TRACE: CAUSAL TRACING AND EDITING VISUAL EVIDENCE RETRIEVAL TO MITIGATE VISION-LANGUAGE MODEL HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) have demonstrated strong capabilities but remain prone to object hallucinations. Existing training-free methods focus on modifying visual processing or answer generation, rather than causally identifying the sources of hallucinations. We are the first to causally trace the information flow underlying hallucinations and discover *Visual Retrieval Heads*: attention heads that ground visual evidence at queried-object positions to causally influence predictions. We also discover that evidence retrieved by these heads significantly contributes to the model's confidence. When this evidence comes from a similar but incorrect region, it can make a false object strongly supported, leading to an overconfident hallucination. Yet, existing *non-causal* methods minimally intervene in this retrieval process. As such, they cannot correct overconfident hallucinations. Based on this insight, we propose *TRACE*, a *causal* intervention method that corrects how visual evidence is gathered by the identified heads. Experiments on object-existence prediction and image captioning benchmarks show that *TRACE* significantly outperforms prior methods by up to 11.7%.

1 INTRODUCTION

Large vision-language models (LVLMs) combine visual perception with language generation to describe scenes and answer questions about images (Liu et al., 2024). Yet, a model may confidently report an object that is absent from the image, a failure known as *object hallucination* (Li et al., 2023). Such errors reveal a mismatch between the model's claims and the visual evidence: the model may retrieve misleading information and treat it as evidence of an object's presence. Mitigating this failure requires understanding how the model retrieves visual evidence to support its object claims.

Existing methods mitigate hallucination by intervening in downstream answer generation or visual inputs. First approach adjusts logits, activations, or attention, as in VCD (Leng et al., 2024), OPERA (Huang et al., 2024), hallucination-head suppression (Yang et al., 2025), ONLY (Wan et al., 2025), VISTA (Li et al., 2025), and ILVAD (Xie et al., 2026). The second improves visual grounding and cross-modal alignment, as in AGLA (An et al., 2025), SHIELD (Huang et al., 2026), CausalMM (Zhou et al., 2025), and PTI (Zhang et al., 2026). In short, existing methods largely intervene in downstream components near answer generation, where hallucinations are ultimately expressed. However, our analysis shows that this is not where hallucinations are formed. In correct and hallucinated cases, the model follows the same information flow, with hallucinations arising when erroneous information emerges at an intermediate stage and propagates to the final answer (Figure 1). Thus, no fixed component is inherently hallucinatory: the same head may contribute to a hallucination in one example but behave normally in another. This exposes a limitation: whether modifying visual inputs or downstream generation, they do not directly target the stage where errors first arise.

To address this limitation, our key idea is to causally reconstruct the information flow underlying hallucinations (Figure 2a). *With* causality, we discover *Visual Retrieval Heads*: at a specific token, these heads retrieve visual evidence that causally influences the prediction (Figure 1). The evidence retrieved by these heads significantly contributes to the model's confidence in its answer (Figure 1b). When this evidence comes from a visually similar but incorrect region, the model can use it as strong support for the presence of an absent object, resulting in an overconfident hallucination.

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.

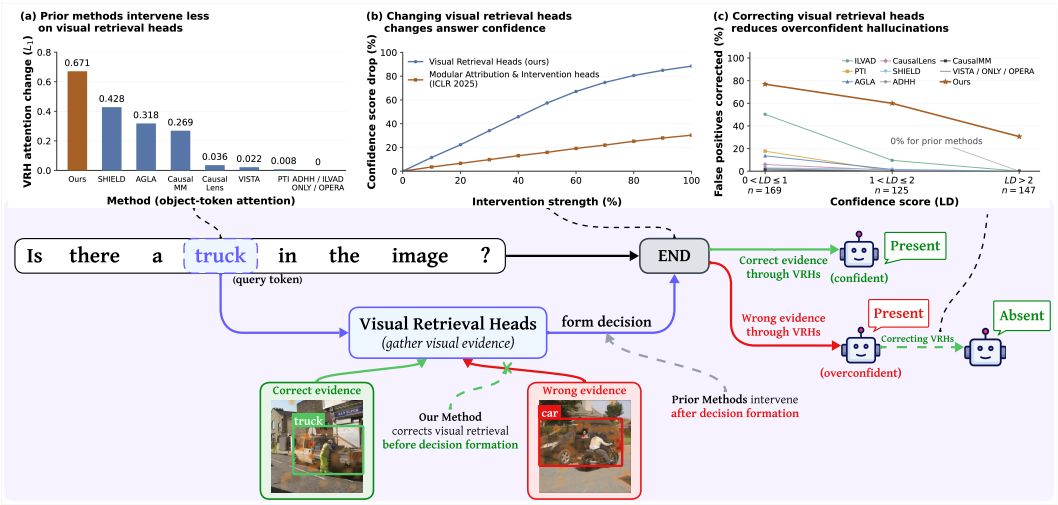


Figure 1: **From evidence retrieval to confident hallucination.** We uncover the causal information flow behind object-existence decisions in LLaVA-1.5-7B (Liu et al., 2024) and discover *Visual Retrieval Heads*: these heads retrieve visual evidence that the model uses to decide object presence. *TRACE* corrects this retrieval before decision formation. (a) *TRACE* intervenes on these heads' attention more than prior methods (mean L_1: 0.671 vs. at most 0.428). (b) On POPE (Li et al., 2023), we use $\text{LD} = \text{logit}(\text{Yes}) - \text{logit}(\text{No})$ as proxy for the model's confidence when predicting an object is present. Intervening on these heads reduces answer confidence more than intervening on the hallucination heads of Yang et al. (2025). (c) *TRACE* corrects up to 30% of confident hallucinations (incorrectly predicting an object is present with $\text{LD} > 2$), versus 0% for the evaluated prior methods.

Lacking causality, prior methods minimally intervene on the specific query token where visual retrieval takes place, as shown in Figure 1a. Therefore, they barely change the visual evidence retrieved at these tokens. As a result, they cannot substantially change the model's confidence and thus fail to correct overconfident hallucinations (Figure 1c).

We therefore propose *TRACE* (**T**argeted **R**etrieval **A**ttention **C**ausal **E**diting), a simple training-free method that causally corrects visual retrieval at the discovered *Visual Retrieval Heads*. For a target object, an external grounding model provides a confidence score and candidate bounding boxes. *TRACE* uses this signal to steer retrieval toward the corresponding object region when visual support exists, and toward a null reference otherwise. By correcting the evidence before it propagates to downstream prediction, *TRACE* directly targets the source of overconfident hallucinations rather than suppressing their final expression.

We apply *TRACE* to both object-existence questions and image captioning. For object-existence questions, *TRACE* corrects the visual evidence for the queried object before the model answers Yes or No. For captioning, *TRACE* verifies each generated object by measuring change in object evidence after visual retrieval correction, then adjusts the generation score. We evaluate *TRACE* on two representative VLMs: LLaVA-1.5 (Liu et al., 2024) and Qwen2-VL (Wang et al., 2024) using POPE (Li et al., 2023), BEAF (Ye-Bin et al., 2024), and CHAIR (Rohrbach et al., 2018). *TRACE* reduces object hallucination in both settings and improves over existing methods by up to 11.7%, outperforming current methods. In short, our contributions are threefold:

- To the best of our knowledge, we are the first to causally trace the information flow underlying hallucinations and discover *Visual Retrieval Heads*: the stage where VLMs ground on visual evidence to causally determine the predictions.
- We find that the evaluated mitigation methods can change answers without substantially changing where the model grounds visual evidence, revealing a gap between answer correction and evidence retrieval.
- We introduce *TRACE*, a training-free method that directly corrects evidence retrieval at the discovered *Visual Retrieval Heads*. *TRACE* greatly reduces object hallucination across both object-existence questions and image captioning, outperforming existing approaches.

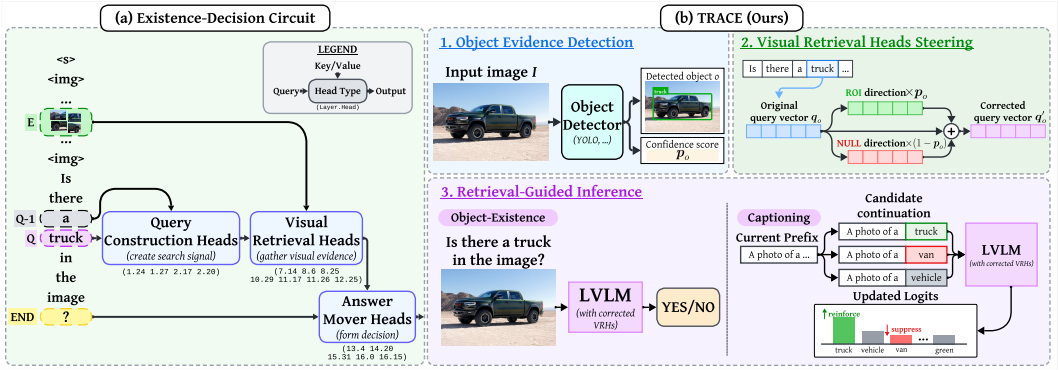


Figure 2: (a) We trace the computational stages in LLaVA-1.5 that implement object-existence decisions. *Query Construction Heads* form object-specific search signals, *Visual Retrieval Heads* use this signal to gather visual evidence from image tokens and *Answer Mover Heads* write presence signals toward the output at END. Arrows indicate query/key/value inputs and writes to the residual stream. (b) *TRACE* uses detector confidence and candidate bounding boxes to steer *Visual Retrieval Heads* toward the ROI when supported, or toward a null reference otherwise. The corrected retrieval is used directly for Yes/No prediction and to verify object candidates during captioning.

2 Related Work

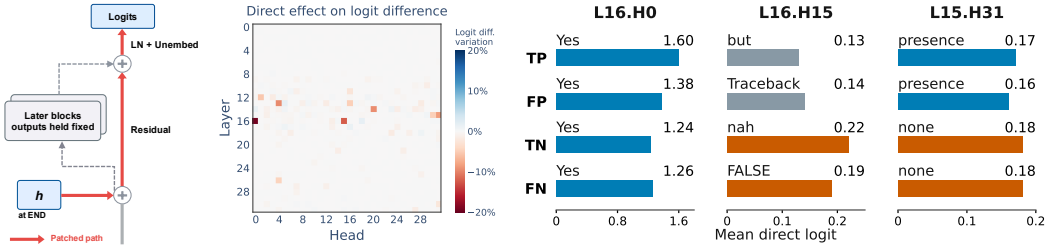
Object Hallucination in Large Vision-Language Models. Object hallucination occurs when LVLMs generate or affirm objects absent from the image (Rohrbach et al., 2018; Li et al., 2023), reducing response reliability. During decoding, VCD (Leng et al., 2024) contrasts predictions from clean and distorted visual inputs, while OPERA (Huang et al., 2024) penalizes over-trusted summary tokens. Earlier in visual processing, AGLA (An et al., 2025) combines global and local attention to improve grounding. CausalLens (Ji et al., 2026) applies sensitivity-guided causal interventions to attention heads. However, these methods do not reconstruct internal computations deeply enough to functionally explain how hallucinations arise. In contrast, we causally trace information flow underlying object hallucination to identify internal components directly driving erroneous predictions.

Mechanistic Interpretability of Vision-Language Models. Mechanistic interpretability seeks to explain learned computations through functional components and their interactions (Nanda et al., 2023; Wang et al., 2022), helping diagnose model failures and understand emergent behavior (Wei, 2022). These computations are often described as *circuits*: subgraphs of the model's computational graph whose components jointly implement a particular behavior (Wang et al., 2022). To identify and validate such circuits, causal interventions such as ablation, causal mediation, and activation/path patching test how individual components contribute to the behavior (Wang et al., 2022). For VLMs, Neo et al. (2025) use ablation and attention knockout to localize object information, while Yang et al. (2025) identify text-biased hallucination heads through causal mediation. FCCT (Li et al., 2026) traces causal effects across visual/text tokens, MHSA, FFNs, and hidden states. Beyond localizing influential components, we reconstruct a head-level *causal circuit* with three functional stages: query construction, visual evidence retrieval, and answer readout.

3 A Causal Analysis of Visual Evidence Retrieval

Task description. We study LLaVA-1.5-7B (Liu et al., 2024) on 10,002 balanced POPE-style questions (Li et al., 2023) from 1,667 COCO images (Lin et al., 2014). False-positive answers are treated as object hallucinations, and $\mathrm{LD} = \mathrm{logit}(\mathrm{Yes}) - \mathrm{logit}(\mathrm{No})$ is a proxy for confidence in object presence. In this work, we primarily trace pathways supporting correct positive answers because these cases let us follow the model's normal computation, from retrieving evidence for a present object to forming its answer. To identify this underlying computation, we use *path patching* (Wang et al., 2022): we replace a head's activation in a true-positive run with one from a length-matched true-negative run along a selected path $h \rightarrow R$. A large average LD drop under path patching reveals these pathways. Following Wang et al. (2022), we intervene only on attention heads, not MLPs, layer norms, or embedding matrices. Details and additional experiments are in Appendices A and B.

Which heads directly write to the output? (Answer Mover Heads). We begin by identifying attention heads that directly influence the output. We apply path patching to $h \rightarrow \text{logits}$ for each head at END position (Figure 3a). Only five heads (L13.H4, L14.H20, L15.H31, L16.H15, L16.H0) show a significant drop on logit difference, especially L16.H0 (Figure 3b). We call these heads *Answer Mover Heads*, i.e., heads that read object-presence signals and write answer-related information at END.



(a) (b) (c)

Figure 3: (a) We search for heads that directly affect the logits using path patching. (b) Result of the path patching experiments. *Answer Mover Heads* are the heads with the strongest direct effects on the logit difference. (c) Applying the logit lens shows that *Answer Mover Heads* write signals related to the answer and object presence, e.g., L16.H0 writes "Yes" in all groups, L16.H15 writes negation in "No" cases, and L15.H31 writes "presence"/"none" in "Yes"/"No" cases. Full table in Appendix C.1.2.

Answer Mover Heads use the object-presence signal at the queried object position to produce the answer. To better understand the function of these heads, we examine their attention patterns at the END position. We aggregate attention mass into seven categories: `image`, `pre_object`, `object`, `post_object`, `qmark`, `after_qmark`, and `other`, illustrated below with formatting tokens omitted:

$$\underbrace{\textit{BOS / template}}_{\texttt{other}} + \underbrace{v_1 \cdots v_N}_{\texttt{image}} + \underbrace{\textit{Is there a}}_{\texttt{pre_object}} \underbrace{\textit{cat}}_{\texttt{object}} \underbrace{\textit{in the image}}_{\texttt{post_object}} \underbrace{\textit{?}}_{\texttt{qmark}} + \underbrace{\textit{ASSISTANT:}}_{\texttt{after_qmark}}$$ [1]

We find that these heads allocate most of their attention to text, with very little directed toward image tokens (Figure 12 in Appendix C.1.1). L16.H0 attends most strongly to "?"[2], whereas L13.H4, L14.H20, and L16.H15 attend more to the object-text token.

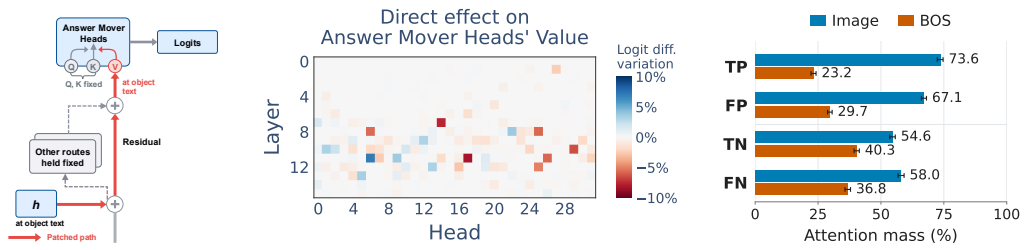
We then use the Logit Lens (nostalgebraist, 2020) to examine these heads' outputs in vocabulary space (Figure 3c). We observe a pattern: these heads write answer-related signals into the residual stream at the END position. L16.H0 writes a "Yes" signal that is strongest in true-positive cases but weakens as the model's prediction shifts toward "No"; L16.H15 writes negation-related signals but only for "No" answers; and L15.H31 writes distinct signals for "Yes" and "No" answers. In contrast, L13.H4 and L14.H20 show no clear answer semantics, suggesting that they provide supporting signals that are not yet directly interpretable as an object-existence decision at these layers.

These results lead us to hypothesise that, at this stage, the object-text representation encodes a signal of the queried object's presence or absence. *Answer Mover Heads* read this signal and translate it into output logits with varying strengths, reflecting the model's confidence. An ablation study designed to test this hypothesis is described in Appendix C.1.4.

Which heads write the object-existence signal? (Visual Retrieval Heads). Given that the *Answer Mover Heads* read the object-existence signal at the queried-object position, we next investigate which heads provide this signal. Our goal is to trace how evidence forms the decision, so we focus on the value read by the *Answer Mover Heads*. Therefore, we patch $h \rightarrow V_{\text{answer mover heads}}$ at the object-text position (Figure 4a), and again report the change in the logit difference. Patching L11.H17, L7.H14, L10.H29, L8.H25, L11.H26, L8.H6, and L12.H25 directly affects the values read by *Answer Mover Heads*, causing a significant drop in logit difference (Figure 4b). We call these heads *Visual Retrieval Heads*.

[1] `ASSISTANT:` is part of LLaVA's conversation template.

[2] We also investigate what information is stored at "?" and find object-text-related contributions from L13.H4, L13.H28, and L14.H20. See Appendix C.1.3 for detailed analysis.



(a) (b) (c)

Figure 4: (a) Diagram of direct-effect experiments on *Answer Mover Heads'* value vectors at object text position. We patch activations from true-negative examples along the red path and measure the resulting change in the final logit difference. (b) Result of the path patching experiments. *Visual Retrieval Heads* are the heads that cause a decrease in logit difference. (c) Mean image and BOS attention at the object token across *Visual Retrieval Heads*; error bars indicate 95% confidence intervals. "No" predictions show lower image attention and higher BOS attention than "Yes" predictions.

Visual Retrieval Heads attend more to visual object regions when predicting presence, and more to BOS when predicting absence. We first examine these heads' attention at the object-text position. At this stage, the model pays more attention to image tokens than to text tokens through these heads (Figure 4c; full statistics in Appendix C.2.1). When the queried object is present and correctly recognized, these heads strongly focus on visual tokens, with concentrated and consistent patterns across heads (73.6% image attention; consensus 0.742). Attention drops slightly when the model hallucinates an object, but remains substantially higher than when it predicts "No." In "No" cases, attention becomes less image-focused and shifts more toward BOS.

We then use COCO bounding-box annotations and measure *region-of-interest* (ROI) mass: the fraction of image attention falling within the annotated object region. The regional breakdown is provided in Figure 15 in Appendix C.2.1. Across all true-positive cases with a visible, annotated target, the target region receives an average of 54.8% of image attention. Figure 5a illustrates attention concentrated on the queried object.

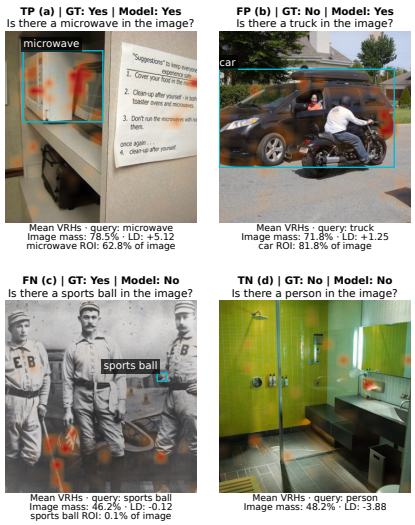


Figure 5: Attention patterns of *Visual Retrieval Heads*.

Visual Retrieval Heads' attention causally determines the model's answer. We test two direct attention interventions at the object text token. First, we shift attention away from correct object evidence in true-positive cases, redirecting it to <s>[3] or another object's ROI. Second, we redirect misallocated attention, including that on <s>, back to the target ROI in false-negative cases, and also to another object's ROI to distinguish correct evidence retrieval from simply increasing image attention. Table 1 reports the resulting logit-difference changes. Controls using random and top image-attending heads, and interventions at the question-mark and END positions, are provided in Appendix C.2.2.

In true cases, redirecting image attention to <s> produces the largest drop in mean true-positive logit difference (4.554 to -0.214), showing that cutting off these heads' image attention weakens the model's support for the object's presence. Steering attention to another object lowers LD to 2.059 without reversing its sign. Explanation for this can be that the target object's information may also be distributed across other visual tokens, potentially through self-attention. For missed objects, steering image and BOS attention to the target raises mean LD from -0.851 to 1.644, whereas steering to another object yields only 0.153. This suggests that attending to the correct region strengthens evidence for hard-to-detect objects more than just increasing image attention.

[3]The model allocates more attention to <s> when answering No; we therefore use it as an operational NULL destination to represent a lack of retrieved visual evidence.

When the model hallucinates, Visual Retrieval Heads attend to semantically similar but incorrect image regions. In hallucination (false-positive) cases, although the queried object is absent, the model still strongly attends to another object, with the most-attended COCO object receiving 53.5% of image attention on average. Notably, the attended objects often share the queried object's COCO supercategory (when present, they receive peak area-corrected attention in 73% of false-positive cases). For example, the model may attend to a car when asked about a truck (Figure 5b), a couch when asked about a chair, or a handbag when asked about a backpack.

We call these similar objects *semantic confusers*. Table 2 reports pair-specific false-positive rates ($\mathrm{FP\ rate} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN}) \times 100\%$) and mean semantic-confuser ROI mass in false-positive cases. These observations explain that when hallucinating, the model uses information from a similar concept as evidence of the queried object's presence.

Table 1: Mean Yes-No logit difference (LD) after attention redirection in the *Visual Retrieval Heads* (VRHs).

Setting	Group	LD
Natural	TP	4.554
VRHs → <s>	TP	−0.214
VRHs → other-object ROI	TP	2.059
Natural	FN	−0.851
VRHs → target ROI	FN	1.644
VRHs → other-object ROI	FN	0.153

Table 2: Top five semantic-confuser pairs by False Positive rate. Full results and definitions: Appendix C.2.3.

Queried object	Semantic confuser	FP rate (%)	Confuser ROI mass (%)
chair	couch	75.0	46.9
bicycle	motorcycle	75.0	54.7
truck	car	74.4	42.4
backpack	handbag	68.2	27.2
cup	bowl	66.7	30.1

When the model misses an object, Visual Retrieval Heads attend weakly to the target regions. 40.9% of false-negative cases have no annotated target region after CLIP center-crop preprocessing. Among the remaining false-negative cases, the target object occupies only 3.7% of the image on average, compared with 10.4% in true-positive cases. These target regions receive an average of 22.7% of image attention, compared with 54.8% in the true-positive cases (Figure 5c). This suggests that the visual signal itself is too weak to attract these heads to the target and retrieve the relevant evidence.

In conclusion, at this stage, the model uses *Visual Retrieval Heads* to gather evidence of the queried object's existence. When an object is present and visually clear, these heads concentrate attention on the correct region and retrieve evidence that supports its presence. Conversely, when the object is absent, attention shifts toward <s>, while the remaining image attention provides insufficient evidence, reinforcing the object's absence.

Hallucination arises when the model misgrounds evidence from a semantically related region as evidence of the queried object's presence. We further investigate the hallucination mechanism underlying this behavior and discover *Query Construction Heads*. Details are provided in Appendix C.3.

4 Causal-Guided Hallucination Mitigation

The findings in Section 3 lead to a question: Can hallucination be reduced by guiding *Visual Retrieval Heads* (*VRHs*) toward the correct object region when the object is present, and away from misleading evidence when it is absent? Motivated by this, we introduce *TRACE* (**T**argeted **R**etrieval **A**ttention **C**ausal **E**diting), with two task-specific variants for object-existence and free-form captioning (Figure 2), both built on the same visual retrieval correction. For object-existence, the target object is provided in the question and the correction directly affects the answer; for captioning, it is used to verify generated object candidates.

4.1 Locating Visual Retrieval Heads

We identify *VRHs* using a small held-out discovery set. For each task, we keep examples where the model is correct on the real image and replacing it with a gray null image sufficiently reduces the target score. For object-existence tasks, the score is the Yes-No logit difference; for captioning, it is the logit of the object-token right before it is generated. For each attention head h, we bidirectionally patch its image-token K/V states between the real and null runs. Let b_r and b_n be the corresponding

target scores and $\Delta = b_r - b_n$. We define

$$N_h = \frac{b_r - b_{r\leftarrow n}^{(h)}}{\Delta}, \qquad R_h = \frac{b_{n\leftarrow r}^{(h)} - b_n}{\Delta}, \qquad S_h = \frac{N_h + R_h}{2}, \tag{1}$$

where N_h measures how much removing the head's visual information weakens the prediction, while R_h measures how much restoring it recovers the prediction. We aggregate S_h across examples and retain the highest-scoring heads with positive effects on at least half of the discovery examples as ***VRHs***. This selects heads whose visual states are both necessary and restorative, and directly affects the model's prediction, rather than heads that merely have high image attention.

4.2 Visual Retrieval Correction

Let o be the object being assessed. Given image I, an external detector provides a confidence s_o and, when available, a bounding box R_o. We convert s_o into a calibrated presence probability $p_o = g(s_o) \in [0, 1]$, where g is fitted on an independent calibration set.

For each visual retrieval head $(l, h) \in$ ***VRHs***, let $a_j^{l,h}$ denote its attention from the object token to token j. For any target token set $\mathcal{S}$, define

$$A_{\mathcal{S}}^{l,h} = \sum_{j\in\mathcal{S}} a_j^{l,h}, \quad \nabla_{q_o^{l,h}} \log A_{\mathcal{S}}^{l,h} = \frac{k_{\mathcal{S}}^{l,h} - \bar{k}^{l,h}}{\sqrt{d}}, \quad k_{\mathcal{S}}^{l,h} = \frac{\sum_{j\in\mathcal{S}} a_j^{l,h} k_j^{l,h}}{A_{\mathcal{S}}^{l,h}}, \quad \bar{k}^{l,h} = \sum_j a_j^{l,h} k_j^{l,h}.$$

Thus, $k_{\mathcal{S}}^{l,h} - \bar{k}^{l,h}$ gives a local direction that moves retrieval toward $\mathcal{S}$. We use two such directions:

$$d_{\mathrm{ROI}}^{l,h} = \mathrm{Norm}\left(k_{\mathrm{ROI}}^{l,h} - \bar{k}^{l,h}\right), \qquad d_{\mathrm{null}}^{l,h} = \mathrm{Norm}\left(k_{\mathrm{BOS}}^{l,h} - \bar{k}^{l,h}\right),$$

where the ROI direction points toward the detected object region and the null direction toward BOS, which serves as a no-evidence reference. Finally, we update the object query as

$$q_o'^{l,h} = q_o^{l,h} + \|q_o^{l,h}\|_2 \left[\alpha p_o d_{\mathrm{ROI}}^{l,h} + \beta(1 - p_o) d_{\mathrm{null}}^{l,h}\right]. \tag{2}$$

A high p_o therefore favors retrieval from the detected region, while a low p_o shifts retrieval toward the null reference. [4]

Discriminative Object Existence. For object-existence questions, the queried object o is given explicitly. We therefore apply Eq. 2 directly to ***VRHs*** before the model produces its Yes/No answer. High detector confidence steers retrieval toward the detected ROI, while low confidence steers it toward BOS. The remaining layers are left unchanged.

Retrieval-Based Verification for Captioning. In captioning, the object is not known beforehand. We first let the model propose an object candidate o, then temporarily append it to the current prefix as the queried object and apply Eq. 2. Let $L_R = \max\{l : (l, h) \in$ ***VRHs***$\}$ be the last layer containing a visual retrieval head. We score the change in object evidence as

$$V_R(o) = w_o^\top \left[N(h_{L_R,o}^{\mathrm{corr}}) - N(h_{L_R,o}^{\mathrm{orig}})\right], \tag{3}$$

where $N(\cdot)$ is the final normalization layer and w_o is the LM-head vector of o. We then update $z'(o) = z(o) + \gamma V_R(o)$, so visually supported objects are reinforced and unsupported ones are suppressed. At most K object candidates are verified at each decoding step.

5 Experiments

5.1 Experimental Setup

Models. We evaluate TRACE on two open-source LVLMs: LLaVA-1.5-7B (Liu et al., 2024) and Qwen2-VL-7B (Wang et al., 2024).

Benchmarks and Metrics. We evaluate object hallucination on three benchmarks. **POPE** (Li et al., 2023) uses binary object-existence questions; we report Accuracy and F1 on the random, popular,

[4] If no ROI is returned for o, we set $p_o = 0$; the ROI steering term $\alpha p_o d_{\mathrm{ROI}}^{l,h}$ is therefore omitted.

Table 3: Results on POPE-COCO Benchmark. Best and second-best results are highlighted in green and gray, respectively.

Method	LLaVA-1.5-7B						Qwen2-VL-7B					
	Random		Popular		Adversarial		Random		Popular		Adversarial	
	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
Greedy	88.4	88.5	84.8	85.5	78.1	80.3	89.0	87.8	87.7	86.6	86.5	85.5
VCD	87.8	87.7	84.3	84.6	79.0	80.5	88.7	87.5	87.0	85.9	85.8	84.8
OPERA	89.8	89.8	86.2	86.3	80.4	81.6	88.8	88.0	87.8	86.7	86.5	85.4
AGLA	87.2	86.5	84.2	83.5	80.9	81.1	86.4	84.3	85.3	83.3	82.8	81.3
ONLY	88.9	88.8	85.7	86.1	79.3	81.0	81.5	77.4	81.2	77.1	81.1	77.0
CausalLens	88.3	88.7	84.1	85.2	76.9	79.9	89.7	88.7	88.5	87.5	86.8	85.9
SHIELD	90.3	89.8	85.2	84.1	82.5	81.9	**91.4**	**90.9**	88.9	88.5	80.9	79.0
TRACE	**90.7**	**89.9**	**89.7**	**88.9**	**88.2**	**87.6**	**91.4**	90.8	**89.6**	**89.0**	**87.8**	**87.3**

and adversarial splits. **BEAF** (Ye-Bin et al., 2024) evaluates whether predictions respond appropriately to object removal. We report True Understanding (TU), InDecision (ID), and $\mathrm{F1}_{\mathrm{TUID}}$. **CHAIR** (Rohrbach et al., 2018) evaluates object hallucination in free-form captioning. We report CHAIR_i, the fraction of hallucinated object mentions, and CHAIR_s, the fraction of captions containing hallucinated objects. Detailed metric definitions are provided in Appendix E.

External Visual Model. We use YOLO26 (Jocher et al., 2026) as the external object detector. For each queried object, it provides a confidence score and, when available, a bounding box. We map the raw confidence to a presence probability using a logistic calibrator fitted on an independent COCO calibration set disjoint from the evaluation images. The detector and calibrator remain fixed during evaluation.

Baselines and Implementation Details. We compare TRACE with representative training-free hallucination mitigation methods, including VCD (Leng et al., 2024), OPERA (Huang et al., 2024), AGLA (An et al., 2025), ONLY (Wan et al., 2025), and CausalLens (Ji et al., 2026), SHIELD (Huang et al., 2026). The intervention strengths α and β are selected separately for each LVLM using held-out development data and are then fixed across benchmarks. Additional model configurations, hyperparameter settings for evaluated methods, and decoding settings are provided in Appendix D.

5.2 Main Results

Results on Object-Existence Hallucination Evaluation. Table 3 reports results on POPE-COCO under the random, popular, and adversarial settings. TRACE consistently improves over greedy decoding on all three splits and both LVLMs. On LLaVA-1.5-7B, the largest gain appears on the adversarial split, where Accuracy/F1 increases from 78.1/80.3 to 88.2/87.6. TRACE also reaches 90.7/89.9 and 89.7/88.9 on the random and popular splits. On Qwen2-VL-7B, TRACE obtains 91.4/90.8, 89.6/89.0, and 87.8/87.3 on the three splits, respectively. It achieves the best Accuracy/F1 on both the popular and adversarial splits, while tying the best Accuracy on the random split. Overall, the gains are largest on the harder adversarial setting, showing that better visual retrieval helps when negative objects are more difficult to reject.

Table 4: Results on BEAF and CHAIR. Best and second-best results are highlighted in green and gray, respectively.

Method	BEAF						CHAIR			
	LLaVA-1.5-7B			Qwen2-VL-7B			LLaVA-1.5-7B		Qwen2-VL-7B	
	TU↑	ID↓	$\mathrm{F1}_{\mathrm{TUID}}$ ↑	TU↑	ID↓	$\mathrm{F1}_{\mathrm{TUID}}$ ↑	C_s ↓	C_i ↓	C_s ↓	C_i ↓
Greedy	30.09	5.46	45.65	57.08	3.89	71.62	50.6	14.6	25.6	6.6
VCD	34.70	8.90	50.26	57.22	5.17	71.38	55.2	16.7	24.0	7.3
OPERA	34.51	6.04	50.48	53.71	4.22	68.83	48.6	14.6	22.2	5.7
AGLA	56.70	5.10	70.99	60.23	4.52	73.87	46.4	13.6	21.2	5.8
ONLY	33.80	5.32	49.82	61.19	2.98	75.05	53.0	15.9	21.4	5.7
CausalLens	52.59	3.95	67.96	52.19	4.55	67.48	45.2	13.3	22.7	5.8
SHIELD	41.90	5.73	58.02	49.49	4.72	65.15	36.6	10.3	21.6	5.8
TRACE	**67.00**	**2.84**	**79.30**	**71.82**	**2.18**	**82.83**	**35.4**	**10.0**	**20.2**	**5.3**

Results on Change-Aware Hallucination Evaluation. Table 4 reports results on BEAF. TRACE achieves the best results on all three metrics for both LVLMs. On LLaVA-1.5-7B, TU increases from 30.09 to 67.00 and ID decreases from 5.46 to 2.84, giving an $F1_{TUID}$ of 79.30. On Qwen2-VL-7B, TRACE reaches 71.82 TU, 2.18 ID, and 82.83 $F1_{TUID}$, compared with 57.08, 3.89, and 71.62 under greedy decoding. These results show that TRACE responds better to relevant visual changes while remaining stable under irrelevant ones. Results for the remaining BEAF indicators are reported in Appendix F.1.

Results on Open-Ended Caption Hallucination Evaluation. Table 4 also reports CHAIR results for free-form image captioning. On LLaVA-1.5-7B, TRACE reduces $CHAIR_s$ from 50.6 to 35.4 and $CHAIR_i$ from 14.6 to 10.0, achieving the lowest values among all compared methods. On Qwen2-VL-7B, TRACE reduces $CHAIR_s$ from 25.6 to 20.2 and $CHAIR_i$ from 6.6 to 5.3, again giving the best results. This shows that TRACE reduces both the fraction of captions containing hallucinated objects and the fraction of generated object mentions that are hallucinated. Full captioning results, including caption length, object recall, and inference-time comparisons, are provided in Appendices F.2 and F.3.

5.3 Ablation Studies

We ablate the intervention position, head selection, and number of intervened heads (Figure 6). Ablations are conducted on subsets of the standard POPE and BEAF benchmarks. Ablations of the retrieval-correction components and detector-confidence calibration are provided in Appendix F.4.

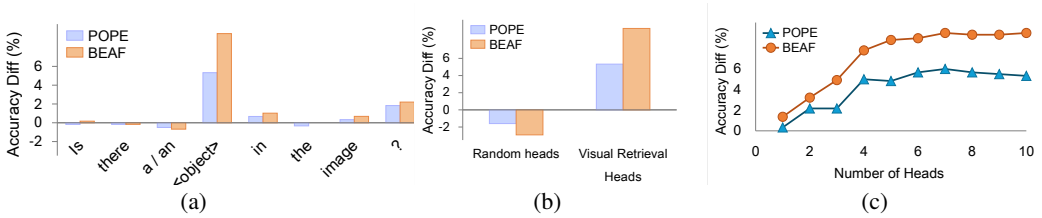


Figure 6: Accuracy changes relative to greedy decoding on subsets of the POPE and BEAF benchmarks. (a) Intervening at the queried-object position yields the largest gain on both sets. (b) Targeting *Visual Retrieval Heads* improves accuracy, whereas intervening on random heads reduces it below the baseline. (c) Gains increase as more retrieval heads are added and largely saturate after six to seven heads, indicating that effective correction is concentrated in a small set of targeted heads.

Intervention location. We first vary the token position at which retrieval correction is applied. As shown in Figure 6a, the queried object position gives the largest gain on both benchmarks, while surrounding tokens yield substantially smaller changes. This supports our causal analysis that the object position is the main site at which *VRHs* retrieve object-specific visual evidence. On the other hand, replacing *VRHs* with randomly selected heads drops performance below the greedy baseline on both benchmarks (Figure 6b), showing that the effect does not arise from modifying arbitrary attention heads.

Number of retrieval heads. We next vary the number of visual retrieval heads being intervened on. Figure 6c shows a clear improvement as the first few heads are added. Adding more heads produces diminishing gains, with performance largely stabilizing after six to seven heads on both benchmarks. Thus, the retrieval correction is concentrated in a small set of heads.

6 Conclusion

This work shows how causal understanding can guide targeted and efficient hallucination mitigation. By uncovering the computations behind erroneous predictions, we identify where and how to intervene, achieving substantial improvements by correcting a small set of causally relevant components. These findings highlight the value of *causal analysis* for improving model reliability and guiding future architecture design. Future work will extend this investigation to reasoning, open-ended generation, and interactions between vision encoders and language decoders.

AI USE STATEMENT

We used generative AI tools to assist with translation, manuscript drafting and editing, LaTeX formatting, figure preparation and plotting code, and discussion of hypotheses and experimental interpretations. The authors take responsibility for the final content of this work, including all AI-assisted text, claims, code, and figures.

REPRODUCIBILITY STATEMENT

Our source code is included in the supplementary materials. Section 5.1 describes the evaluation benchmarks. Appendix A details dataset construction and path-patching procedures, while Appendix D provides model configurations, hyperparameter settings, and evaluation prompts.

REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926. IEEE, 2025.

ARENA. Path patching. Chapter 1: Transformer Interpretability, Indirect Object Identification, n.d. URL `https://learn.arena.education/chapter1_transformer_interp/21_ioi/4-path-patching/`. Accessed September 22, 2026.

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

Yiyang Huang, Liang Shi, Yitian Zhang, Yi Xu, and Yun Fu. Shield: Suppressing hallucinations in lvlm encoders via bias and vulnerability defense. In *International Conference on Learning Representations*, volume 2026, pp. 125123–125143, 2026.

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*, pp. 4199–4209, 2026.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026.

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

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

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

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

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2026-09-23.

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.

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

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. *arXiv preprint arXiv:2211.00593*, 2022.

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

J Wei. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Yutong Xie, Zhenglin Hua, Ran Wang, Wing WY Ng, Xizhao Wang, and Yuheng Jia. Finding the correct visual evidence without forgetting: Mitigating hallucination in lvlms via inter-layer visual attention discrepancy. *arXiv preprint arXiv:2605.20965*, 2026.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *International Conference on Learning Representations*, volume 2025, pp. 51546–51568, 2025.

Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision*, pp. 232–248. Springer, 2024.

Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. Prefill-time intervention for mitigating hallucination in large vision-language models. *arXiv preprint arXiv:2604.25642*, 2026.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *International Conference on Learning Representations*, volume 2025, pp. 54415–54439, 2025.

A PRELIMINARY

A.1 TRANSFORMER ARCHITECTURE

LLaVA-1.5-7B connects a CLIP vision encoder to a Vicuna language decoder through a two-layer MLP projector (Liu et al., 2024). The projector maps 576 visual representations into the decoder's embedding space; these visual tokens precede the textual question. The decoder has 32 layers, each containing 32 attention heads and an MLP that read from and write to the residual stream. For head (ℓ, h) with dimension d_h, the attention pattern and output are

$$A^{\ell,h} = \mathrm{softmax}\left(\frac{Q^{\ell,h}(K^{\ell,h})^\top}{\sqrt{d_h}}\right), \qquad H^{\ell,h} = A^{\ell,h} V^{\ell,h} W_O^{\ell,h}, \tag{4}$$

with causal masking implicit. Thus, $QK^\top$ determines which image or text positions the head reads from, whereas V and W_O determine what it writes to the residual stream. A final normalization and vocabulary projection convert the last residual state into next-token logits.

A.2 DATASET CONSTRUCTION

We select 1,667 COCO 2014 training images with at least three annotated object categories (Lin et al., 2014). Each image yields three present-object and three absent-object questions, totaling 10,002 balanced questions. Absent objects are selected from frequently co-occurring categories. We use the POPE-style template "Is there a/an ⟨object⟩ in the image?" (Li et al., 2023). Predictions are grouped as true positives (present/Yes), true negatives (absent/No), false positives (absent/Yes), and false negatives (present/No). We treat false positives as object hallucinations.

LLaVA-1.5-7B achieves 79.82% accuracy on this set, with 4,654 true positives, 3,330 true negatives, 1,671 false positives, and 347 false negatives.

A.3 PATH-PATCHING IMPLEMENTATION DETAILS

Choice of replacement examples. Our goal is to remove the signal supporting object presence along the patched path while preserving the object-existence question format. We therefore pair a true-positive example with a true-negative example having the same tokenized sequence length and question template, but a different image and queried object. The true-positive run serves as the *clean* baseline, while the true-negative run supplies replacement activations from an input where the object is absent and the model correctly answers No.

Path patching isolates the contribution of a head h to selected downstream components R, rather than allowing its replacement to affect every route to the output (Wang et al., 2022). Figure 7 gives a conceptual overview.

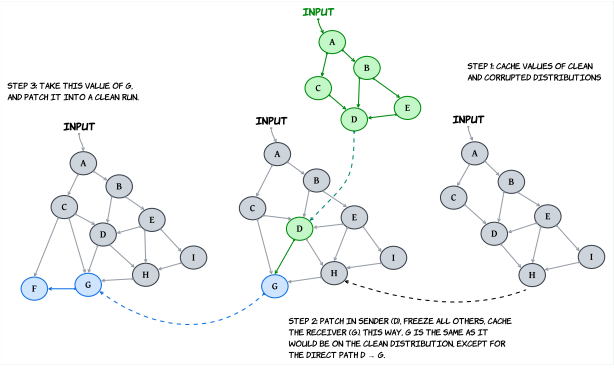


Figure 7: Conceptual overview of path patching, reproduced from the ARENA path-patching tutorial (ARENA, n.d.). Gray denotes clean activations, green denotes replacement activations, and blue denotes recomputed receiver activations and their downstream effects. In our setting, the clean and replacement inputs are true-positive and true-negative examples, respectively.

Implementation. Following Wang et al. (2022), we retain paths through the residual stream and MLPs while blocking routes through intermediate attention heads. Here, R is a set of selected head inputs (Q, K, or V), or the final residual state. Figure 8 illustrates the procedure in a Transformer.

1. **Cache both runs.** Record head outputs before W_O and the baseline logit differences for the true-positive and true-negative inputs. Retain pairs with positive and negative LD, respectively, and align the intervention sites using the question template.
2. **Isolate the path.** Rerun the true-positive input, replacing only h at the selected positions with its true-negative activation. Freeze all other head outputs to their clean values, but recompute MLPs and normalization. Save the resulting inputs to the selected receivers strictly downstream of h before overwriting their head outputs.
3. **Replay the receivers.** In a fresh true-positive run, insert the saved receiver inputs jointly at the selected positions, and let the remaining computation proceed normally. Measure the final Yes-No logit difference. For $h \rightarrow \text{logits}$, read the logits directly from the final residual state obtained in step 2.

The intervention sites are END for direct output effects, the question mark for question-mark readout, and the object token in object-site scans. Only the selected receiver component is replaced; the other components are recomputed normally during replay. We average the per-pair normalized effect

$$e = \frac{\text{LD}_{\text{patched}} - \text{LD}_{\text{clean}}}{\text{LD}_{\text{clean}} - \text{LD}_{\text{replacement}}} \tag{5}$$

over valid pairs for each head. Negative values indicate reduced support for object presence under this contrast. Additional contrasts are reported in Appendix B.1.

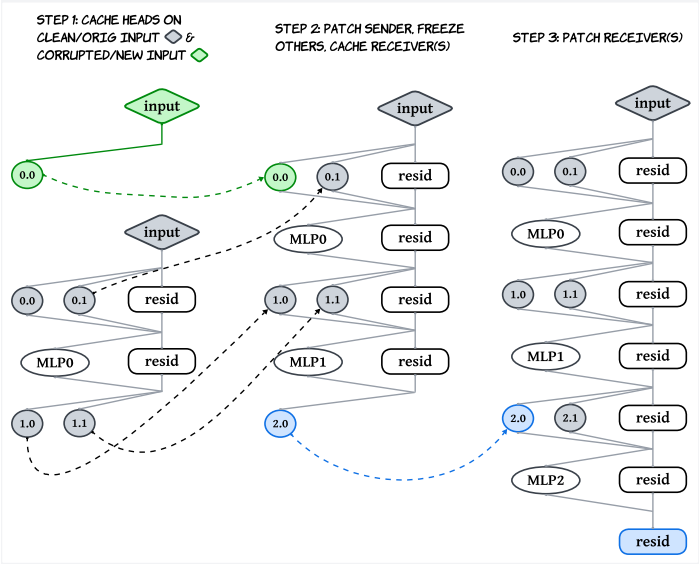


Figure 8: Transformer path-patching procedure, reproduced from ARENA (ARENA, n.d.). Cache the original and replacement runs, recompute the receiver with the selected head replaced and other heads frozen, then replay the receiver in the original run. Unfilled MLP nodes are recomputed. The head numbers are illustrative, not the heads identified in our experiments.

B More Patching Experiments

B.1 Alternative patching contrasts in LLaVA

Figure 9 compares TN-to-TP, TN-to-FP, and TP-to-TN patching. The same five *Answer Mover Heads* have the largest direct effects in all three contrasts, while *Visual Retrieval Heads* show partially overlapping effects and vary in strength. This supports a shared answer-readout stage, with more variation in the upstream retrieval pathways.

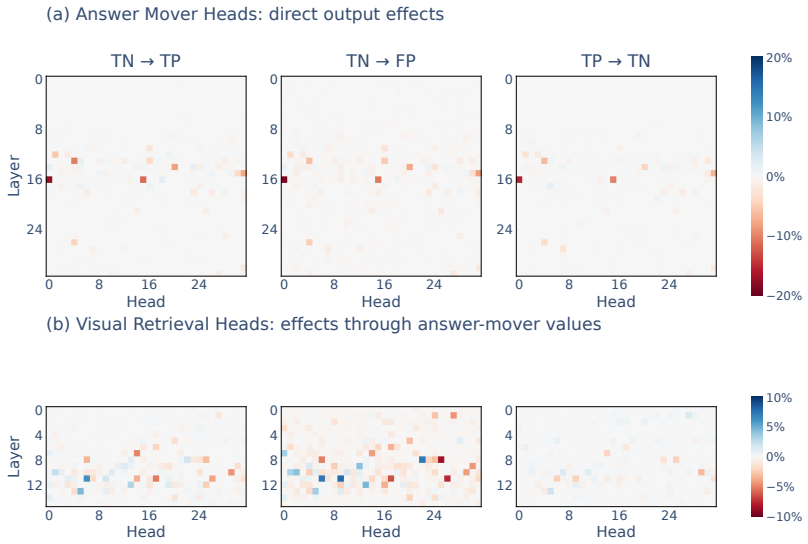
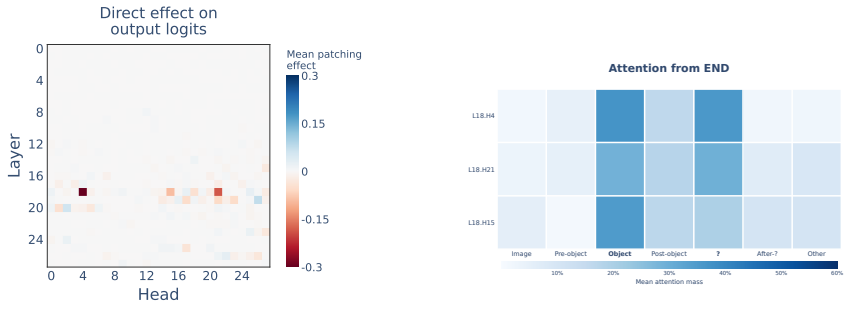


Figure 9: Causal effects across patching contrasts. (a) Direct effects on the output. *Answer Mover Heads* are the heads causing the largest direct effects. (b) Effects through *Answer Mover Heads'* value vectors. *Visual Retrieval Heads* are observed with partially overlapping heads across contrasts.

B.2 Qwen2-VL Patching Experiments

In Qwen2-VL-7B (Wang et al., 2024), we patch $h \rightarrow \text{logits}$ at END and identify a compact set of *Answer Mover Heads*: L18.H4, L18.H21, and L18.H15 produce the largest logit drops (Figure 10a). These prominent heads are fewer and more localized than in LLaVA, all lying in the same layer. We hypothesize that grouped-query attention (GQA) may contribute to this compact organization, but we have not verified this architectural explanation. Their attention from END focuses on the object text and question mark (Figure 10b).



(a) (b)

Figure 10: *Answer Mover Heads* in Qwen2-VL. (a) Result in direct patching effects at END. (b) Attention from END in *Answer Mover Heads* (L18.H4, L18.H21, and L18.H15).

We then patch $h \rightarrow V_{\text{answer mover heads}}$ at the object and question-mark tokens, identifying *Visual Retrieval Heads*, including L14.H8, L14.H11, L14.H13, and L15.H1 (Figure 11a). This scan also

shows early effects at L0.H3 and L0.H10. We further patch $h \rightarrow Q_{\text{VRHs}}$ at the last object token, supporting their role as *Query Construction Heads*, alongside L0.H21 (Figure 11b). Attention visualizations show that the *Visual Retrieval Heads* attend to the queried objects (Figure 11c).

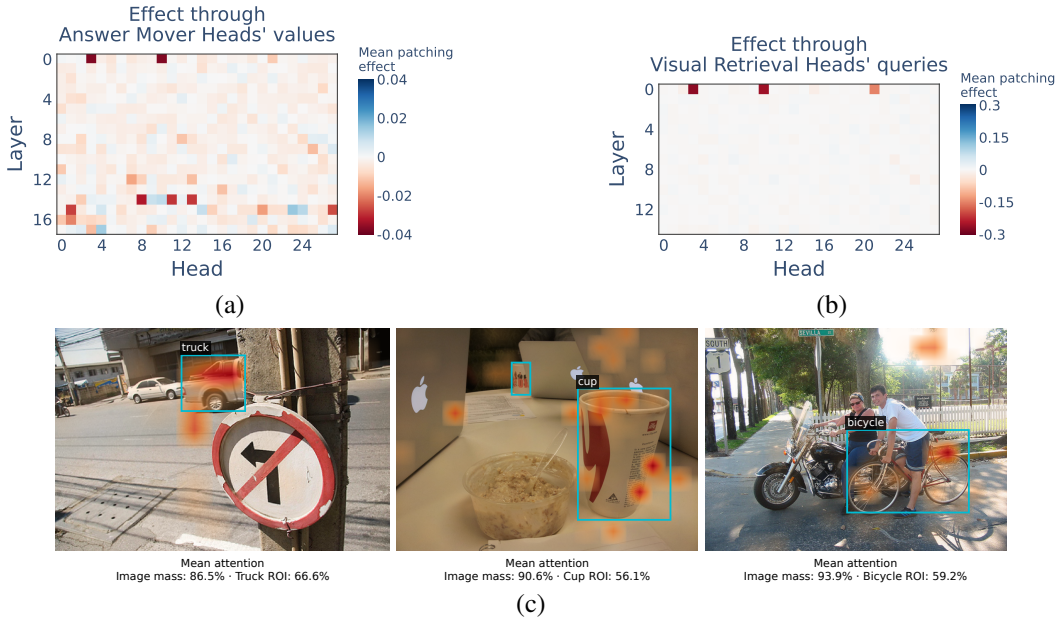


Figure 11: *Visual Retrieval Heads* and *Query Construction Heads* in Qwen2-VL. (a) Result in path-patching effects through the values of *Answer Mover Heads* at the object and question-mark tokens. (b) Effects through the queries of *Visual Retrieval Heads* (L14.H8, L14.H11, L14.H13, L15.H1) at the last object token. (c) Attention examples for a truck, a cup, and a bicycle. Heatmaps use stage-specific color scales and omit unmeasured trailing layers.

Together, these results support a shared mechanism for constructing object-specific queries, retrieving visual support, and reading out an existence decision across Vision-Language Models.

C Further Results on the Causal Analysis

This appendix provides supplementary results for the causal analysis in Section 3, followed by an investigation of query construction.

C.1 Answer-Mover Stage

C.1.1 Attention Patterns of the Answer Mover Heads

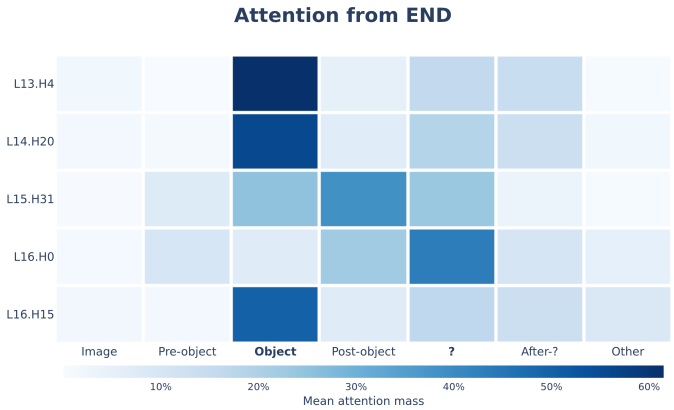


Figure 12: Mean attention mass from END to the seven token-position categories for *Answer Mover Heads*.

C.1.2 Full Vocabulary Decoding of the Answer Mover Heads

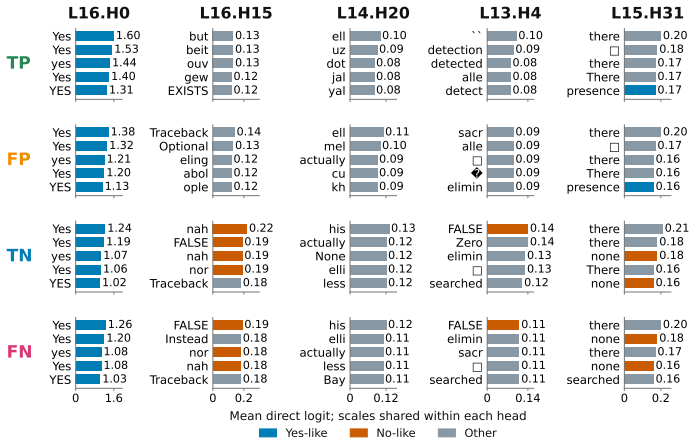
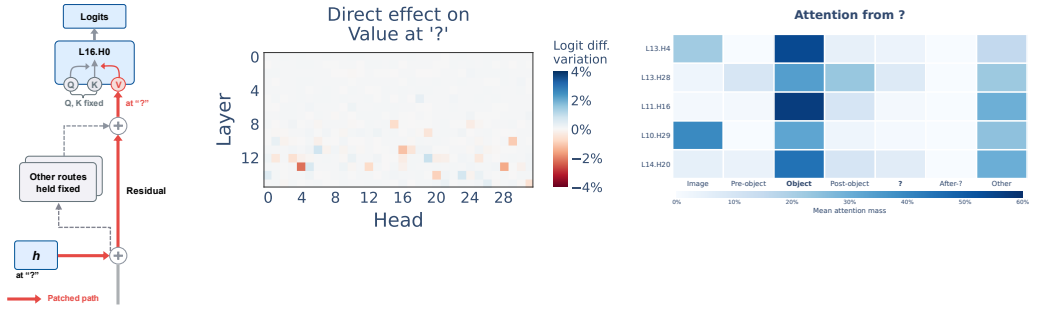


Figure 13: Mean direct logits for the top-five tokens of each head across TP, FP, TN, and FN cases on LLaVA-1.5-7B. These colors group related expressions, not exact synonyms. Axis scales are shared across outcomes within each head, but differ between heads.

C.1.3 Tracing Information Through the Question Mark

Since patching L16.H0's direct path to the logits produces the largest drop in the logit difference, and this head attends most strongly to "?", we trace which heads write the information it reads at this position. Specifically, we patch $h \to V_{\text{L16.H0}}$ at this position and measure the change in the logit difference (Figure 14).

The five strongest negative effects come from L13.H4, L13.H28, L11.H16, L10.H29, and L14.H20. Two of these heads (L13.H4 and L14.H20) also contribute to the direct readout at END. Together with L13.H28, they primarily attend to the object text, suggesting that object-related information reaches the output both directly and through "?". L10.H29 attends to image tokens and also participates in the visual-retrieval stage examined in Section 3. This backtrace suggests that "?" serves as an intermediate site combining object-text information and visual evidence for the final readout.



(a) (b) (c)

Figure 14: (a) Diagram of the direct-effect experiments on the value vector of head 16.0 at the question mark position. (b) Result in path patching experiments. Negative values indicate reduced support for Yes. Only measured sender layers (0-15) are shown. (c) Mean attention mass from "?" to token-position categories for the five heads with the strongest negative patching effects.

C.1.4 Testing the Decision Signal at the Object-Text Position

To test whether the object-text representation already carries an object-existence decision or merely encodes object identity, we keep queries and keys fixed and replace only the value vectors read by *Answer Mover Heads* from the object-text positions. We test four settings: replacing values in Yes-answer examples with those from No-answer or other Yes-answer examples, either asking about the same object or a different object. The results are reported in Table 5.

Table 5: Mean logit difference between Yes and No, when patching value vectors in *Answer Mover Heads*.

Setting	LD [95% CI]
Natural	3.252 [2.915, 3.598]
No → Yes, same object	1.351 [1.137, 1.567]
No → Yes, different object	1.343 [1.132, 1.557]
Yes → Yes, same object	3.163 [2.874, 3.469]
Yes → Yes, different object	3.196 [2.938, 3.466]

Values from No-answer examples substantially reduce the logit difference between Yes and No regardless of object identity, whereas values from other Yes-answer examples leave it largely unchanged. This supports the hypothesis that the object-text representation already carries a decision-related signal about object existence, which the *Answer Mover Heads* read and translate into the final prediction.

C.2 Visual Retrieval Stage

C.2.1 Image Attention

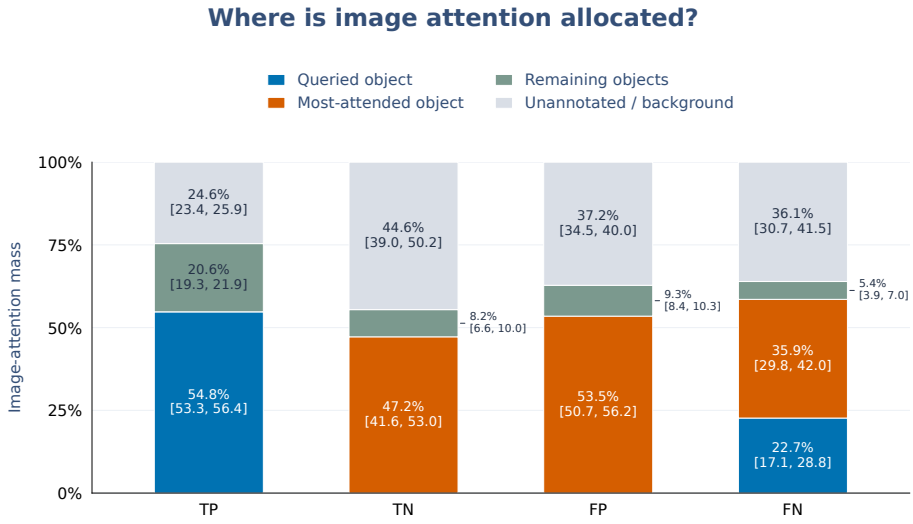


Figure 15: Fraction of image attention from the object token allocated to the queried object, the most-attended object, remaining annotated objects, and unannotated or background regions across TP, TN, FP, and FN cases. Confidence intervals (CIs) are shown.

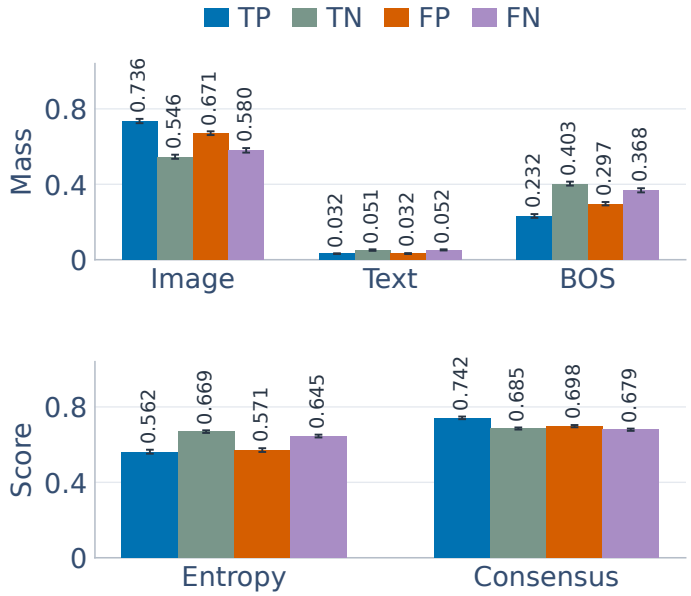


Figure 16: Full attention statistics at the object token for *Visual Retrieval Heads*: image, text (excluding BOS), and BOS attention mass, normalized entropy, and consensus. Lower normalized entropy indicates more concentrated attention; higher consensus indicates greater agreement between heads' attention patterns. Statistics are averaged across heads within each case; error bars indicate 95% confidence intervals.

C.2.2 HEAD AND POSITION STEERING CONTROLS

We use the same two intervention settings at the object token: shifting attention away from correct visual evidence in TP cases, and redirecting image and BOS attention toward the target ROI in FN cases.

We compare the seven *Visual Retrieval Heads* (VRHs) with two control sets: random non-VRH heads and top image-attending heads. Random heads are sampled once with seed 42, matching the VRH count in each layer. Top image-attending heads are selected by ranking all language-model heads by their mean attention mass over image tokens on 100 image-disjoint calibration examples, separately for each query position, and taking the top seven. All head sets are fixed during evaluation. We intervene at the object token, the question-mark token, and END token (the final colon of the assistant generation prefix from which the first answer token is predicted). Table 6 reports the resulting LD for each head set and intervention position.

Table 6: Mean Yes-No logit difference (LD) after attention redirection.

Group	Intervention	Natural	Object token			Question mark		END	
			VRHs	Random	Top-image	VRHs	Top-image	VRHs	Top-image
TP	Image → BOS	4.554 [4.223, 4.870]	−0.214 [−0.494, 0.072]	4.455 [4.127, 4.770]	3.756 [3.468, 4.038]	3.845 [3.537, 4.145]	3.824 [3.525, 4.107]	4.531 [4.204, 4.844]	1.729 [1.357, 2.088]
	Image → other-object ROI		2.059 [1.765, 2.350]	4.460 [4.133, 4.775]	4.441 [4.109, 4.766]	4.086 [3.778, 4.385]	4.548 [4.221, 4.862]	4.537 [4.210, 4.853]	4.614 [4.282, 4.930]
FN	Image + BOS → target ROI	−0.851 [−1.055, −0.658]	1.644 [1.262, 2.049]	−0.373 [−0.580, −0.162]	−0.884 [−1.100, −0.675]	−0.244 [−0.450, −0.047]	−0.856 [−1.061, −0.660]	−1.220 [−1.426, −1.025]	−0.797 [−1.000, −0.604]
	Image + BOS → other-object ROI		0.153 [−0.182, 0.509]	−0.690 [−0.907, −0.489]	−0.871 [−1.079, −0.672]	−0.662 [−0.850, −0.487]	−0.860 [−1.062, −0.669]	−1.308 [−1.526, −1.103]	−0.815 [−1.023, −0.621]

Table 6 shows that interventions on *Visual Retrieval Heads* at the object token have the strongest effects: redirecting image attention to BOS lowers TP mean LD from 4.554 to -0.214, while steering attention to the target raises FN mean LD from -0.851 to 1.644. Random heads and top image-attending heads show smaller effects at this position, and neither control restores a positive mean LD in FN cases. Effects are also weaker at the question-mark and END positions, suggesting that these heads' role depends on where they retrieve evidence, not simply how much they attend to the image.

C.2.3 RELATED-OBJECT STATISTICS

Table 7: Full pair-specific hallucination rates and attention. For each pair, FP rate is $\mathrm{FP}/(\mathrm{FP}+\mathrm{TN}) \times 100\%$ among queries where the queried object is absent and the confuser is visible after preprocessing. Confuser ROI mass is the mean percentage of image attention allocated to the confuser's annotated region in FP cases; brackets indicate 95% confidence intervals.

Queried object	Semantic confuser	FP rate (%)	Confuser ROI mass (%) [95% CI]
chair	couch	75.0	46.9 [35.0, 60.6]
bicycle	motorcycle	75.0	54.7 [40.6, 68.8]
truck	car	74.4	42.4 [31.1, 53.7]
backpack	handbag	68.2	27.2 [16.6, 39.0]
cup	bowl	66.7	30.1 [15.7, 47.0]
bottle	cup	65.0	32.3 [19.9, 45.6]
handbag	backpack	57.7	24.5 [13.6, 37.4]
car	truck	53.8	49.4 [28.2, 71.6]
bowl	cup	50.0	33.1 [24.7, 43.2]
bus	truck	40.0	32.0 [15.4, 47.1]
dining table	chair	37.9	29.2 [17.8, 42.0]
skis	snowboard	33.3	42.3 [42.3, 42.3]
snowboard	skis	22.2	48.2 [26.2, 70.3]
motorcycle	bicycle	16.7	62.6 [33.2, 92.0]
couch	chair	7.1	13.7 [13.7, 13.7]

C.2.4 False Negative Target Visibility Audit

We audit target visibility after CLIP preprocessing in 347 FN cases. The target is cropped out in 40.9% of cases, with no COCO annotation mismatches recorded in this audit. The remaining 59.1% retain a valid visible target. These targets occupy 3.7% of the processed image on average and receive 22.7% of image attention (Table 8).

Table 8: False Negative target visibility after CLIP preprocessing.

Metric	FN
Evaluated cases	347
No target after crop	142 (40.9%)
Valid visible target	205 (59.1%)
Mean target area (% of processed image)	3.7 [2.6, 5.0]
Target ROI mass (% of image attention)	22.7 [17.1, 28.8]

C.2.5 Additional Attention Visualizations

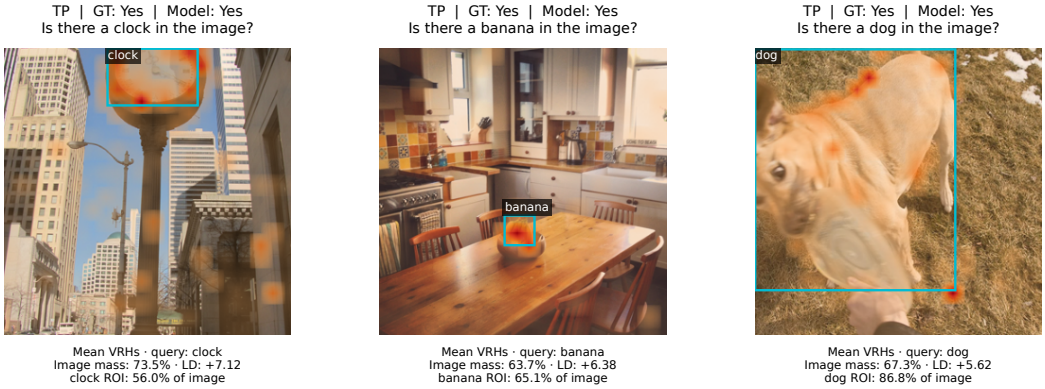


Figure 17: True Positive examples. *Visual Retrieval Heads* ground the queried object in its correct image region.

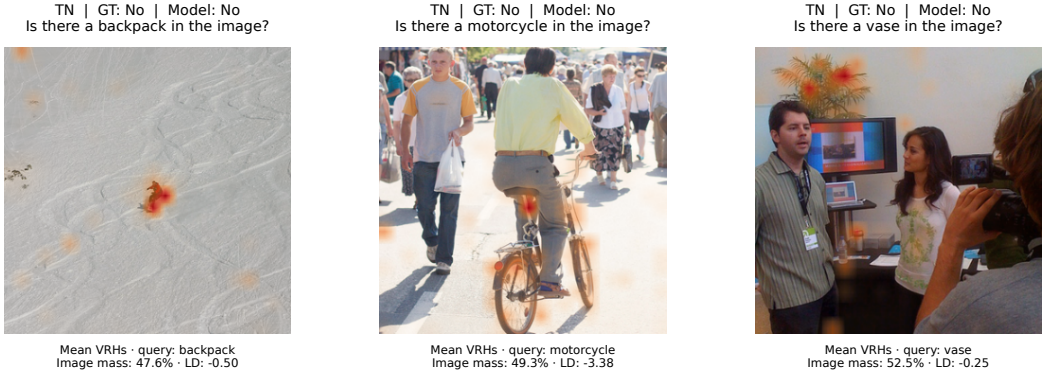


Figure 18: True Negative examples. No target ROI is present, but attention can still fall on related visual concepts without producing a positive answer.



Figure 19: False Positive examples. Attention often falls on semantic confusers; other cases suggest retrieval of related visual concepts from structurally similar objects rather than the queried object.



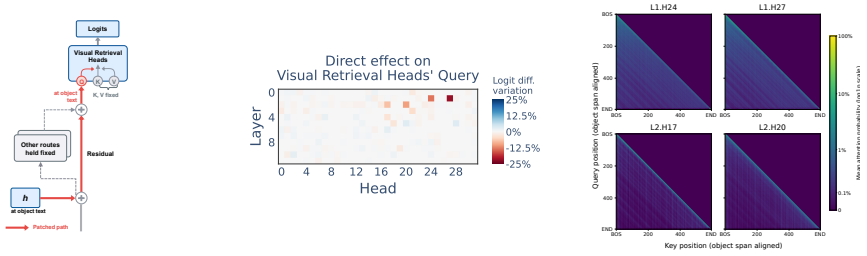
Figure 20: False Negative examples. Missed targets lie outside the CLIP crop (yellow dashed outlines) or are small and visually indistinct, receiving little attention.

C.3 WHY DOES THE MODEL ATTEND TO SEMANTICALLY RELATED REGIONS? (QUERY CONSTRUCTION HEADS)

Since *Visual Retrieval Heads* gather evidence and their attention destinations strongly affect the answer, we next trace which heads direct their attention, rather than the content they read.

Two pathways can affect the attention map at this position: one modifies the object-text representation projected by W_Q, while the other modifies image representations projected by W_K. However, causal masking prevents visual tokens from accessing the subsequent question, so the LLM decoder just further refines the visual information already encoded by CLIP and mapped into its embedding space. These representations therefore remain question-independent for a fixed image prefix, whereas object-text queries specify what to retrieve. We have tried path patching $h \rightarrow K, V_{\text{VRHs}}$ at image-token positions but found no standout heads whose patching substantially reduced LD.

We thus path patch $h \rightarrow Q_{\text{VRHs}}$ at the object-text token and measure the change in the logit difference (Figure 21a-b). Patching L1.H27, L1.H24, L2.H20, and L2.H17 directly alters the *Visual Retrieval Heads'* queries and produces large drops in the logit difference. We refer to these heads as *Query Construction Heads*.



(a) (b) (c)

Figure 21: (a) Diagram of the path-patching experiment on the queries of *Visual Retrieval Heads* at the object-text position. (b) Result of the query path patching, showing sender layers 0-11. Negative values indicate reduced support for Yes. (c) Mean attention maps of *Query Construction Heads* (L1.H24, L1.H27, L2.H17, L2.H20), with object-span-aligned query and key positions and a log1p color scale.

Query Construction Heads write a visual-search signal at the object-text position. Their attention maps show a clear band near the diagonal (Figure 21c), indicating local attention to the current and nearby preceding tokens. This reminds the attention patterns of *Previous Token Heads* in the IOI circuit (Wang et al., 2022).

We then perform patching under four settings: No-to-Yes and Yes-to-Yes, each with the same or a different queried object, and measure the resulting attention in the *Visual Retrieval Heads*. This tests whether patching changes the answer by transferring a "No" signal or by changing which object the model searches for.

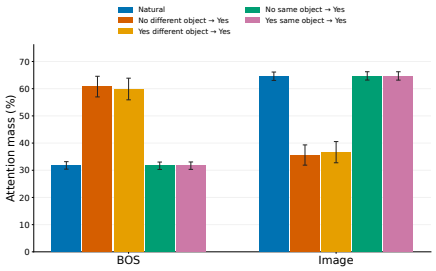


Figure 22: *Visual Retrieval Heads'* attention to BOS and image tokens after query-path patching. Different-object patching shifts image attention to BOS; same-object patching barely changes it.

As expected, different-object patching reduces image attention from 64.6% to 35.6%/36.7% and mean LD from 3.544 to 1.576/1.629 , whereas same-object patching leaves both nearly unchanged. LD may remain positive because patching disrupts the task-specific search signal while leaving object identity largely intact in the residual stream. These results support the role of *Query Construction Heads* in processing the prompt to form an object-specific search signal, rather than transferring an answer signal.

We further hypothesize that these heads shape the object-text representation so that its projection through W_Q aligns with relevant image keys. However, the analysis below suggests that this alignment-related contribution is relatively modest.

When the model hallucinates an object, it often builds a search signal that matches semantic confusers. For each FP example, we ask two separate questions about the same image, one about the queried object and one about its semantic confuser. We measure cosine similarity between two object-tokens' embeddings, representations before and after the *Visual Retrieval Heads*' W_Q projection, and attention maps, using random-token controls at the three representation stages.

Table 9 shows that object embeddings are initially distinct, but become increasingly similar before and after W_Q, exceeding random-token controls and producing similar attention maps. This shows that these object representations is being developed into similar search signals, shift their attention to related visual concepts.

Table 9: Mean cosine similarity for semantic-confuser pairs on the same image. Embedding, Before W_Q, and After W_Q report semantic-confuser / random-token similarity; attention-map similarity is reported only for semantic-confuser pairs.

Object pair	Embedding	Before W_Q	After W_Q	Attention map
chair-couch	0.083 / 0.052	0.755 / 0.512	0.832 / 0.621	0.584
bicycle-motorcycle	0.200 / 0.046	0.750 / 0.503	0.842 / 0.649	0.611
truck-car	0.076 / 0.027	0.756 / 0.508	0.844 / 0.590	0.689
backpack-handbag	0.106 / 0.049	0.778 / 0.511	0.844 / 0.642	0.690
bottle-cup	0.044 / 0.023	0.711 / 0.491	0.810 / 0.581	0.561

We now revisit our hypothesis that *Query Construction Heads* shaped the search queries to match the image. We replace FP queries with queries for the same object from clean TP examples, keeping the FP image keys and values unchanged. If correctly grounded images produce more discriminative queries, this replacement should reduce attention to the confuser in hallucinated cases.

Contrary to this expectation, the replacement and original queries have a mean cosine similarity of 0.916, and the attention maps remain similar after replacement (0.753). We also construct queries from the FP examples after blocking object-token attention to image tokens, then reinsert them into the original VRHs. These queries remain similar to the original queries (0.839), as do the resulting attention maps (0.717).

These results suggest that image information influences query construction but does not fully explain the similarity between confuser queries. Instead, the similarity between these search signals appears to stem mainly from the words' own representations in the language space.

We also compare LLaVA with its base language model, Vicuna. Queries for semantic-confuser pairs are similarly aligned in both models (0.848 vs. 0.847), and each object's query changes little between them (cosine similarity 0.972), suggesting that these similar search signals largely originate in the language backbone rather than multimodal fine-tuning. Figure 23 illustrates how both query replacements preserve attention to a car when the model is asked about an absent truck.

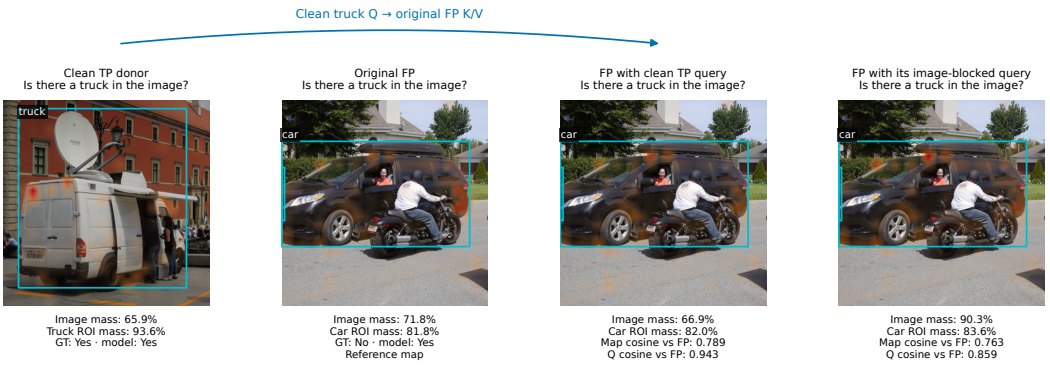


Figure 23: Query replacement for a hallucinated truck. Attention remains focused on the car after inserting either a clean TP query or a query constructed with object-token attention to the image blocked, while retaining the original FP keys and values.

When both objects are present, their search queries remain distinguishable, directing more attention to the target than to its semantic confuser. Surprisingly, the model can still distinguish the two objects when they co-occur (Figure 24). Even after transferring queries from an FP example for the same object into an image containing both the target and its semantic confuser, attention still favors the target: The target's share of image attention falls from 70.9% to 58.8% but remains above the confuser's (15.4%). Query and attention-map similarities remain high (0.912 and 0.754), again showing that image context influences retrieval without overriding the preference for the visible target.

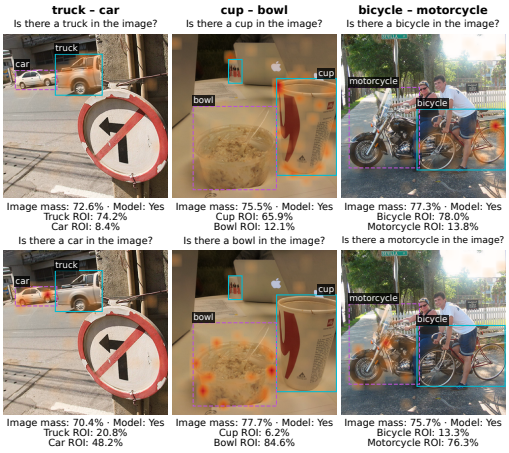


Figure 24: Attention distinguishes co-occurring semantic confusers. Each column shows the same image queried about either member of a pair; attention favors the object named in the question.

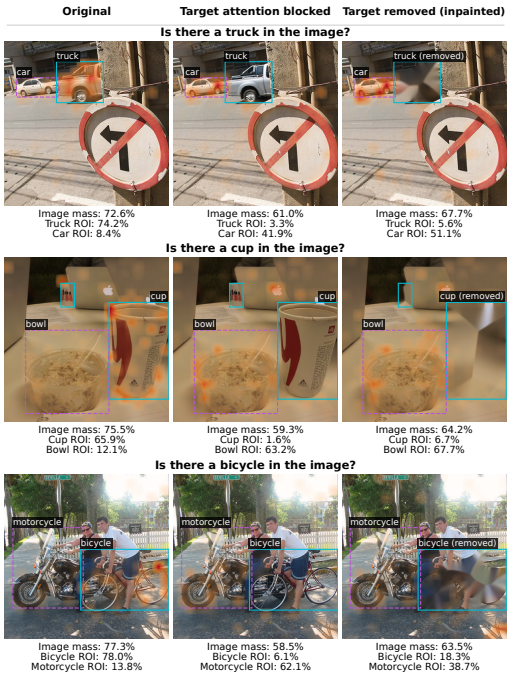


Figure 25: Attention shifts toward semantic confusers when target evidence is unavailable. Columns show the original image, target-region attention blocking, and target removal by inpainting for three object pairs.

However, blocking target-region attention or removing the target through inpainting increases confuser attention by amounts equivalent to 97.4% and 79.5% of the original target attention mass, respectively (Figure 25). One possible explanation is that, when both objects are present, small differences between their queries allow each query to match its target's visual representation more closely, directing attention to the correct region. When the target disappears, its visual evidence is no longer available to attend to; the semantic confuser then has the highest query-key dot product among the remaining image regions, so softmax redistributes attention toward it. The model then uses information from that region as evidence for the queried object in the next stage.

Together, these findings suggest that **hallucination reflects an interaction between similar search queries, available visual evidence, and competition among regions, rather than broad queries alone**. This raises a further question for open-ended generation: *can preceding text both reshape the query for an upcoming object and provide competing textual evidence that attracts attention?* We leave this investigation to future work.

C.3.1 Fine-tuning query construction heads

We further test whether fine-tuning the identified *Query Construction Heads* can reduce hallucinations. We freeze LLaVA-1.5-7B and train additive weight updates only for the Q/K or V/O projection blocks of L1.H24, L1.H27, L2.H17, and L2.H20, using POPE-style COCO training queries with an image-disjoint validation split. We use two losses, separately or combined (hybrid): an answer-label

loss for the correct Yes/No prediction and a localization loss for attention allocation. The localization loss supervises seven downstream *Visual Retrieval Heads* at the object token, favoring the target ROI when present and BOS when absent over competing object regions. All objectives include a KL penalty against the frozen model's next-token distribution.

Table 10: LLaVA-1.5-7B results on POPE-COCO adversarial (3,000 questions).

Method / updates	Objective	Accuracy	F1
Greedy		78.10	80.30
VCD (Leng et al., 2024)		79.03	80.50
OPERA (Huang et al., 2024)		80.40	81.59
AGLA (An et al., 2025)		80.92	81.05
SHIELD (Huang et al., 2026)		82.50	81.90
V/O	Label	84.87	84.11
V/O	Localization	79.43	81.47
V/O	Hybrid	**85.30**	**84.39**
Q/K	Label	81.27	82.20
Q/K	Localization	79.63	81.34
Q/K	Hybrid	80.83	81.89

Table 10 shows that V/O hybrid tuning achieves the best accuracy (85.30%) and F1 (84.39%) among the compared methods, exceeding SHIELD by 2.80 and 2.49 percentage points, respectively. This updates only 0.39% (4/1,024) of the model's attention heads and only their V/O blocks, equivalent to approximately 0.059% of the model's parameters. V/O tuning with answer supervision, alone or combined with localization, outperforms the corresponding Q/K settings, while localization alone provides smaller gains over greedy decoding. These results suggest that refining the information written by *Query Construction Heads* can sharpen the object-search signal, guiding *Visual Retrieval Heads* to attend more precisely to relevant visual evidence.

D Implementation Details

D.1 Visual Retrieval Head Discovery

Discovery protocol. We provide additional implementation details for the visual retrieval head (VRH) discovery described in Sec. 4.1. For each task, we construct a small held-out discovery set containing examples for which the model is correct on the real image and the task-relevant score decreases sufficiently when the image is replaced by a same-size gray null image. We then intervene on each attention head independently by bidirectionally patching its image-token K/V states between the real and null runs. Heads are ranked using the necessity, restoration, and combined causal scores defined in Sec. 4.1. We retain heads with positive effects on at least half of the discovery examples and report the ten highest-ranked heads as *VRHs*.

For LLaVA-1.5-7B, the intervention is applied directly to individual attention heads. For Qwen2-VL-7B, which uses grouped-query attention, discovery is instead performed at the query-head level. Each candidate query head is reconstructed using a private copy of its associated K/V head, and only the corresponding head slice before the output projection is replaced. This isolates the tested query head from other query heads sharing the same physical K/V head.

For object-existence discovery, we use 100 examples per model. Captioning heads are discovered from 100 discovery samples per model. We first let the model generate a caption for each image, identify the onset positions of grounded object mentions in the generated sequence, and then perform head discovery at these object-token onsets. In all cases, the minimum real-null visual gap is 0.25, and the minimum positive-effect consistency is 0.5.

Discovered VRHs. Tables 11 and 12 report the top-10 discovered VRHs and their aggregated causal scores for object existence and captioning. We denote the head with layer index l and head index h by Ll.Hh.

S, N, and R are aggregated independently across discovery examples, therefore their reported aggregate values do not satisfy the corresponding per-example algebraic relation exactly.

Table 11: Top-10 VRHs in LLaVA-1.5-7B for **(a)** object-existence prediction and **(b)** captioning at grounded object-token onsets.

(a) Object existence

#	Head	S	N	R
1	L7.H14	.05359	.05744	.05096
2	L11.H17	.04610	.06495	.04055
3	L8.H6	.04490	.04362	.03751
4	L11.H26	.04490	.02566	.06414
5	L8.H25	.03700	.05047	.02386
6	L10.H7	.02729	.01343	.04028
7	L11.H31	.02582	.01957	.02797
8	L10.H4	.02087	.01494	.02566
9	L10.H29	.02003	.02566	.01429
10	L8.H24	.01778	.01312	.02639

(b) Captioning

#	Head	S	N	R
1	L19.H6	.05000	.02985	.10000
2	L15.H22	.03125	.00000	.04878
3	L17.H18	.03125	.00000	.04167
4	L14.H29	.02976	.00000	.03571
5	L16.H5	.02439	.00000	.03659
6	L20.H12	.02439	.00000	.04762
7	L20.H5	.01961	.00000	.01961
8	L24.H24	.01829	.00000	.03261
9	L17.H30	.01515	.00000	.01829
10	L11.H6	.01042	.00000	.01887

Table 12: Top-10 VRHs in Qwen2-VL-7B for **(a)** object-existence prediction and **(b)** captioning at Qwen-native grounded object-token onsets. Head indices refer to query heads.

(a) Object existence

#	Head	S	N	R
1	L15.H1	.07279	.05955	.10319
2	L14.H13	.05085	.01918	.07157
3	L14.H8	.03854	.00000	.06366
4	L16.H1	.03619	.00000	.07207
5	L15.H20	.03267	.02633	.03399
6	L14.H11	.03126	.02264	.05159
7	L4.H26	.03002	.02601	.02106
8	L14.H16	.02624	.00000	.04830
9	L12.H13	.02601	.00000	.04038
10	L15.H15	.02404	.01527	.03425

(b) Captioning

#	Head	S	N	R
1	L14.H22	.03617	.03451	.00667
2	L13.H16	.03111	.00000	.01647
3	L4.H24	.02206	.00000	.04902
4	L14.H23	.01923	-.01581	.09402
5	L25.H5	.01720	.00980	-.02372
6	L1.H21	.01471	.03827	.00980
7	L9.H3	.01471	-.00791	.02314
8	L27.H19	.01411	.00000	.00000
9	L27.H5	.00984	.00000	.00000
10	L9.H12	.00980	.04137	.00000

D.2 TRACE CONFIGURATION

We evaluate TRACE on LLaVA-1.5-7B and Qwen2-VL-7B. The model-specific inference settings are summarized in Table 13. For LLaVA-1.5-7B, we use $\alpha = 0.8$ and $\beta = 0.4$. For Qwen2-VL-7B, we use $\alpha = 0.7$ and $\beta = 0.4$. The detector confidence is mapped to the probability p_o using a logistic calibrator.

For both object-existence and captioning tasks, TRACE operates on the top-10 task-specific VRHs identified during head discovery. For free-form captioning, we verify at most $K = 5$ object candidates at each decoding step and set the verifier strength to $\gamma = 8$ for both models. The parameters K and γ are used only for captioning; object-existence evaluation uses only the retrieval correction controlled by α and β.

Table 13: TRACE inference settings used in our experiments.

Model	α	β	# VRHs	K	γ	Calibration
LLaVA-1.5-7B	0.8	0.4	10	5	8	Logistic
Qwen2-VL-7B	0.7	0.4	10	5	8	Logistic

D.3 BASELINE CONFIGURATIONS

We use the released implementations of the evaluated training-free baselines and retain their method-specific inference procedures and hyperparameters. For model-method combinations not directly supported by the released implementation, we adapt the inference procedure to the target LVLM while preserving the corresponding intervention and hyperparameters.

Table 14 summarizes the main method-specific hyperparameters used in our experiments. Auxiliary implementation constants are reported separately below to keep the table compact.

Table 14: Core configurations of the evaluated baseline methods.

Method	Core parameters	Values
VCD	$(\alpha_{\text{cd}},\ \beta_{\text{cd}},\ t_{\text{noise}})$	$(1.0,\ 0.1,\ 500)$
OPERA	$(\sigma,\ r,\ N_{\text{can}},\ \lambda_{\text{pen}},\ \beta,\ N_{\text{beam}})$	$(50,\ 15,\ 5,\ 1,\ 5,\ 5)$
AGLA	$(\alpha,\ \beta)$	$(1.0,\ 0.1)$
ONLY	$(\alpha_{+},\ \alpha_{-},\ \beta,\ \gamma_{\text{JS}})$	$(3.0,\ 1.0,\ 0.1,\ 0.2)$
CausalLens	$L,\ (\lambda_{\text{causal}},\ \gamma_{\text{mix}}),$ top-k	10–20, $(0.15,\ 0.15),\ 0.5$
SHIELD	$(\alpha_{\text{cd}},\ \beta_{\text{cd}}),\ \theta,\ (\gamma_{\text{gain}},\ \gamma_{\text{reduce}})$	$(2.0,\ 0.35),\ 0.002,\ (3.0,\ 3.0)$

Additional implementation settings. For VCD, we use `do_sample=False`. AGLA uses BLIP-ITM-large with an input size of 384, Grad-CAM block 6, and a 24×24 Grad-CAM grid. ONLY uses enhancement-layer index 0. For CausalLens, the system/image lengths are $35/576$; we use variance-based sensitivity, no absolute head-count cap, and head-mean normalization. For SHIELD, we use 32 bias samples, a CW configuration $(\text{lr}, \epsilon, \text{steps}, c) = (0.02, 0.14, 30, 12)$, gain/reduce fractions of $0.55/0$, a bias weight of 0.01, and a noise step of 500.

D.4 Evaluation Protocol

Object-existence evaluation. For POPE, all methods receive the original benchmark questions without additional postfix instructions. In particular, CausalLens is evaluated using the same POPE questions as the other methods, without its additional answer-format postfix. We report results on the random, popular, and adversarial splits. For BEAF, all methods are evaluated on the same original-manipulated image pairs and their associated Yes/No questions.

Free-form captioning. For CHAIR, all methods use the prompt "Describe this image." and generate at most 128 new tokens. We evaluate all generated captions with the CHAIR evaluator.

Decoding. We preserve decoding procedures that are part of each baseline method. For example, OPERA uses beam search with $N_{\text{beam}} = 5$. The greedy baseline applies no hallucination-mitigation intervention.

E Evaluation Metric Definitions

E.1 POPE

POPE evaluates object-existence hallucination using binary Yes/No questions. Treating "Yes" as the positive class, we report Accuracy and F1. Let P and R denote precision and recall:

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN},$$
$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \qquad F1 = \frac{2PR}{P + R}.$$

Higher values indicate better performance.

E.2 BEAF

BEAF evaluates whether a model responds appropriately to controlled changes in the visual scene. Let $c_o, c_m \in \{0, 1\}$ indicate whether the model answer is correct on the original and manipulated images, respectively, and let $R \in \{0, 1\}$ indicate whether the question refers to the object removed by the manipulation. Let D_r denote the number of examples with $R = r$.

True Understanding (TU). TU measures questions about the removed object for which the model is correct both before and after manipulation:

$$TU = 100\frac{|\{c_o = 1, c_m = 1, R = 1\}|}{D_1}.$$

Ignorance (IG). IG measures cases in which the model is incorrect both before and after the manipulation:

$$IG = 100\frac{|\{c_o = 0, c_m = 0, R = 1\}|}{D_1}.$$

Stubbornness (SB). BEAF separates stubbornness toward positive and negative answers:

$$SB_p = 100\frac{|\{c_o = 1, c_m = 0, R = 1\}|}{D_1},$$
$$SB_n = 100\frac{|\{c_o = 0, c_m = 1, R = 1\}|}{D_1}.$$

Thus,

$$TU + IG + SB_p + SB_n = 100.$$

Lower IG, SB_p, and SB_n indicate better performance.

InDecision (ID). For questions unrelated to the removed object, the correct answer should not change. ID measures the fraction for which correctness changes after manipulation:

$$ID = 100\frac{|\{c_o = 1, c_m = 0, R = 0\}| + |\{c_o = 0, c_m = 1, R = 0\}|}{D_0}.$$

Lower ID indicates greater stability under irrelevant visual changes.

F1$_{TUID}$. BEAF combines successful adaptation to relevant changes (TU) and stability under irrelevant changes $(100 - ID)$ using their harmonic mean:

$$F1_{TUID} = \frac{2TU(100 - ID)}{TU + (100 - ID)}.$$

Higher F1$_{TUID}$ indicates better overall change-aware performance.

E.3 CHAIR

CHAIR evaluates object hallucination in free-form image captions. CHAIR$_s$ measures the fraction of captions containing at least one hallucinated object:

$$CHAIR_s = \frac{\#\{\text{captions containing a hallucinated object}\}}{\#\{\text{generated captions}\}}.$$

CHAIR$_i$ measures the fraction of generated object mentions that are hallucinated:

$$CHAIR_i = \frac{\#\{\text{hallucinated object mentions}\}}{\#\{\text{all generated object mentions}\}}.$$

To complement the precision-oriented CHAIR metrics, object recall measures the fraction of ground-truth objects that are correctly mentioned in the generated captions:

$$\text{Recall}_{\text{CHAIR}} = \frac{\#\{\text{ground-truth objects correctly mentioned}\}}{\#\{\text{ground-truth objects}\}}.$$

Lower values are better for CHAIR$_s$ and CHAIR$_i$, while higher values are better for $\text{Recall}_{\text{CHAIR}}$.

F Additional Experimental Results

F.1 Full BEAF Results

Table 15 reports the complete BEAF results, including the standard Accuracy and F1 scores and all six change-aware indicators. The full results show that the improvement of TRACE is not limited to a single aggregate metric: it improves both recognition accuracy and the model's response to controlled visual changes.

Table 15: Full results on BEAF. Best and second-best results within each model are highlighted in green and gray, respectively.

Model	Method	Acc.↑	F1↑	TU↑	IG↓	SB_p ↓	SB_n ↓	ID↓	$F1_{TUID}$ ↑
LLaVA-1.5-7B	Greedy	78.50	72.49	30.09	0.17	63.78	5.96	5.46	45.65
	VCD	79.18	72.24	34.70	1.35	56.13	7.82	8.90	50.26
	OPERA	79.85	73.92	34.51	0.50	59.16	**5.83**	6.04	50.48
	AGLA	85.19	76.37	56.70	0.79	24.73	17.78	5.10	70.99
	SHIELD	84.18	76.70	41.90	0.45	46.91	10.74	5.73	58.02
	ONLY	80.06	73.57	33.80	0.17	58.89	7.14	5.32	49.82
	CausalLens	86.72	78.32	52.59	0.28	30.20	16.93	3.95	67.96
	TRACE	**90.72**	**84.91**	**67.00**	**0.06**	**22.95**	10.01	**2.84**	**79.30**
Qwen2-VL-7B	Greedy	88.52	81.69	57.08	0.17	30.14	12.59	3.89	71.62
	VCD	87.79	80.90	57.22	0.63	31.52	10.62	5.17	71.38
	OPERA	87.77	81.12	53.71	0.17	35.88	10.24	4.22	68.83
	AGLA	87.17	78.65	60.23	0.46	20.78	18.53	4.52	73.87
	SHIELD	86.73	80.18	49.49	0.28	41.62	**8.61**	4.72	65.15
	ONLY	88.43	79.15	61.19	0.34	15.02	23.45	2.98	75.05
	CausalLens	87.19	80.71	52.19	0.22	38.70	8.89	4.55	67.48
	TRACE	**91.26**	**84.74**	**71.82**	**0.05**	**13.49**	14.62	**2.18**	**82.83**

TRACE provides the strongest overall change-aware performance on both models. On LLaVA-1.5-7B, it increases TU from 30.09 under greedy decoding to 67.00 while reducing ID from 5.46 to 2.84, resulting in an $F1_{TUID}$ of 79.30. Compared with the strongest competing TU result from AGLA (56.70), TRACE improves TU by 10.30 points. It also reduces SB_p from 63.78 to 22.95, indicating substantially fewer cases in which the model incorrectly preserves a positive response after the relevant visual evidence is removed.

On Qwen2-VL-7B, TRACE similarly raises TU from 57.08 to 71.82 and reduces ID from 3.89 to 2.18, yielding an $F1_{TUID}$ of 82.83. Relative to the strongest competing $F1_{TUID}$ of 75.05 from ONLY, this is a 7.78-point improvement. TRACE also achieves the best Accuracy, F1, IG, and SB_p on both models. The only metric on which it does not obtain the lowest value is SB_n, where OPERA performs best on LLaVA and SHIELD performs best on Qwen2-VL. Overall, the full BEAF results indicate that TRACE primarily improves the model's ability to react to relevant visual changes while also remaining more stable to irrelevant ones.

F.2 Full CHAIR Results

Table 16 reports the full CHAIR results together with object recall and average caption length. Caption length is measured in tokens using the tokenizer of each model, so length comparisons are most meaningful within the same model family.

TRACE achieves the lowest hallucination rates on both models. On LLaVA-1.5-7B, it reduces $CHAIR_s$ from 50.6 to 35.4 and $CHAIR_i$ from 14.6 to 10.0, while also improving over SHIELD at 36.6 and 10.3. On Qwen2-VL-7B, TRACE reduces the corresponding scores from 25.6 and 6.6 to 20.2 and 5.3.

Importantly, these gains are not explained by shorter captions. TRACE keeps the average caption length close to greedy decoding on both LLaVA (103.50 vs. 102.54 tokens) and Qwen2-VL (112.10 vs. 112.45 tokens). TRACE therefore reduces hallucination while largely preserving the original generation length.

Table 16: **Full results on CHAIR.** Lower CHAIR_s and CHAIR_i indicate fewer hallucinations, while higher Recall is better. Avg. Len. denotes the average number of generated caption tokens. Best and second-best CHAIR and Recall results within each model are highlighted in green and gray, respectively.

Model	Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Recall↑	Avg. Len.
LLaVA-1.5-7B	Greedy	50.6	14.6	77.38	102.54
	VCD	55.2	16.7	76.05	105.57
	OPERA	48.6	14.6	76.86	104.25
	AGLA	46.4	13.6	78.74	100.61
	ONLY	53.0	15.9	73.51	111.03
	CausalLens	45.2	13.3	**79.79**	100.95
	SHIELD	36.6	10.3	77.89	110.96
	TRACE	**35.4**	**10.0**	74.63	103.50
Qwen2-VL-7B	Greedy	25.6	6.6	**66.13**	112.45
	VCD	24.0	7.3	62.50	100.28
	OPERA	22.2	5.7	64.17	107.08
	AGLA	21.2	5.8	57.35	126.47
	ONLY	21.4	5.7	56.62	124.80
	CausalLens	22.7	5.8	64.28	110.22
	SHIELD	21.6	5.8	59.25	124.73
	TRACE	**20.2**	**5.3**	62.50	112.10

This improvement comes with a moderate reduction in object recall, from 77.38 to 74.63 on LLaVA and from 66.13 to 62.50 on Qwen2-VL. Overall, TRACE provides a strong reduction in hallucinated objects without relying on aggressively shortening the generated captions.

F.3 Inference Overhead Analysis

We compare the decoding cost of TRACE with greedy decoding and other training-free methods on CHAIR using LLaVA-1.5-7B. All methods are evaluated on a single NVIDIA L40S-48GB GPU with batch size 1. As shown in Figure 26, TRACE takes 6.06 seconds per sample, compared with 2.77 seconds for greedy decoding. It is faster than OPERA and SHIELD, and has a similar runtime to VCD and AGLA. Overall, TRACE achieves stronger hallucination reduction while maintaining a decoding cost competitive with existing training-free methods.

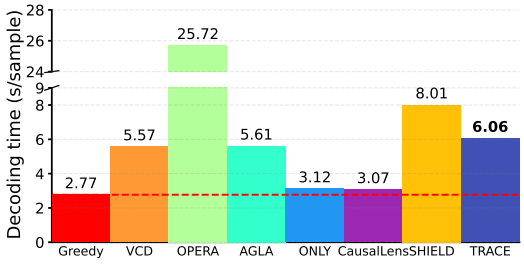


Figure 26: Average decoding time per sample on CHAIR using LLaVA-1.5-7B.

F.4 Additional Ablation Results

Retrieval correction components. We ablate the ROI and null terms in Eq. 2 on a held-out synthetic POPE-like object-existence set. Combining both gives the best result, reaching 95.00% accuracy and 94.81 F1, while removing either component leads to lower performance (Table 17). The two directions therefore play different roles: ROI steering strengthens evidence for detected objects, while null steering is more crucial when target evidence is absent.

Detector confidence calibration. On the same synthetic POPE-like set, we ablate detector confidence calibration. The calibrated score $p = g(s)$ reaches 95.00% accuracy and 94.81 F1, outperforming the raw score $p = s$ (Table 18), showing that calibration provides a better control signal for TRACE.

Table 17: Retrieval correction components.

ROI	Null	Acc.↑	F1↑
✓		91.62	92.03
	✓	93.50	93.12
✓	✓	**95.00**	**94.81**

Table 18: Detector confidence calibration.

Score	Acc.↑	F1↑
Raw $p = s$	92.00	91.40
Calibrated $p = g(s)$	**95.00**	**94.81**

Choice of confidence calibration. TRACE maps the detector score s_o to a calibrated presence probability $p_o = g(s_o)$. We compare several lightweight calibration functions: the raw identity mapping, linear logistic calibration, a log-transformed logistic variant, isotonic regression, and a small MLP. All calibrators are fit using the same calibration data and are kept fixed during downstream inference.

For calibration, we construct a balanced training split containing 512 examples (128 each from TP, TN, FP, and FN cases), together with a separate image-disjoint held-out split of 256 examples (64 per group). The held-out split is used for calibration-model selection and is not used to fit the calibrators.

For the linear-logistic calibrator used in our main experiments,

$$p_o = \sigma(as_o + b),$$

where a and b are learned on the calibration split. The other calibrators differ only in the mapping $g(\cdot)$ and leave the TRACE intervention unchanged.

Table 19 summarizes the average downstream performance of each calibration choice. Calibration consistently improves over directly using the raw detector score. We use linear logistic calibration in all main experiments because it provides the highest mean Accuracy while remaining competitive in F1.

Table 19: Effect of the detector-confidence calibration function. Results are averaged across the three object-existence evaluation partitions. † denotes the calibration used in the main experiments.

Calibration	Acc.↑	F1↑	ΔAcc.
Raw identity	86.66	85.08	–
Linear logistic†	**89.49**	88.74	**+2.83**
Log-scaled logistic	89.26	89.06	+2.60
Isotonic	89.33	**89.15**	+2.67
Small MLP	89.13	88.96	+2.47

All learned calibration functions substantially outperform the raw detector confidence, showing that the detector score is informative but not directly well calibrated for controlling the retrieval intervention. Linear logistic calibration achieves the highest mean Accuracy (89.49), improving the raw mapping by 2.83 points. Isotonic regression gives the highest mean F1 (89.15), but the difference among the learned calibrators is small. We therefore use the simpler linear-logistic mapping throughout the main experiments.

G HYPERPARAMETER SENSITIVITY

The object-existence sensitivity analysis is conducted on a separate held-out sensitivity set constructed from COCO validation images and disjoint from the data used for the ablation experiments. Captioning sensitivity is evaluated on 100 COCO validation images. We vary one parameter at a time while keeping the remaining parameters fixed.

Table 20: Sensitivity to retrieval strengths.

α	Acc.↑	F1↑	β	Acc.↑	F1↑
0.0	93.50	93.12	0.0	91.62	92.03
0.2	94.63	94.41	0.2	95.40	94.81
0.4	94.88	94.68	**0.4**	**95.50**	**95.37**
0.6	95.40	95.35	0.6	95.38	95.20
0.8	**95.50**	**95.37**	0.8	94.50	94.26
1.0	95.20	94.81	1.0	94.25	93.98

Table 21: Sensitivity to caption verification.

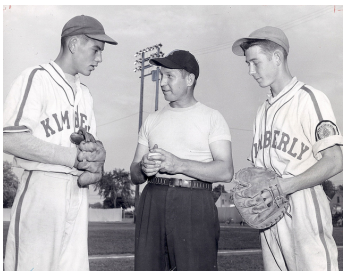
K	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	γ	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$
1	46.0	13.04	2	41.0	11.61
3	41.0	10.78	4	**37.0**	9.92
5	**37.0**	9.79	6	41.0	10.25
7	38.0	**9.52**	**8**	**37.0**	**9.79**
9	38.0	9.80	10	38.0	10.03

Retrieval strengths. We vary α with $\beta = 0.4$, and β with $\alpha = 0.8$. The boundary settings correspond to the individual correction components: $\alpha = 0$ gives null-only correction, while $\beta = 0$ gives ROI-only correction. Both sweeps peak at the setting used in the main experiments, $(\alpha, \beta) = (0.8, 0.4)$, which obtains 95.50% accuracy and 95.37 F1 on the held-out sensitivity set.

Caption verification. We vary K with $\gamma = 8$, and γ with $K = 5$. Increasing K from 1 to 5 substantially reduces hallucination. Larger values provide only minor changes: $K = 7$ slightly improves CHAIR_i but worsens CHAIR_s, while also requiring more object verifications. We therefore use $K = 5$. For γ, $\gamma = 8$ ties the lowest CHAIR_s while achieving the lowest CHAIR_i. We use $(K, \gamma) = (5, 8)$ in the main experiments.

H QUALITATIVE RESULTS

LLaVA-1.5-7B: FP → TN



baseball bat: Yes → No

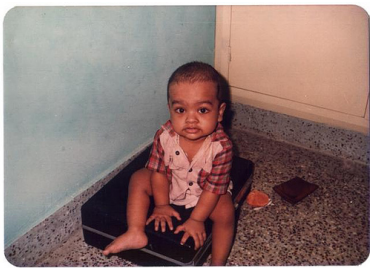


cup: Yes → No

LLaVA-1.5-7B: FN → TP



tennis racket: No → Yes



suitcase: No → Yes

Figure 27: Qualitative object-existence examples on LLaVA-1.5-7B.

Qwen2-VL-7B: FP → TN



handbag: Yes → No



spoon: Yes → No

Qwen2-VL-7B: FN → TP



traffic light: No → Yes



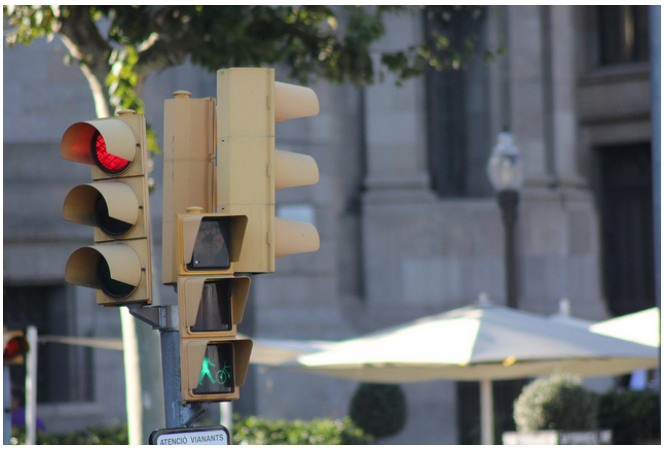
fork: No → Yes

Figure 28: Qualitative object-existence examples on Qwen2-VL-7B.

Table 22: **Qualitative captioning examples on LLaVA-1.5-7B.** Hallucinated objects in greedy captions are highlighted in red. TRACE removes these unsupported objects while preserving the main visual content.



Prompt	Describe this image.
Greedy	The image features a dining table with a plate of food and a glass of milk. The plate contains a variety of food items, including a hot dog, a slice of pizza, and a piece of cake. There are also two strawberries on the plate, adding a touch of color to the meal. The glass of milk is placed next to the plate, completing the meal.
TRACE	The image features a dining table with a plate of food and a glass of milk. The plate contains a variety of food items, including a hot dog, a slice of cheese, and a strawberry. The hot dog is placed towards the center of the plate, while the slice of cheese is located on the left side. The strawberry is situated on the right side of the plate. The glass of milk is positioned on the right side of the table, close to the edge. The overall scene appears to be a casual meal setting, with the food items and the glass of milk creating



Prompt	Describe this image.
Greedy	The image features a traffic light with multiple lights on it, including a red light. The traffic light is located near a building, and there is a street light nearby. The scene also includes a dining table and a chair, suggesting that the area might be a restaurant or a similar establishment.
TRACE	The image features a traffic light with multiple lights on it, including a red light. The traffic light is located near a building, and there is a street light nearby. The scene also includes a street umbrella, which is open and providing shade to the area.

Table 23: **Qualitative captioning examples on Qwen2-VL-7B.** Hallucinated objects in greedy captions are highlighted in red. TRACE removes these unsupported objects while preserving the main visual content.



Prompt	Describe this image.
Greedy	The image depicts a woman standing in a kitchen. She is wearing a brown long-sleeve shirt and denim shorts. The kitchen has a white stove and oven, a white dishwasher, and a white refrigerator. The floor is tiled, and there is a windowed with curtains in the background. The woman appears to be cleaning or wiping down the counter.
TRACE	The image depicts a woman standing in a kitchen, looking at a small object in her hand. The kitchen is equipped with various appliances and utensils. On the left side, there is a white dishwasher, and next to it, a stove with a white oven. Above the the stove, there are several knives in a knife block. On the countert, there are several bottles, a roll of paper towels, and a few other other items. The kitchen also features a sink with a window above it, and a white radiator on the right side of the image. The floor is tiled, and the woman is wearing a brown long-slee



Prompt	Describe this image.
Greedy	The image depicts two individuals standing in a kitchen, engaging in cooking activities. The person on the left is wearing a white long-sleeved shirt and jeans, while the person on the right is dressed in a white shirt and jeans. They are both smiling and appear to be enjoying their time together. The kitchen features a stove with a pan on it, and there are various kitchen items around, such as a cutting board and a knife. The background includes a tiled backsplash and a cabinet with a microwave on top. The overall atmosphere of the image suggests a friendly and casual cooking session.
TRACE	The image depicts two individuals standing in a kitchen, engaged in cooking. The kitchen features a stove with a pan on it, and the individuals appear to be preparing a meal together. The person on the left is wearing a white shirt and jeans, while the person on the the right is wearing a light-colored shirt and jeans. The kitchen has a tiled backsplash and a white cabinet.