG↑
LLaVA-CAM (Zhang et al., 2025)	$61.88_{\pm 2.32}$	$62.78_{\pm 2.27}$	$0.90_{\pm 0.91}$	$14.91_{\pm 1.08}$	$88.57_{\pm 0.90}$	$11.60_{\pm 2.03}$
iGOS++ (Xing et al., 2025)	$61.47_{\pm 5.01}$	$62.81_{\pm 4.70}$	$1.34_{\pm 1.23}$	$12.48_{\pm 1.93}$	$88.74_{\pm 1.73}$	$5.00_{\pm 2.84}$
TAM (Li et al., 2025b)	$56.00_{\pm 2.49}$	$66.91_{\pm 2.18}$	$10.91_{\pm 1.09}$	$36.22_{\pm 1.64}$	$94.60_{\pm 0.69}$	$42.40_{\pm 3.13}$
EAGLE (Chen et al., 2026)	$46.99_{\pm 5.32}$	$76.66_{\pm 3.36}$	$29.67_{\pm 3.26}$	$25.50_{\pm 2.87}$	$88.29_{\pm 2.12}$	$40.00_{\pm 6.38}$
VIRAH (Ours)	$\mathbf{40.86}_{\pm 3.31}$	$\mathbf{87.47}_{\pm 4.15}$	$\mathbf{46.61}_{\pm 1.87}$	$\mathbf{39.27}_{\pm 2.13}$	$\mathbf{96.23}_{\pm 2.01}$	$\mathbf{47.50}_{\pm 4.19}$

demonstrate the robustness of our method across diverse visual concepts, including objects with different shapes, textures, and levels of visual complexity.



Figure 6: Qualitative attribution example for a `person`.

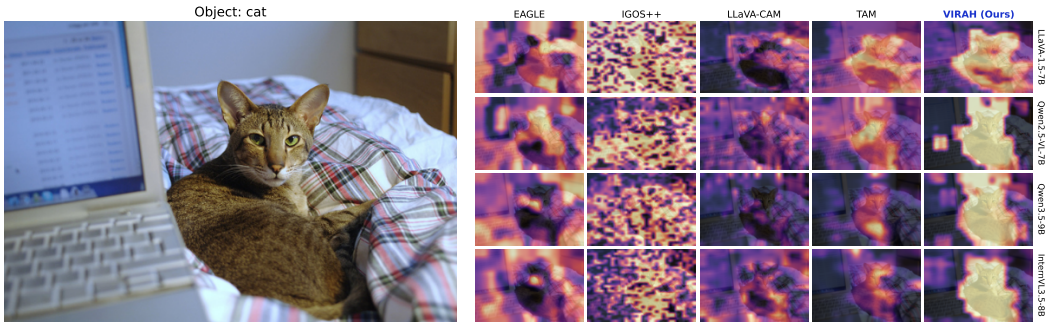

Figure 7: Qualitative attribution example for a `cat`.

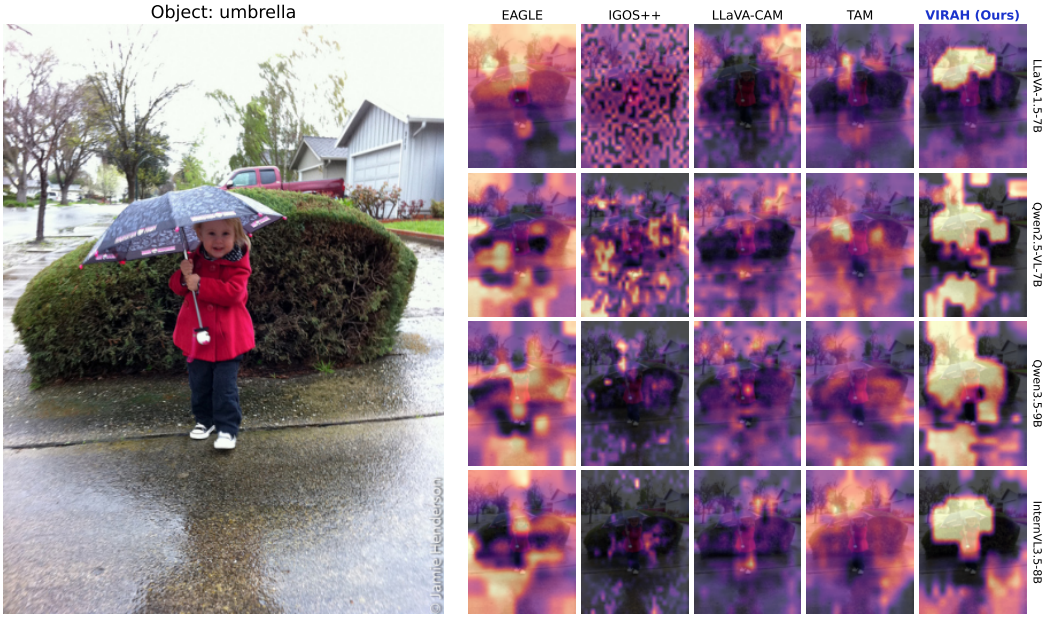

Figure 8: Qualitative attribution example for an `umbrella`.

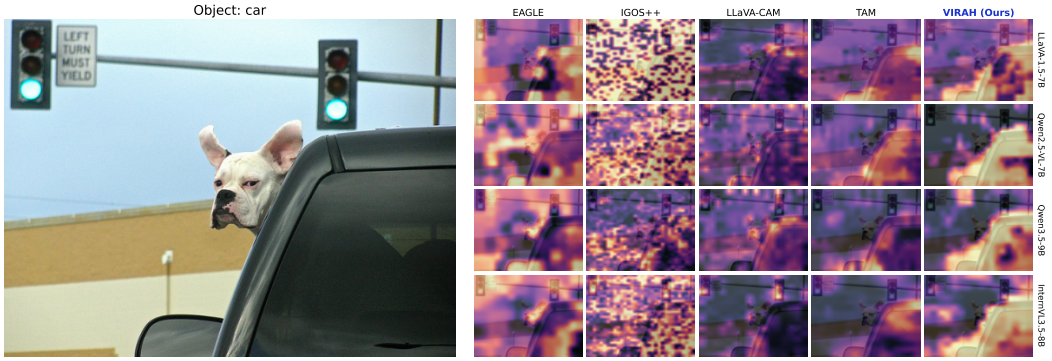

Figure 9: Qualitative attribution example for a `car`.

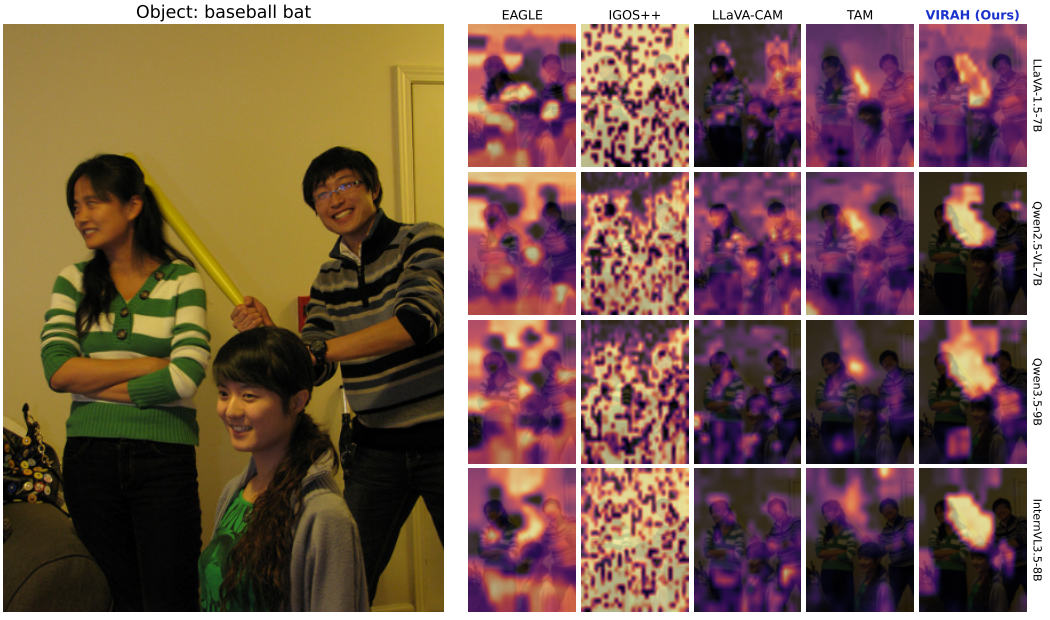

Figure 10: Qualitative attribution example for a `baseball bat`.

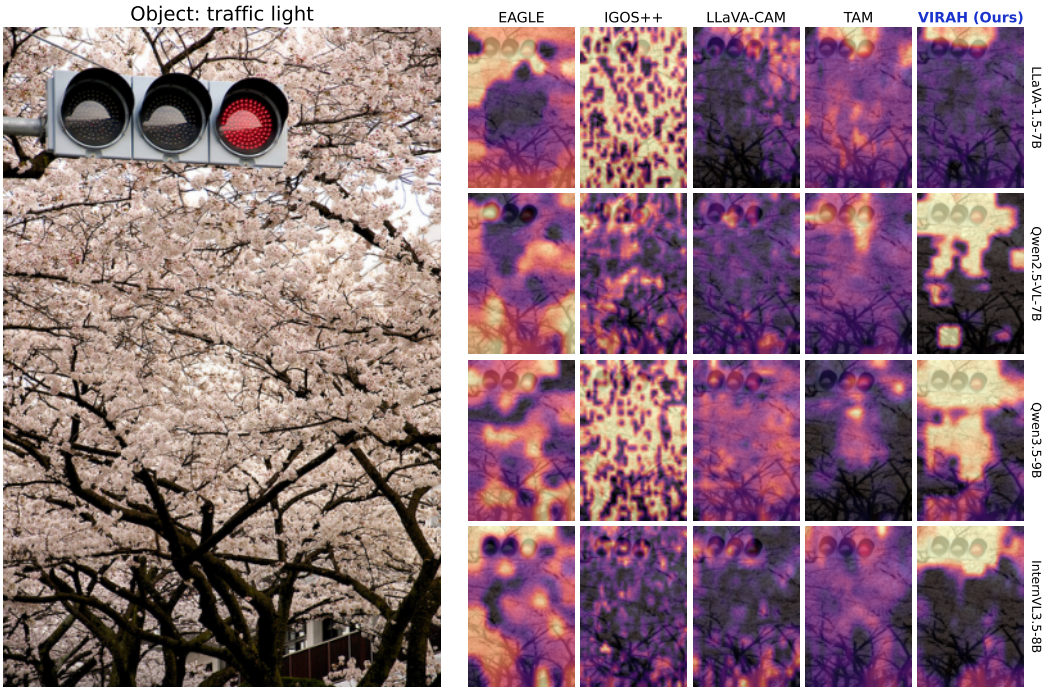

Figure 11: Qualitative attribution example for a `traffic light`.

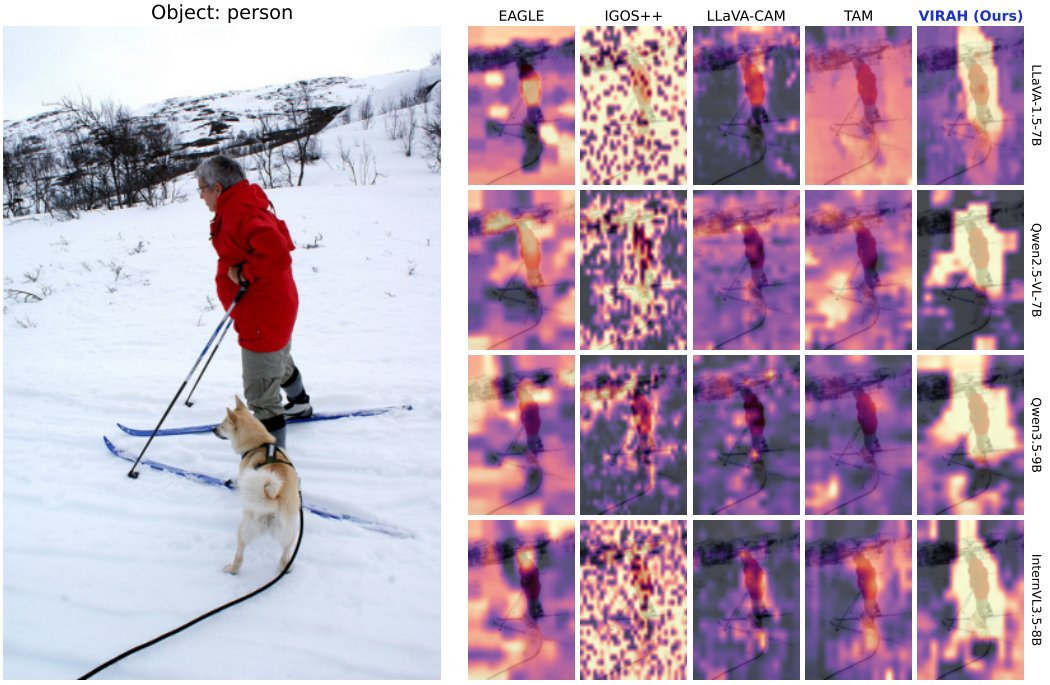

Figure 12: Qualitative attribution example for a `person`.

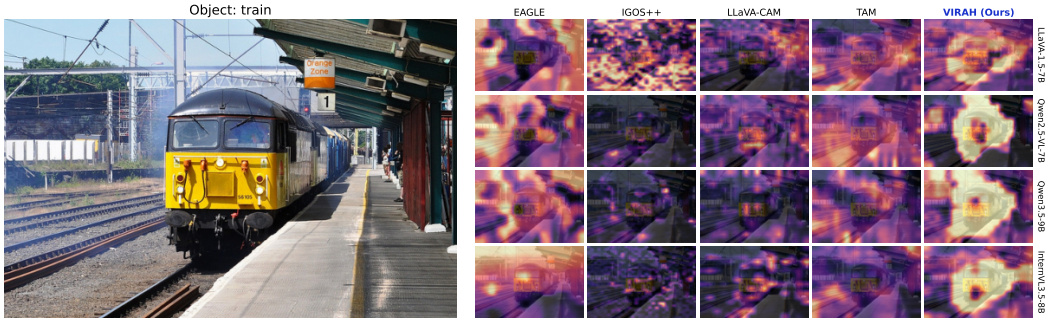

Figure 13: Qualitative attribution example for a `train`.

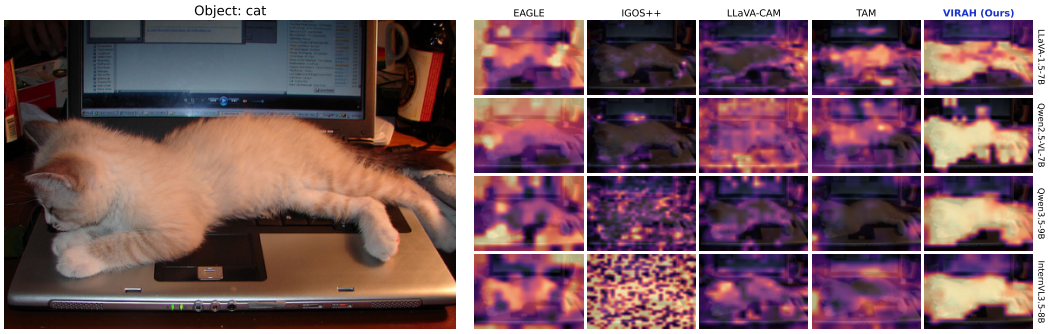

Figure 14: Qualitative attribution example for a `cat`.

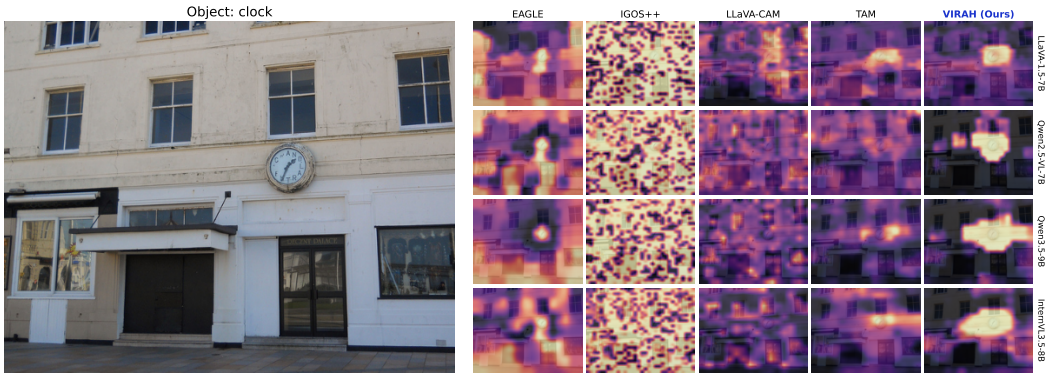

Figure 15: Qualitative attribution example for a `clock`.

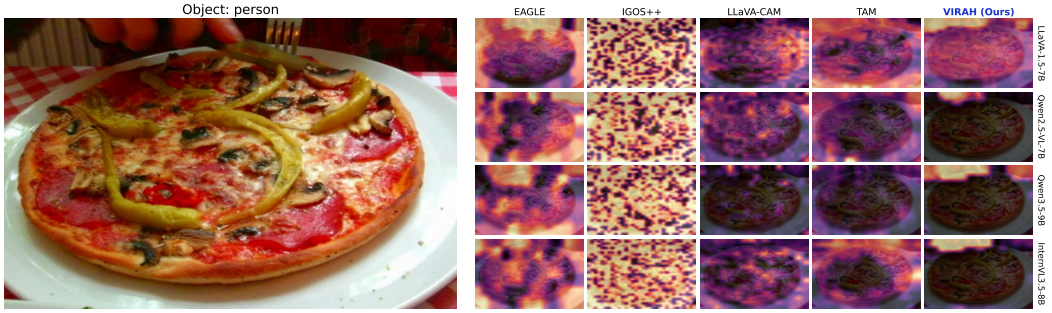

Figure 16: Qualitative attribution example for a `person`.

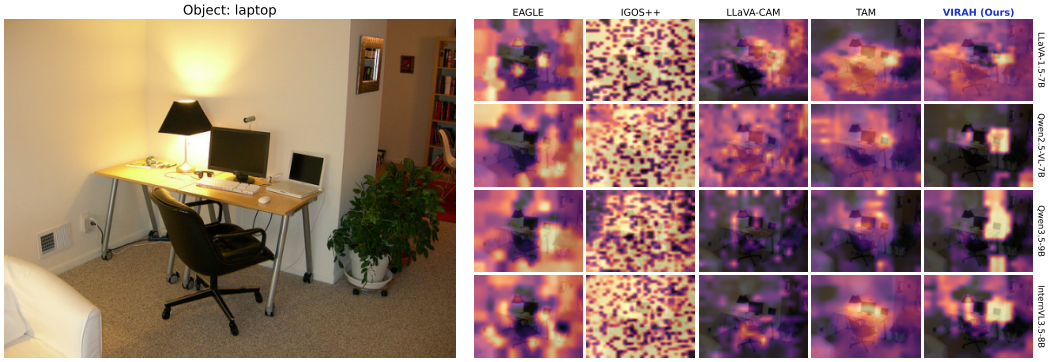

Figure 17: Qualitative attribution example for a `laptop`.

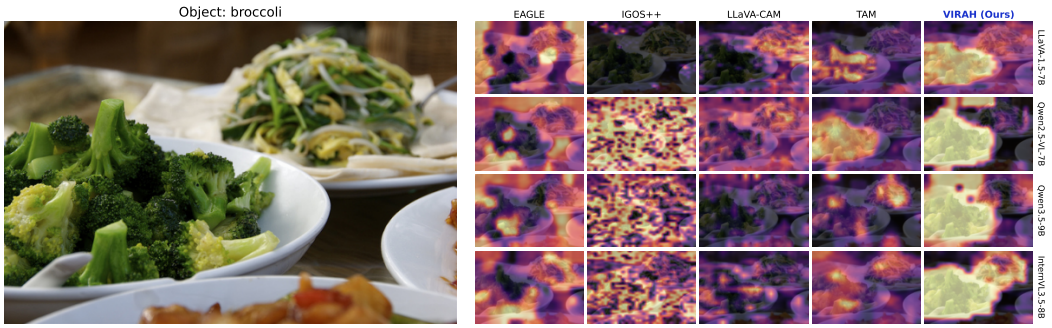

Figure 18: Qualitative attribution example for a `broccoli`.

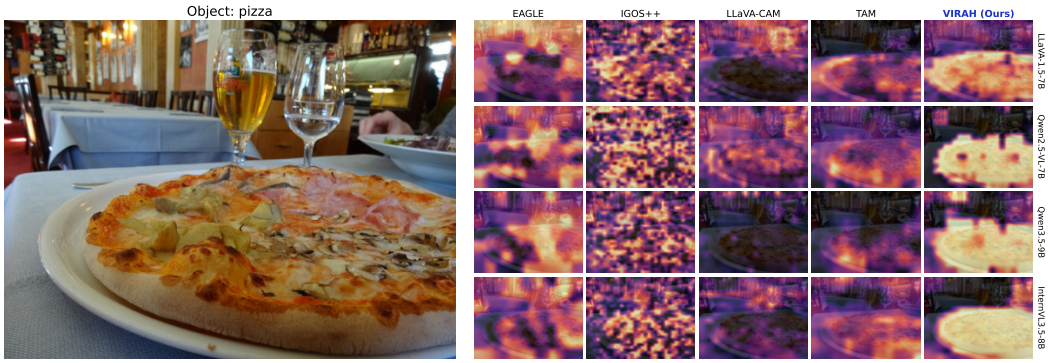

Figure 19: Qualitative attribution example for a `pizza`.

Table 7: Hallucination results on POPE (Li et al., 2023b) evaluated using Precision, Recall, and F1-score across Qwen2.5-VL-7B (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025), and LLaVA1.5-7B (Liu et al., 2024). Bold indicates the best numerical result in each column.

	Qwen2.5-VL-7B (Bai et al., 2025)			Qwen3.5-9B (Qwen Team, 2026)			InternVL3.5-8B (Wang et al., 2025)			LLaVA1.5-7B (Liu et al., 2024)		
Method	Precision↑	Recall↑	F1-score↑	Precision↑	Recall↑	F1-score↑	Precision↑	Recall↑	F1-score↑	Precision↑	Recall↑	F1-score↑
SADT (Wang et al., 2026a)	96.77	77.78	**86.24**	50.00	**100.00**	66.67	85.57	96.67	90.78	88.10	82.22	85.06
Saliency-Guided (Zhang et al., 2026a)	70.65	**81.11**	75.52	74.74	81.11	77.80	85.15	95.56	90.05	54.43	81.85	65.38
OPERA (Huang et al., 2024)	**100**	66.67	80	87.8	80	83.72	86.54	**100**	92.78	83.72	80	81.82
VCD (Leng et al., 2024)	**100**	73.33	84.62	88.89	88.89	**88.89**	86.54	**100**	92.78	78	**86.67**	82.11
SPARC (Jung et al., 2025)	**100**	73.33	84.62	88.64	86.67	87.64	86.54	**100**	92.78	76.6	80	78.26
LTS-FS (Nullu) (Dang et al., 2026)	**100**	66.67	80	**90**	80	84.71	90	**100**	94.74	78.26	80	79.12
LTS-FS (VTI) (Dang et al., 2026)	94.29	73.33	82.5	88.64	86.67	87.64	86.54	**100**	92.78	75	**86.67**	80.41
PTI (Zhang et al., 2026b)	94.29	73.33	82.5	88.64	86.67	87.64	86.54	**100**	92.78	80	80	80
CausalLens (Ji et al., 2026)	94.29	73.33	82.5	84.78	86.67	85.71	88.24	**100**	93.75	70.91	**86.67**	78
Grounding-Ids (Hasani et al., 2026)	96.00	71.11	81.70	51.04	**100.00**	67.58	76.30	97.78	85.71	**88.30**	**86.67**	**87.48**
VIRAH (Ours)	94.44	73.33	83.52	**90**	86.67	87.92	**90.90**	**100**	**95.00**	87.39	**86.67**	86.85

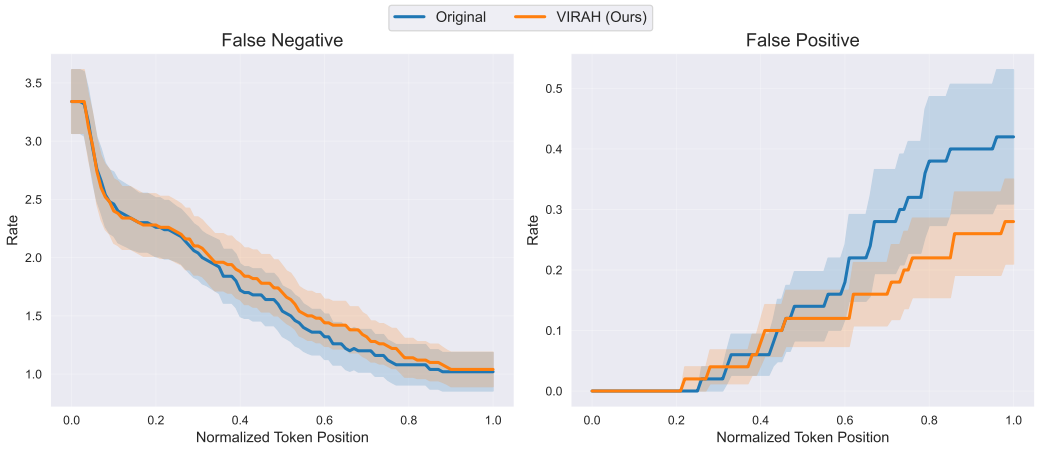

Figure 20: Evolution of object hallucination throughout caption generation. The plots show the cumulative number of false positive object mentions and remaining false negative object classes at normalized token positions. The x-axis represents generation progress, expressed as the fraction of the caption's total token count, where 0 and 1 correspond to the beginning and end of caption generation, respectively. Uncertainty-guided visual-token knockout reduces the accumulation of unsupported object mentions in later generation stages while maintaining comparable final ground-truth object coverage.

B.3 Hallucination Mitigation Analysis.

Aggregate metrics summarize caption quality but do not reveal when unsupported object mentions emerge during generation. We therefore examine false positives and false negatives at successive caption prefixes, comparing ordinary full-image generation (Original) with VIRAH. Figure 20 shows the mean cumulative number of unsupported COCO object mentions and the mean number of ground-truth COCO object classes not yet mentioned. Generation progress is expressed as a fraction of each caption's total token count.

The two conditions exhibit similar false-positive counts early in generation, but their trajectories separate more clearly as the captions develop. Ordinary generation continues to accumulate unsupported mentions, whereas knockout limits their accumulation, particularly in the latter part of the caption. At completion, the mean false-positive count decreases from approximately 0.42 to 0.28 mentions per caption. The false-negative curves provide complementary evidence: knockout leaves slightly more ground-truth classes unmentioned at several intermediate prefixes, but reaches comparable coverage by the end of generation. Thus, the observed reduction in unsupported mentions does not come at the expense of final ground-truth object coverage. Together, these trajectories suggest that the intervention primarily curbs unsupported elaboration as generation proceeds, while preserving the model's ability to describe the annotated objects.

C Implementation Details

Our implementation involves two components: attribution and hallucination mitigation. The corresponding details are described separately below.

C.1 Attribution Details

For attribution, we perform visual-token knockout in the final vision-encoder layer and use a spatial radius of one for the selected patch operator, corresponding to a 3×3 region around each selected patch. These choices follow the ablation results in Section A.

We formulate the target object, property, or relation as a yes/no question and append the instruction: "Answer with exactly one word: yes or no." The attribution score is computed as the probability assigned to the next-token answer "yes."

During autoregressive generation, textual token representations can incorporate information from spatially scattered image regions, including regions with little apparent relevance to the target concept. As these representations condition subsequent tokens, their visual contributions can accumulate and jointly influence later predictions. Attribution for a later token may therefore reflect a mixture of visual contributions carried through earlier tokens, making it harder to isolate the evidence supporting the target concept. Consistent with this interpretation, our qualitative observations show increasingly dispersed and entangled attribution patterns as generation proceeds. We therefore formulate attribution as an immediate yes/no response, assessing visual support before these dependencies accumulate across a generated description. The qualitative comparison in Figures 21–24 shows that this formulation produces more focused attribution maps that align more closely with the relevant visual evidence.

C.2 Hallucination Mitigation Details

Hallucination mitigation uses two independently configurable depths: (1) the depth used to extract local regions for uncertainty estimation, and (2) the depth at which high-uncertainty patch tokens are removed during caption generation. We set both depths to the final vision-encoder layer and use a spatial radius of one for the selected patch operator. These choices follow the findings of the ablation study in Section A.

For regional uncertainty estimation, we use $K = 3$ uncertainty clusters, estimate uncertainty from a single generated token, and allow up to 256 new tokens during caption generation. The influence of these hyperparameters is further evaluated in the ablation study (Section A).

The regional uncertainty prompt is: "What object exists in the image? Answer with only the object name." Mitigated captions for all methods are generated using the prompt: "Describe the image in details."

D Runtime and Computational Complexity

We further compare the computational cost of VIRAH with existing attribution methods, with particular emphasis on EAGLE (Chen et al., 2026), the strongest-performing competing method in our attribution evaluations. We first provide an order-level comparison and then report wall-clock runtime under the same hardware and model configuration.

D.1 Complexity of VIRAH.

Let N denote the number of visual patches and let $k = \max_i |\mathcal{R}_i|$ denote the maximum number of retained patches in each local neighborhood. In our experiments, $k = 1$ for the No-Padding setting and $k = 9$ for the Radius-1 Box (3×3) setting. VIRAH processes the complete set of N visual tokens only once up to the intervention depth q. This prefix computation is shared across all regional evaluations, and the resulting representations are reused for every patch center. For each of the N patch centers, only the tokens in $\mathcal{R}_i$ are propagated through the remaining vision blocks and language

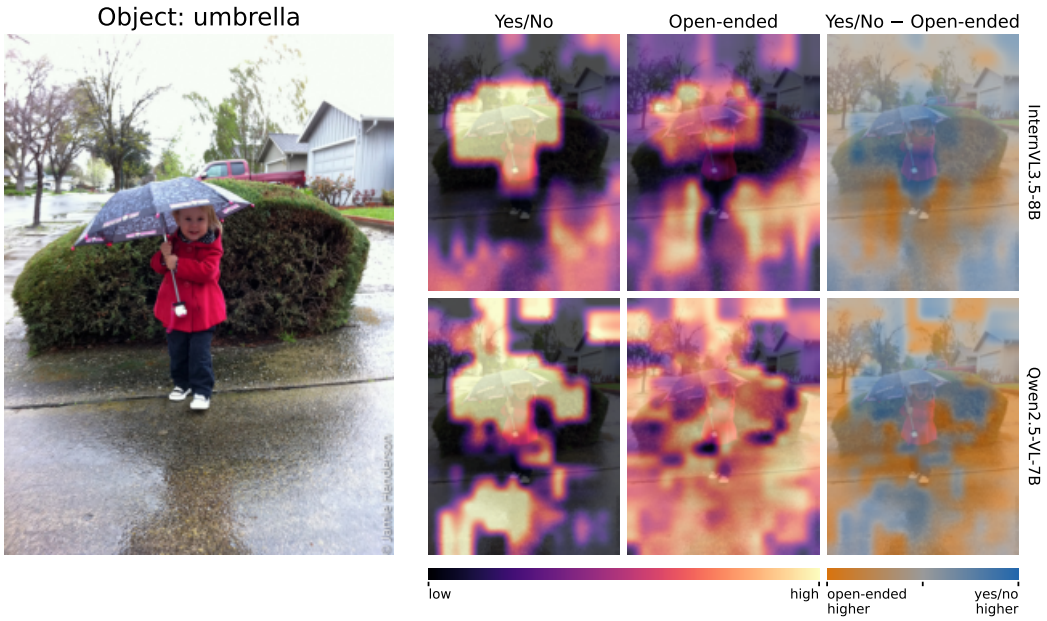

Figure 21: Qualitative comparison between open-ended generation and yes/no question prompt for a `umbrella`.

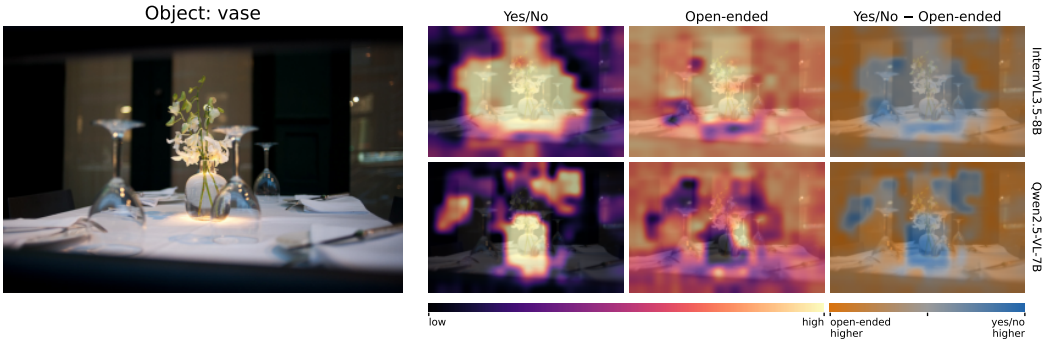

Figure 22: Qualitative comparison between open-ended generation and yes/no question prompt for a `vase`.

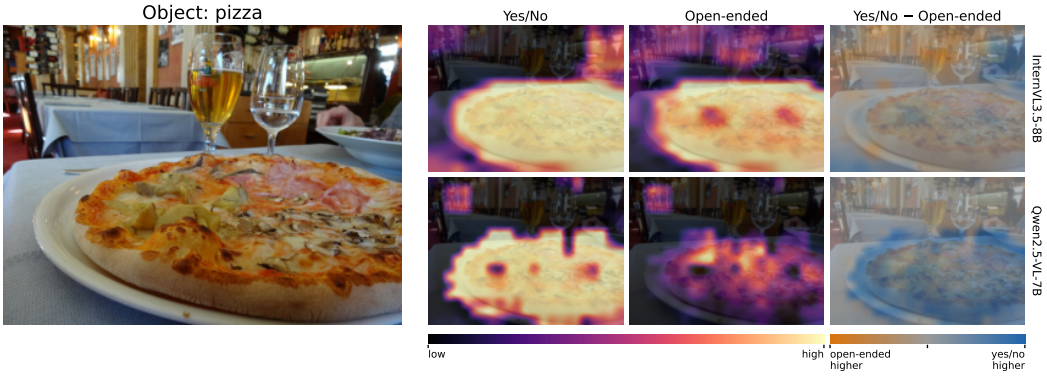

Figure 23: Qualitative comparison between open-ended generation and yes/no question prompt for a `pizza`.

decoder. Its computation can therefore be decomposed as

$$C_{\text{VIRAH}} = C_{\text{pre}}(N, q) + \sum_{i=1}^{N} C_{\text{post}}(|\mathcal{R}_i|, q), \tag{10}$$

where C_{pre} denotes the shared computation before intervention and C_{post} denotes the computation after intervention.

As a simple measure of regional visual-token workload,

$$\sum_{i=1}^{N} |\mathcal{R}_i| \leq Nk. \tag{11}$$

Thus, increasing the neighborhood from a single patch to a 3×3 neighborhood increases the number of retained token positions processed after intervention by at most a factor of nine. Since k is a small fixed constant in our experiments, the number of regional token evaluations remains linear in N. In practice, the wall-clock increase is additionally affected by batching, fixed language-decoder costs, and GPU utilization; we therefore report measured latency in Table 8.

D.2 Comparison with EAGLE.

EAGLE first partitions an image into M SLICO superpixels and performs a greedy search over combinations of these regions. Its greedy procedure requires

$$\frac{M(M+1)}{2} \tag{12}$$

model evaluations and therefore has $O(M^2)$ inference complexity (Chen et al., 2026). With the $M = 64$ configuration used in its main experiments, this corresponds to 2,080 model evaluations per image. In contrast, VIRAH requires N regional evaluations and has a regional workload of $O(Nk)$. Moreover, the computation of all N visual tokens up to the intervention depth q is performed only once and shared across these evaluations, whereas each candidate subset considered by EAGLE requires a separate model evaluation.

To make the comparison concrete, the average area of one EAGLE superpixel can be expressed in units of ViT patches. For an image containing N visual patches and partitioned into M superpixels, the average superpixel area corresponds to

$$\bar{s}_{\text{EAGLE}} \simeq \frac{N}{M} \tag{13}$$

ViT patch areas. With $N = 576$ and $M = 64$, one EAGLE superpixel therefore corresponds to approximately nine ViT patches on average, which matches the maximum neighborhood size used by VIRAH ($k = 9$). Under these settings, VIRAH performs 576 regional evaluations with at most nine retained visual tokens after a single shared prefix computation, whereas EAGLE performs 2,080 separate model evaluations during its greedy search. EAGLE superpixels are irregular and non-overlapping, while the neighborhoods in VIRAH are fixed and overlapping.

D.3 Comparison with Other Attribution Methods.

Moreover, LLaVA-CAM derives visual attribution from internal model activations and requires a single attribution pass, making it computationally efficient but dependent on model activations. TAM also derives attribution from activation information during generation, while IGOS++ formulates attribution as an iterative optimization problem requiring repeated forward/backward evaluations. In contrast, VIRAH directly intervenes on visual token representations, avoids iterative optimization and external region generation, and reuses the shared computation before the intervention depth across all regional evaluations.

D.4 Wall-clock runtime.

We measure end-to-end attribution time using the same model, images, prompts, numerical precision, and GPU for all methods. Model loading and disk I/O are excluded, while method-specific preprocessing is included.

Table 8: Computational efficiency comparison of attribution methods. We report the dominant evaluation cost and average attribution time per image under the Qwen2.5-VL-7B model.

Method	Evaluation strategy	# Evaluations	Order	Time (s/image)
LLaVA-CAM	Activation-based	51	$O(1)$	$19.9_{\pm\ 0.48}$
TAM	Activation-based	1	$O(1)$	$1.1_{\pm\ 0.21}$
IGOS++	Iterative optimization	T	$O(T)$	$60.2_{\pm\ 0.43}$
EAGLE ($M = 64$)	Superpixel greedy search	$\frac{M(M+1)}{2}$	$O(M^2)$	$481.0_{\pm\ 4.42}$
VIRAH ($k = 1$)	Regional evaluation	N	$O(Nk) = O(N)$	$38.7_{\pm\ 1.97}$
VIRAH ($k = 9$)	Regional evaluation	N	$O(Nk) = O(N)$	$41.9_{\pm\ 2.11}$

Table 8 reports the resulting per-image latency over Qwen2.5-VL-7B-Instruct. The $k = 9$ configuration incurs additional computation relative to $k = 1$ because more retained tokens are propagated after intervention, but the shared full-image prefix is computed only once. In contrast, EAGLE's greedy search requires repeated evaluations of separately constructed visual inputs. Under this controlled setting, VIRAH is approximately $11\times$ faster than EAGLE.

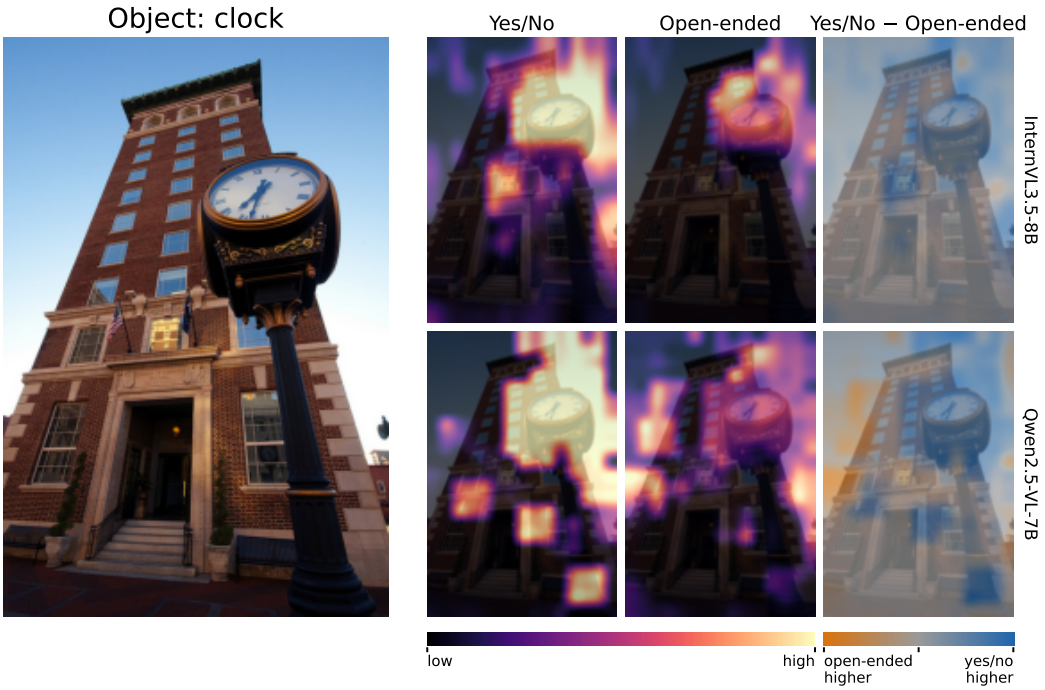

Figure 24: Qualitative comparison between open-ended generation and yes/no question prompt for a `clock`.