LOOKING IS NOT GROUNDING: RETHINKING ATTENTION-BASED HALLUCINATION DETECTION AND MITIGATION IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models frequently hallucinate objects that are not supported by the visual input, limiting their reliability in visually grounded generation. Existing attention-based approaches often treat visual attention distributions as evidence of object grounding, yet our analyses show that attention is systematically influenced by decoding context and semantic associations, while direct attention manipulation can induce response bias rather than reliably improve object verification. Motivated by this mismatch, we turn to alternative model-internal signals and introduce RASE, which combines **R**elational **A**ttention structures with native **S**emantic **E**vidence to reliably quantify visual grounding. During generation, high-risk object mentions are corrected through rollback and reselection from the model's original prediction distribution. Experiments across four LVLM backbones and multiple benchmarks show that RASE consistently improves object-level hallucination detection and mitigation, and generalizes across model architectures and datasets. Our code is available at `https://anonymous.4open.science/r/HalluAttn-1`.

1 INTRODUCTION

Large vision-language models (LVLMs) have demonstrated remarkable capabilities in visual understanding and language generation, enabling a wide range of applications that require reasoning over visual information. Yet LVLMs still frequently hallucinate objects absent from the input image, undermining their reliability in applications that require faithful visual grounding. Consequently, substantial efforts have been devoted to understanding and mitigating object hallucinations (Leng et al., 2024; Huang et al., 2024; Chen et al., 2024) with attention-based approaches (Yin et al., 2025; Hoang et al., 2026; Nguyen et al., 2026) emerging as a promising direction due to their ability to expose how models allocate information across visual and textual representations.

Attention distributions during generation have been widely treated as evidence of whether an object prediction is supported by visual input. Detection methods derive hallucination scores from attention statistics or learned attention patterns (Jiang et al., 2025b; Hoang et al., 2026; Zhang et al., 2025), whereas mitigation methods amplify or redistribute visual attention to strengthen reliance on image evidence (Yin et al., 2025; Kang et al., 2025). Despite their different mechanisms, these approaches share an implicit premise that the distribution of visual attention reflects object-specific grounding. However, our pilot studies suggest that attention-based signals are shaped not only by object-specific visual evidence, but also by decoding context and learned semantic associations. Moreover, directly manipulating attention can induce directional shifts in response preference rather than consistently improving object verification. These findings suggest that attention distributions reflect context-dependent information routing rather than object-specific visual support, making them an unreliable basis for hallucination detection and mitigation.

Motivated by this mismatch between attention distribution and object-specific visual support, we turn to alternative model-internal signals and introduce RASE, a framework that couples relational attention structures with native semantic evidence from the frozen LVLM. As shown in Figure 1, for every candidate object mention, RASE captures relational patterns across attention heads and extracts layer-wise semantic evidence from the LVLM's native visual representations, fusing them

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

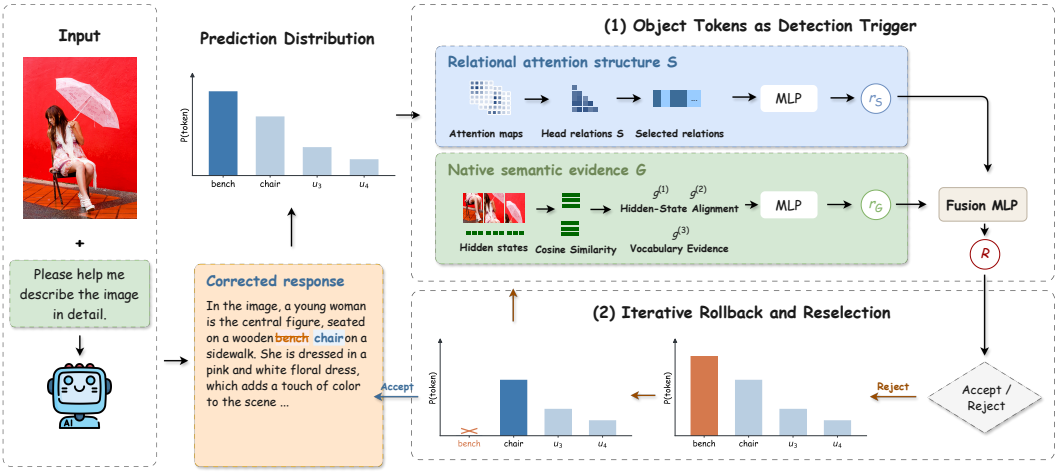


Figure 1: **Overview of RASE.** For each candidate object mention, RASE jointly evaluates relational attention structure and native semantic evidence from the LVLM to estimate its hallucination risk. High-risk candidates trigger rollback and reselection from the original prediction distribution, while accepted candidates are retained and generation continues.

into an object-level hallucination score during the forward pass processing the object token. During online generation, high-risk candidates trigger a rollback to the preceding state and reselection from the original prediction distribution to avoid direct attention manipulation, followed by identical risk checks for subsequent candidates.

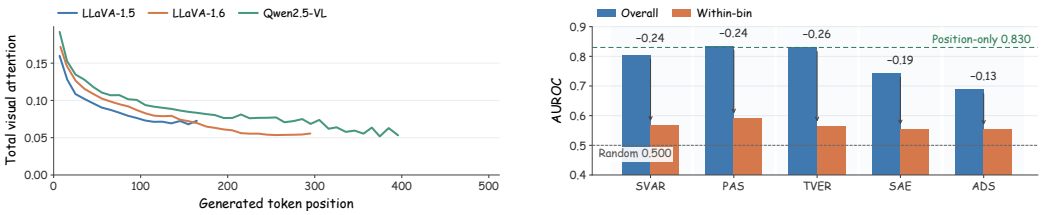
Experiments across four LVLM backbones show that RASE yields a consistently stronger object-level hallucination signal and translates this advantage into effective generation-time correction. For object-level detection, RASE achieves an average AUROC of **94.76**, outperforming existing baselines by a substantial margin. For mitigation, it consistently reduces object hallucinations across three benchmarks, demonstrating strong cross-architecture and cross-dataset generalizability while validating the utility of our grounding signal for online intervention.

Our contributions are summarized as follows:

- We revisit the assumption that visual attention reflects object-specific grounding, showing that attention-based hallucination signals are confounded by decoding context and semantic associations, while attention intervention can bias predictions rather than improve object verification.
- We introduce **RASE**, which combines relational attention structures with native semantic evidence from the frozen LVLM to estimate object-level hallucination risk, and uses this risk to guide rollback and reselection from the original prediction distribution.
- We extensively evaluate RASE across multiple LVLM backbones and benchmarks, demonstrating strong object-level hallucination detection and consistent hallucination reduction across diverse architectures and data distributions.

2 Related Work

Object Hallucination in LVLMs. Large vision-language models (LVLMs) can generate detailed visual descriptions and answer questions requiring visual reasoning, yet they may also produce objects or content that is unsupported by the visual input. Existing evaluations study this failure under complementary settings. CHAIR measures unsupported object mentions in open-ended image captions, whereas POPE formulates object hallucination as binary object-presence assessment (Rohrbach et al., 2018; Li et al., 2023). AMBER broadens evaluation to both generative and discriminative hallucination settings, while nocaps provides an open-vocabulary captioning setting for assessing generalization to novel object categories. As LVLMs extend from images to videos, hallucination evaluation has also expanded to dynamic content such as actions and events. MH-Bench and VidHalluc focus on motion and temporal inconsistencies, while Vript-HAL and ARGUS



(a) Visual attention by position (b) Position-controlled detection

Figure 2: **Generation-position bias in attention-based hallucination detection.** (a) Total visual attention systematically decreases with generation position across LVLMs. (b) Controlling for position substantially reduces the AUROC of five attention-based detectors, while generation position alone remains strongly predictive of hallucination.



(a) Object association prior (b) Attention intervention

Figure 3: **(a) Object association prior.** Absent objects can still attend to semantically associated regions through contextual co-occurrence or category similarity. **(b) Directional effects of attention intervention.** Scaling highly or weakly attended visual tokens induces systematic shifts in affirmative and negative response probabilities rather than consistent object verification.

evaluate hallucinated and omitted content in open-ended video descriptions (Kong et al., 2025; Li et al., 2025; Yang et al., 2024; Rawal et al., 2025).

Attention-Based Hallucination Detection and Mitigation. Attention-based hallucination detection has progressed from scalar statistics toward more structured analyses of cross-modal attention. Existing detectors exploit attention at different levels of granularity, from aggregate visual or contextual attention statistics to learned cross-modal attention structures (Jiang et al., 2025b; Hoang et al., 2026; Zhang et al., 2025). More recent work further augments these signals with semantic consistency between generated tokens and attended visual regions (Nguyen et al., 2026). For mitigation, attention-based methods directly intervene on internal visual pathways by amplifying image attention, rescaling cross-modal interactions, or redistributing attention toward informative visual tokens (Liu et al., 2024c; Yin et al., 2025; Kang et al., 2025). Together, these methods broaden the use of attention from scalar and spatial statistics to semantically informed signals, while leaving open how attention distribution relates to object-specific visual evidence.

3 Looking Is Not Grounding

Many attention-based hallucination methods interpret where and how strongly the model attends as evidence of visual grounding. This interpretation assumes that attention is primarily determined by object-specific visual support and that strengthening such attention should improve object verification. We examine whether these assumptions hold in practice and find three recurring limitations. Attention-based signals are confounded by generation position and can remain strong for semantically associated but absent objects, while attention intervention often shifts response preference rather than improving grounding.

Generation-Position Bias. Attention-based hallucination detection typically assumes that changes in attention reflect changes in visual grounding. However, attention also varies systematically during autoregressive generation. Figure 2a reports how the mean total visual attention of generated object tokens changes as the decoding context grows longer across three LVLMs on the

same COCO val2017 (Lin et al., 2014). This makes attention-derived scores naturally susceptible to generation-position bias, even when the underlying visual evidence remains unchanged. To quantify this effect, we evaluate SVAR, PAS, TVER, SAE, and ADS (Jiang et al., 2025b; Hoang et al., 2026; Wan et al., 2025; Song et al., 2026; Nguyen et al., 2026) under both standard and position-controlled settings. For the latter, we partition generated object tokens into bins according to their generation positions and evaluate detection performance within each bin. As shown in Figure 2b, their AUROC drops by 0.24, 0.24, 0.26, 0.19, and 0.13, respectively, after controlling for generation position, while generation position alone achieves an AUROC of 0.830. These results suggest that existing attention-based detectors partly benefit from position-related variation in attention rather than relying solely on visual grounding.

Object Association Prior. Even when attention is directed toward semantically relevant visual content, it does not necessarily provide object-specific evidence for the queried object. As illustrated in Figure 3a, an absent object can still attract substantial attention through learned associations with the visual scene. One source is **contextual co-occurrence**. When a traffic light is absent from a road scene, the model may nevertheless attend to roads, vehicles, or other traffic-related regions that frequently co-occur with traffic lights. Another source is **category similarity**. When queried about a bicycle, for example, the model may focus on a motorcycle because the two categories share visual structure, semantic attributes, and typical scene contexts. In both cases, the attended regions are relevant to the queried concept but do not constitute direct visual support for its presence. Thus, high or semantically meaningful attention can reflect learned object associations rather than genuine object-specific grounding.

Directional Effects of Attention Intervention. Attention-based mitigation methods often assume that increasing attention to selected visual tokens helps the model rely more on visual evidence and therefore reduces hallucination. However, it is unclear whether such intervention actually improves object verification or simply changes the model's response preference. To study this effect directly, we rank visual tokens by their attention weights and reweight the top 25%. Increasing the weights of highly attended visual tokens consistently raises the probability of an affirmative response, whereas increasing the weights of weakly attended visual tokens produces the opposite effect. As shown in Figure 3b, attention intervention therefore produces a clear directional shift in the output rather than a consistent improvement in visual grounding.

Table 1: Directional prediction changes induced by ClearSight.

Vanilla	ClearSight	Transition	Count
No	Yes	FN $\rightarrow$ TP	**83**
No	Yes	TN $\rightarrow$ FP	**53**
			+30

We observe the same pattern in ClearSight (Yin et al., 2025), which scales visual attention scores before softmax and thereby sharpens the resulting attention distribution, increasing the relative weight of highly attended visual tokens while suppressing weakly attended ones. On the POPE adversarial split, all 136 predictions changed by ClearSight move from *No* to *Yes*. Among them, 83 false negatives become true positives, while 53 true negatives become false positives, resulting in a spurious net gain of 30 correct predictions. This suggests that part of the reported improvement depends on the composition of the evaluation data, as the induced preference for affirmative answers yields a net gain only because more affected examples are false negatives than true negatives, rather than because object verification becomes more accurate.

4 METHOD

As demonstrated in Section 3, raw visual attention distributions alone do not provide reliable evidence of object-specific grounding. Rather than treating attention as a direct grounding signal, we retain the relational structure across attention heads and complement it with native semantic evidence from the model's visual representations. Based on this design, we propose RASE, which combines these two signals for hallucination detection and mitigation, as illustrated in Figure 1. For each candidate object mention, RASE fuses the structural and semantic evidence into an object-level hallucination risk score, which guides rollback and reselection during decoding.

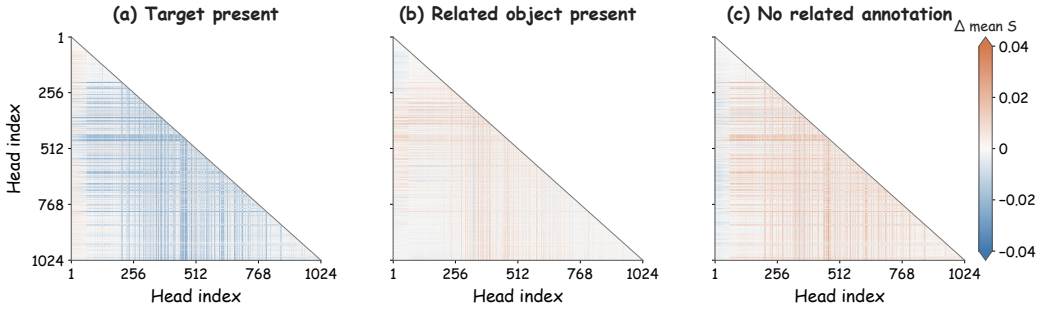


Figure 4: **Cross-head visual attention similarity under different levels of target-related visual support.** Each panel shows the mean pairwise similarity over 300 images after centering each head pair across the three conditions. Similarity increases for a subset of head pairs as target-specific visual support weakens, indicating reduced differentiation in their visual attention patterns.

4.1 Relational Attention Structure

Although attention magnitude and spatial allocation alone do not provide reliable evidence of visual grounding, this does not imply that attention encodes no information related to hallucination. Prior studies have shown that attention heads in LVLMs differ in their sensitivity to visual and linguistic context (Bi et al., 2025; He et al., 2025). This motivates us to decouple head-relational patterns from absolute magnitude, examining whether inter-head geometries remain diagnostic.

Inter-Head Similarity. Once an object token y_t is generated, we collect the attention distribution from each head during the subsequent forward pass that consumes y_t. For each attention head h, we normalize its attention over the visual token set $\mathcal{V}$, removing the influence of total visual-attention mass while preserving the relative allocation pattern within the visual sequence. We then measure the similarity between each pair of heads with the Bhattacharyya similarity (Bhattacharyya, 1943):

$$S_{ij} = \sum_{v \in \mathcal{V}} \sqrt{p_i(v) p_j(v)}, \qquad 1 \le i < j \le H. \tag{1}$$

where H is the total number of attention heads across decoder layers. A larger S_{ij} indicates greater spatial similarity between the visual attention distributions of heads i and j. As shown in Figure 4, the pairwise similarities form the relational matrix $\mathbf{S}$, where some head pairs exhibit lower similarity when the target object is present and shift toward higher similarity as target-specific visual support weakens. These shifts are more evident in certain head pairs, indicating that changes in target-specific visual support are captured more strongly by these relations. This motivates us to isolate them as a structural signal for hallucination detection.

Grounded-Reference Relation Selection. Crucially, Figure 4 reveals that head pairs exhibiting the lowest similarity under grounded generation also undergo the most obvious distributional shifts as visual evidence degrades. This suggests that head-pair relations exhibiting the strongest differentiation under grounded generation are particularly sensitive to the loss of visual support and provide useful structural signals for hallucination detection. Guided by this observation, we directly identify these relations from grounded training mentions. For each head pair, we compute its mean similarity

$$\bar{S}_{ij} = \frac{1}{|\mathcal{D}|} \sum_{n \in \mathcal{D}} S_{ij} \tag{2}$$

where $\mathcal{D}$ denotes the grounded mentions in the training set. We retain the bottom 1% of head pairs with the smallest $\bar{S}_{ij}$. For each candidate object, the similarities of the selected head pairs form the structural feature vector. A lightweight MLP then maps this to a scalar structural risk score r_S, which is combined with the semantic evidence in the final detector.

4.2 NATIVE SEMANTIC EVIDENCE

While r_S monitors relational information routing across attention heads, the LVLM's internal hidden states provide complementary evidence regarding the semantic alignment between the generated object token and the visual representations. To capture this evidence across the model's representational depth, we extract three diagnostic semantic features from each decoder layer, summarizing local similarity, attention–semantic alignment, and native vocabulary support.

Hidden-State Semantic Compatibility. To verify whether the visual sequence contains regions semantically aligned with the generated object, we first evaluate the cosine similarity between the text hidden state $\mathbf{h}_t$ and each visual hidden state $\mathbf{h}_v$.

$$c(v) = \cos(\mathbf{h}_t, \mathbf{h}_v)\,. \tag{3}$$

We then characterize semantic compatibility from two complementary perspectives. On the one hand, to assess whether the generated object is supported by visual representations, we aggregate the cosine similarities over the subset of visual positions with the strongest alignment.

$$g^{(1)} = \frac{1}{|\mathcal{C}|}\sum_{v\in\mathcal{C}} c(v), \qquad \mathcal{C} = \{v \in \mathcal{V} \mid c(v) \text{ ranks in the top } 5\%\}\,. \tag{4}$$

On the other hand, to verify whether the model actively attends to semantically compatible visual evidence, we measure whether the model's attention focuses on visually relevant content. Letting $w(v)$ denote the head-averaged visual attention distribution, we compute:

$$g^{(2)} = \sum_{v\in\mathcal{V}} w(v)c(v) - \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} c(v). \tag{5}$$

Native Vocabulary Evidence. We further assess whether semantically relevant visual representations support the generated object in the model's vocabulary space. We average the probability assigned to the generated object token y over the most semantically relevant visual tokens.

$$g^{(3)} = \frac{1}{|\mathcal{C}|}\sum_{v\in\mathcal{C}} \mathrm{Softmax}\left(\mathrm{Decode}(\mathbf{h}_v)\right)_y, \qquad \mathcal{C} = \{v \in \mathcal{V} \mid c(v) \text{ ranks in the top } 5\%\} \tag{6}$$

Finally, we concatenate the three signals across all L decoder layers to obtain

$$\mathbf{G} = \left[g^{(1)}, g^{(2)}, g^{(3)}\right]_{\ell=1}^{L} \in \mathbb{R}^{3L}. \tag{7}$$

The resulting vector $\mathbf{G}$ summarizes layer-wise semantic evidence from the LVLM's native visual representations. Similarly, a lightweight MLP maps this vector to a scalar semantic risk score r_G, which is subsequently fused with the structural risk score r_S in the final detector.

4.3 RISK-GUIDED ROLLBACK AND RESELECTION

Having obtained the structural risk r_S and semantic risk r_G, we combine the two signals using a single-layer fusion MLP to produce the final hallucination risk R. This forms an object-level detector that jointly considers relational attention structure and native semantic evidence. As revealed in Section 3, directly manipulating attention can introduce directional response bias. We therefore intervene at token selection, rolling back whenever the detector assigns high risk to a candidate object and selecting an alternative from the model's original prediction distribution.

Object Tokens as Detection Trigger. During autoregressive generation, each newly generated object token is treated as a candidate and is accepted into the output only after validation by the hallucination detector. This token is consumed in the next forward pass to extract the diagnostic signals and evaluate its hallucination risk without modifying the normal decoding pipeline. If the candidate is flagged as hallucinated, decoding rolls back to the state before the token was selected. Otherwise, the token is accepted and generation continues normally.

Iterative Rollback and Reselection. When a candidate is flagged as hallucinated, decoding returns to the state before the token was selected. The rejected token is added to a prefix-specific ban set $\mathcal{B}(Y_{<t})$ and excluded from subsequent selection. A replacement token is then reselected from the saved original prediction distribution after masking all rejected tokens,

$$y'_t = \mathrm{Decode}\left(p_t(\cdot) \mid y \notin \mathcal{B}(Y_{<t})\right), \tag{8}$$

where $\mathrm{Decode}(\cdot)$ denotes the backbone's original decoding strategy.

5 Experiments

Evaluation Protocol. We evaluate RASE on four LVLM backbones: LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-v1.6-Mistral-7B (Liu et al., 2024b), Qwen2.5-VL-7B (Bai et al., 2025b), and Qwen3-VL-8B (Bai et al., 2025a). For object-level detection, RASE and all learned baselines are trained on 20,000 images sampled from the COCO 2017 training set (Lin et al., 2014). For each backbone, we fit a separate detector and perform all calibration, including operating-threshold selection, using training data only.

Datasets, Metrics, and Baselines. For object-level hallucination detection, we compare RASE against IC (Jiang et al., 2025a), SVAR (Jiang et al., 2025b), GLSim (Park & Li, 2025), PAS (Hoang et al., 2026), ADS+CGC (Nguyen et al., 2026), and LLCC (Wang et al., 2026), with AUROC, accuracy, precision, and F1 reported. For hallucination mitigation, we evaluate open-ended image captioning on COCO 2017 validation set, AMBER-G (Wang et al., 2023), and the NoCaps (Agrawal et al., 2019) validation set. On COCO, we report CHAIR_s, CHAIR_i, and correct-object coverage. On AMBER-G, we report the official CHAIR, Hal, Cover, and Cog metrics. On NoCaps, we use the same CHAIR and coverage metrics as on COCO. We compare against VCD (Leng et al., 2024), OPERA (Huang et al., 2024), PAI (Liu et al., 2024c), ClearSight (Yin et al., 2025), PTI (Zhang et al., 2026), and SADT (Wang et al., 2026).

5.1 Comparison with Hallucination Detectors

We first evaluate whether RASE provides a reliable object-level signal for distinguishing grounded from hallucinated mentions across different LVLM backbones. The results are reported in Table 2.

Overall Performance. RASE achieves an average AUROC of 94.76 across the four backbones, outperforming the best baseline on each backbone by 7.46–12.39 points. The advantage remains pronounced at the selected operating thresholds, where RASE reaches an average precision of 96.46 and improves over the strongest baseline by 10.29–12.41 points, together with consistent gains in accuracy and F1. The large improvement in precision indicates that RASE is substantially less likely to accept hallucinated object mentions as visually grounded, providing a reliable signal for subsequent rollback-based intervention.

Cross-Backbone Consistency. Beyond its overall performance, RASE exhibits little variation across backbone architectures, with AUROC differing by only 2.41 points between its highest and lowest results. Several baselines are considerably more sensitive to the underlying LVLM. For example, GLSim decreases from 83.27 AUROC on LLaVA-1.5 to 36.41 on Qwen3-VL. This stability suggests that the grounding signal learned by RASE is less sensitive to backbone-specific representations and remains effective across both LLaVA and Qwen model families.

5.2 Comparison with Hallucination Mitigation Methods

We next evaluate whether the object-level hallucination signal learned by RASE translates into effective generation-time mitigation. Figure 5 compares RASE with vanilla decoding and existing mitigation methods across four LVLM backbones on COCO, AMBER-G, and NoCaps.

Overall Performance and Transfer. RASE consistently reduces object hallucination across all twelve dataset–backbone settings while largely preserving the correct-object coverage of vanilla decoding. On COCO, it reduces CHAIR_s and CHAIR_i by an average of 25.44 and 6.74 percentage points relative to vanilla decoding, respectively, while decreasing coverage by only 1.37 points

Table 2: Object-level hallucination detection on the COCO 2017 validation set across four LVLM backbones. **Bold** and underlined values indicate the best and second-best results, respectively.

Backbone	Metric	IC	SVAR	GLSim	PAS	ADS+CGC	LLCC	RASE (Ours)
LLaVA-1.5	AUROC	59.23	80.20	83.27	83.57	85.99	53.95	**93.45**
7B	Accuracy	75.01	75.27	77.87	77.93	81.00	73.70	**84.70**
	Precision	75.01	75.29	79.40	80.77	85.51	75.54	**95.97**
	F1	85.72	85.82	86.59	86.29	87.65	84.56	**89.07**
LLaVA-v1.6	AUROC	77.67	76.31	74.42	73.89	85.51	56.47	**95.86**
Mistral-7B	Accuracy	85.12	83.65	84.67	82.73	85.26	81.10	**89.03**
	Precision	85.10	84.29	85.11	82.74	87.42	82.65	**97.71**
	F1	91.70	90.89	91.42	90.54	91.51	89.53	**93.05**
Qwen2.5-VL	AUROC	59.79	74.25	62.79	73.50	82.04	49.68	**94.43**
7B	Accuracy	81.95	81.99	82.02	81.99	83.54	20.63	**87.49**
	Precision	81.98	81.99	82.14	82.16	85.83	59.75	**96.88**
	F1	90.08	90.10	90.09	90.08	90.51	16.75	**91.98**
Qwen3-VL	AUROC	60.83	76.72	36.41	69.74	83.36	43.76	**95.30**
8B	Accuracy	74.43	76.13	74.37	74.36	80.56	50.41	**88.60**
	Precision	74.43	77.15	74.37	74.37	82.88	69.76	**95.29**
	F1	85.32	85.74	85.30	85.30	87.69	63.82	**92.07**

on average. The same trained detectors transfer effectively to AMBER-G and NoCaps without benchmark-specific adaptation. On AMBER-G, RASE reduces Hal by 36.90 points on average and achieves the lowest CHAIR, Hal, and Cog across all four backbones. On NoCaps, it reduces CHAIR_s and CHAIR_i by an average of 29.25 and 7.07 points, respectively, and achieves the lowest CHAIR_i on every backbone. Across all dataset–backbone settings, the reduction in correct-object coverage relative to vanilla never exceeds 2.41 points. These results show that the learned hallucination signal can reliably guide generation-time correction across both LVLM architectures and image-captioning distributions.

Hallucination–Coverage Trade-off. Lower hallucination rates do not necessarily indicate better captioning when they are achieved by suppressing valid object mentions. This trade-off is evident for several competing methods. On LLaVA-1.5, for example, PAI obtains a slightly lower COCO CHAIR_s than RASE (21.66% vs. 22.70%), but its correct-object coverage drops to 62.35%, compared with 75.29% for RASE. The trade-off becomes more pronounced on Qwen3-VL, where PAI reaches 18.06% CHAIR_s compared with 25.06% for RASE, while retaining only 56.45% coverage compared with 76.59% for RASE. A similar pattern appears on NoCaps, where lower caption-level hallucination can also come at the cost of reduced object coverage. Thus, RASE does not minimize every hallucination metric in every setting, but generally achieves a more favorable balance between hallucination reduction and preservation of grounded object content.

5.3 Extension to Video Understanding

To evaluate whether RASE generalizes beyond static-image settings, we extend it to video using Qwen2.5-VL-7B and evaluate object-level hallucination mitigation on Vript-HAL. The video detector is trained on VidOR and operates on native visual tokens from 16 sampled frames, while the same rollback-based intervention is retained during generation. As shown in Table 3, RASE achieves the highest precision and F1 among all compared methods. Relative to vanilla decoding, it improves object precision from 72.39% to 84.06% while nearly preserving recall (43.15% vs. 42.49%), resulting in an F1 increase from 54.07% to 56.45%. These results show that the grounding signal captured by

Table 3: Video object-hallucination mitigation on Vript-HAL with Qwen2.5-VL-7B.

Method	Precision ↑	Recall ↑	F1 ↑
Vanilla	72.39	43.15	54.07
VCD	68.49	46.46	55.36
MotionCD	69.42	42.30	52.57
RASE (Ours)	84.06	42.49	56.45

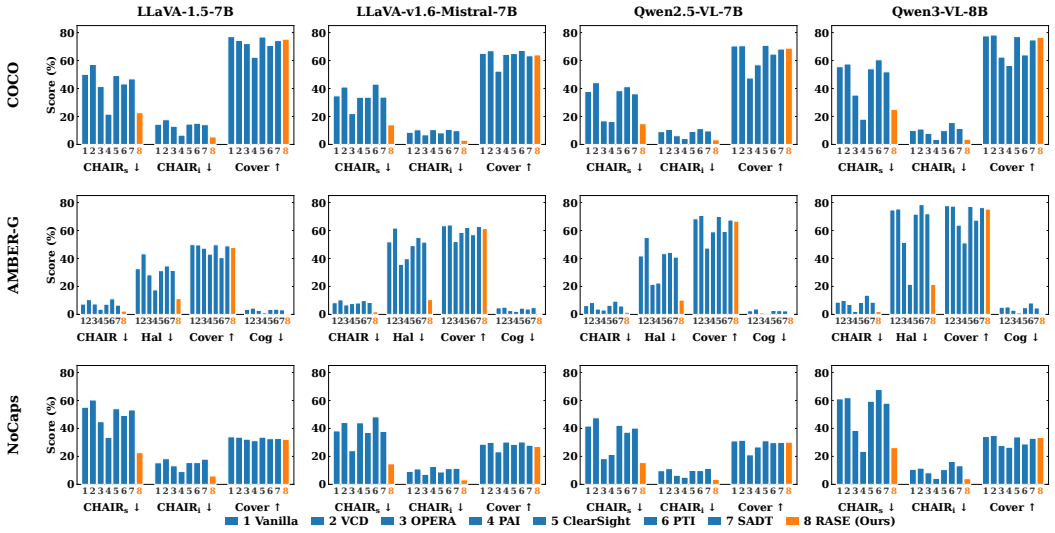


Figure 5: Object-hallucination mitigation results on COCO, AMBER-G, and the NoCaps validation set across four LVLM backbones. COCO and NoCaps report CHAIR$_s$, CHAIR$_i$, and correct-object coverage, while AMBER-G reports CHAIR, Hal, Cover, and Cog. Lower is better for hallucination metrics, whereas higher is better for coverage.

RASE remains effective in the video setting and can reduce unsupported object generation without substantially sacrificing object recall.

5.4 Ablation of Structural and Semantic Evidence

Table 4 evaluates the contributions of relational attention structure (S) and native semantic evidence (G). Both signals are individually effective, with S alone outperforming G alone across all metrics (94.80 vs. 93.47 AUROC and 91.71 vs. 90.50 F1), suggesting that cross-head relational structure provides a particularly strong cue for object grounding. Combining the two yields the best performance, reaching 95.86 AUROC, 97.71 precision, and 93.05 F1. Compared with S alone, adding G further improves AUROC, precision, and F1 by 1.06, 1.17, and 1.34 points, respectively. These results indicate that native semantic evidence provides complementary information beyond the structural signal, while relational attention structure forms the stronger standalone component of RASE.

Table 4: Detector feature ablation on LLaVA-v1.6-Mistral-7B.

Input	AUROC	Precision	F1
S only	94.80	96.54	91.71
G only	93.47	96.23	90.50
$S + G$	**95.86**	**97.71**	**93.05**

6 Conclusion

We revisit whether visual attention can serve as reliable evidence of object-specific grounding in LVLMs and find that this assumption does not hold consistently in practice. Motivated by this gap, we introduce RASE, which derives a more reliable grounding signal from model-internal structure and semantic evidence and uses it to support generation-time correction. Across diverse models and evaluation settings, RASE consistently improves hallucination detection and mitigation while preserving grounded visual content. Its effectiveness beyond the original image setting further suggests that the learned signal is not tied to a particular architecture or data distribution. Overall, our findings indicate that reliable visual grounding requires going beyond where a model attends and assessing whether its internal representations provide genuine support for the generated content.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, stylistic refinement, and improving the clarity and readability of the manuscript. All AI-assisted content was carefully reviewed, verified, and revised by the authors, who take full responsibility for the final manuscript.

REFERENCES

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8948–8957, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Agrawal_nocaps_novel_object_captioning_at_scale_ICCV_2019_paper.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 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. *arXiv preprint arXiv:2502.13923*, 2025b.

Anil Kumar Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4135–4144, 2025. doi: 10.1109/CVPR52734.2025.00391.

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*, pp. 7824–7846. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24bi.html.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in LVLMs with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3488–3501. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.175.

Nhat Hoang, 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*, 2026. URL https://arxiv.org/abs/2511.11502.

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. URL https://arxiv.org/abs/2311.17911.

Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2410.02762.

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. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.03321.

Ming Kong, Xianzhou Zeng, Luyuan Chen, Yadong Li, Bo Yan, and Qiang Zhu. MHBench: Demystifying motion hallucination in VideoLLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4401–4409, 2025. doi: 10.1609/aaai.v39i4.32463. URL https://ojs.aaai.org/index.php/AAAI/article/view/32463.

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. URL https://arxiv.org/abs/2311.16922.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025. URL https://arxiv.org/abs/2412.03735.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

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*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision*, 2024c. URL https://arxiv.org/abs/2407.21771.

Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Yutong Xie, Cam-Tu Nguyen, 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2604.04863.

Seongheon Park and Sharon Li. GLSim: Detecting object hallucinations in LVLMs via global-local similarity. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3640a1997a4c9571cea9db2c82e1fc35-Abstract-Conference.html.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. ARGUS: Hallucination and omission evaluation in Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20280–20290, 2025. URL https://arxiv.org/abs/2506.07371.

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. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Jiale Song, Jiaxin Luo, Xue-Song Tang, Kuangrong Hao, and Mingbo Zhao. Segmentation-based attention entropy: Detecting and mitigating object hallucinations in large vision-language models. *arXiv preprint arXiv:2603.16558*, 2026.

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

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. *arXiv preprint arXiv:2311.07397*, 2023. URL https://arxiv.org/abs/2311.07397.

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*, 2026. URL https://arxiv.org/abs/2608.07302.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6903a5aaece71b76623245fc6e32f01b-Abstract-Datasets_and_Benchmarks_Track.html.

Hao Yin, Guangzong Si, and Zilei Wang. ClearSight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.13107.

Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. Prefill-time intervention for mitigating hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2604.25642.

Yudong Zhang, Ruobing Xie, Xingwu Sun, Yiqing Huang, Jiansheng Chen, Zhanhui Kang, Di Wang, and Yu Wang. DHCP: Detecting hallucinations by cross-modal attention pattern in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. URL https://arxiv.org/abs/2411.18659.

A Implementation Details for RASE

Table 5: Model-dependent feature dimensions of RASE. S denotes the number of selected head-pair relations and G the dimension of the semantic feature vector.

Backbone	$L \times h$	S	G
LLaVA-1.5-7B	32×32	5,238	96
LLaVA-v1.6-Mistral-7B	32×32	5,238	96
Qwen2.5-VL-7B	28×28	3,070	84
Qwen3-VL-8B	36×32	6,630	108

Let $H = Lh$ denote the total number of attention heads for a backbone with L decoder layers and h heads per layer. For the structural branch S, we compute pairwise similarities between head-wise visual attention distributions and retain the bottom 1% of head pairs according to their mean similarity on grounded training mentions. The semantic branch G contains three features per decoder layer, giving a $3L$-dimensional representation; the top 5% most cosine-aligned visual tokens are used for the local semantic and vocabulary evidence. All diagnostic features are extracted after consuming the object token of each recognized object mention.

B Additional Object-Association Visualizations

Figure 6 provides additional qualitative examples complementing Figure 3a. Each pair presents an input image alongside the corresponding attention overlay for the queried object. These examples further show that, even when the queried object is absent, the model may attend to semantically related scene content, reinforcing the contextual co-occurrence and category-similarity patterns discussed in the main text.

C Qualitative Examples of Hallucination Mitigation

Figures 7–10 present four selected COCO val2017 examples for each backbone, comparing the saved vanilla caption with the corresponding RASE output generated from the same image and prompt ("Please help me describe the image in detail."). The excerpts preserve the original wording, with `[...]` indicating omitted text. Red highlights unsupported object mentions, while teal highlights selected visual content that is retained after mitigation. Examples were selected based on the stored outputs and COCO/CHAIR annotations, followed by manual inspection of the source images and noun contexts to exclude lexical false positives. These examples are intended to illustrate the removal of specific object hallucinations.

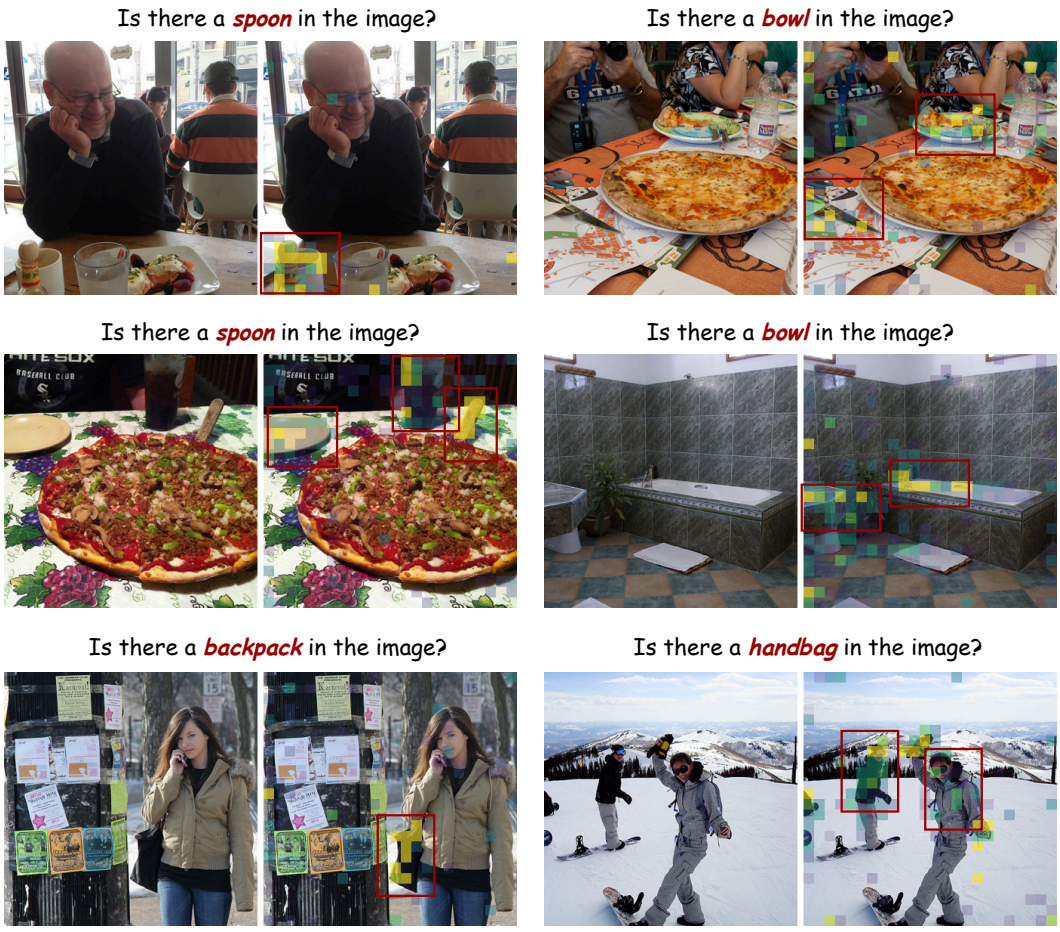


Figure 6: **Additional examples of the object association prior.** Each image is paired with its attention overlay, with red boxes marking selected regions. Attention to related visual content does not by itself establish the presence of the queried object.

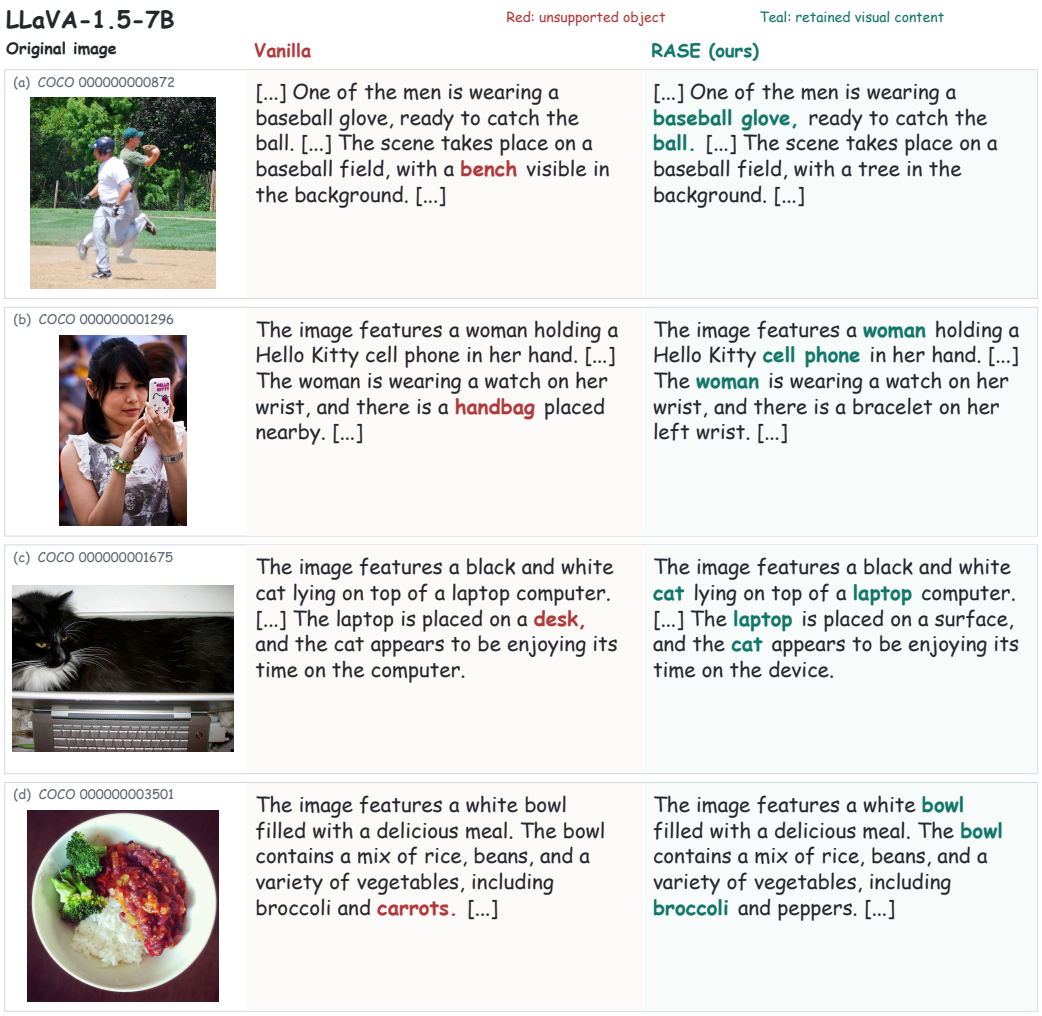


Figure 7: **Hallucination mitigation on LLaVA-1.5-7B.** Four original image–caption pairs illustrate changes from vanilla decoding to RASE. Colors annotate selected phrases only.

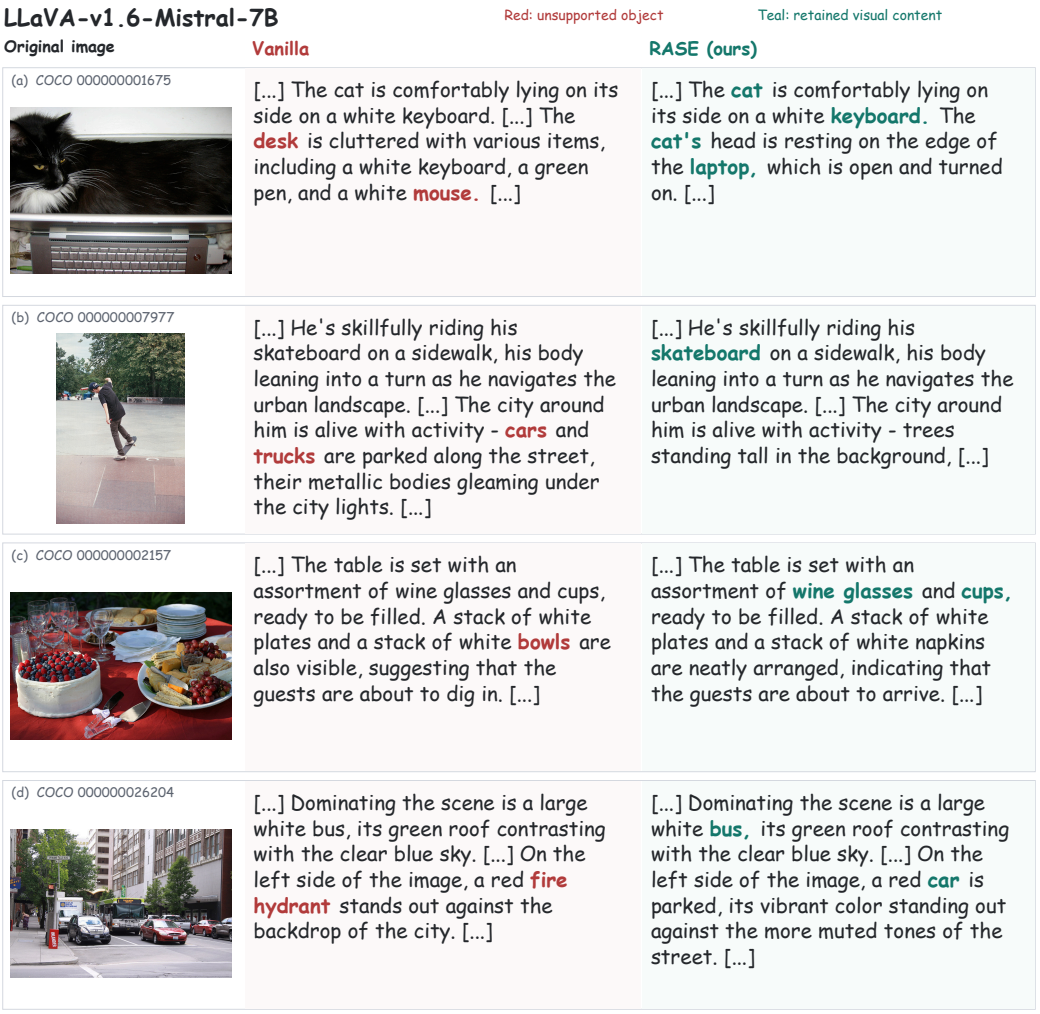


Figure 8: **Hallucination mitigation on LLaVA-v1.6-Mistral-7B.** Four selected examples, with the same layout and annotation convention as Figure 7.

Qwen2.5-VL-7B Red: unsupported object Teal: retained visual content

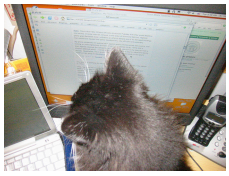
Original image	Vanilla	RASE (ours)
(a) *COCO* 000000004795 	The image shows a close-up of a black and gray cat sitting in front of a computer monitor, partially obscuring it. [...] The background includes some office or home workspace elements, such as a **chair** and possibly other equipment. [...]	The image shows a close-up of a black and gray **cat** sitting in front of a computer **monitor.** [...] To the left of the **monitor,** there is a **laptop,** and to the right, there is a landline **telephone** with a digital display and buttons. [...]
(b) *COCO* 000000005001 	[...] The scene is vibrant and festive, with a group of people gathered around a **table** where a ceremonial ribbon is being cut. [...] One child is holding a pair of scissors, poised to cut the ribbon, while the other child stands nearby, also wearing a helmet. [...]	[...] A group of **people,** including adults and children, are gathered around a red ribbon stretched across a small platform. [...] He is wearing a red plaid shirt and a blue t-shirt with a logo on it, and he is holding a pair of large **scissors.** [...]
(c) *COCO* 000000010092 	[...] The bed is neatly made with white bedding. [...] On either side of the bed, there are small wooden nightstands. One of them holds a lit candle in a black holder, adding a warm glow to the room. The nightstand also has some papers or **books** placed on it. [...]	[...] The **bed** is neatly made with white bedding. [...] On the left side of the room, there is a small wooden nightstand with a lit candle on it, adding a warm glow to the space. The right side of the room has a wooden **table** with **chairs,** suggesting a dining or work area. [...]
(d) *COCO* 000000038048 	Certainly! The image shows an urban street scene with a red fire hydrant prominently placed on a sidewalk. [...] To the right of the fire hydrant, there is a person sitting on a **bench.** [...]	Certainly! The image shows a bright red **fire hydrant** situated on a sidewalk in an urban or suburban setting. [...] To the right of the hydrant, there is a **person** sitting on a low wall or ledge, partially obscured by the hydrant. [...]

Figure 9: **Hallucination mitigation on Qwen2.5-VL-7B.** Four selected examples compare unchanged excerpts from the stored vanilla and RASE outputs.

Qwen3-VL-8B Red: unsupported object Teal: retained visual content

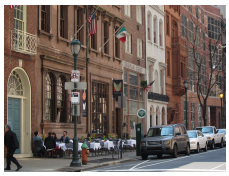
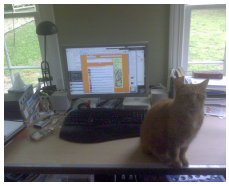
Original image	Vanilla	RASE (ours)
(a) *COCO* 000000055167 	[...] The bus is painted in a bright yellow and blue livery, with a blue front bumper and lower section. [...] The foreground features a well-maintained grassy verge with colorful flowerbeds containing purple, yellow, and red flowers, and a small black **bench.** [...]	[...] The **bus** is painted in a bright yellow and blue livery, with a blue front bumper and lower section. [...] The foreground features a well-maintained grassy area with some colorful flowerbeds containing purple, yellow, and red flowers, and a small black metal garden ornament. [...]
(b) *COCO* 000000067616 	[...] On the sidewalk, several people are seated at outdoor tables under **umbrellas,** enjoying meals or drinks. The tables are covered with white tablecloths, and the chairs are simple, dark metal. [...]	[...] On the sidewalk, several **people** are seated at outdoor **tables** under the shade of trees, enjoying meals or drinks. The **tables** are covered with white tablecloths, and the **chairs** are simple, dark metal. [...]
(c) *COCO* 000000070254 	[...] On the left, a woman in a white t-shirt and dark pants stands holding a large, bright red shopping bag. [...] Next to her, a man in a blue and white plaid shirt sits on a bench, holding a phone or small device. A black **dog** sits beside him on the ground. [...]	[...] On the left, a **woman** in a white t-shirt and dark pants stands holding a large, bright red shopping bag. [...] Next to her, a **man** in a blue and white plaid shirt sits on a bench, holding a small object, possibly a camera or a small device. [...]
(d) *COCO* 000000316015 	This is a candid, slightly cluttered photograph of a home office desk, with a ginger cat as the central subject. [...] Below the monitor is a black keyboard. [...] To the right of the monitor, there is a white **computer mouse.** [...]	This is a candid, slightly cluttered photograph of a home office desk, with a ginger **cat** as the central subject. [...] Below the monitor is a black **keyboard.** [...] A white, cylindrical object, possibly a second speaker or a small fan, is to the right of the **keyboard.** [...]

Figure 10: **Hallucination mitigation on Qwen3-VL-8B.** Four selected examples compare unchanged excerpts from the stored vanilla and RASE outputs.