THE EVIDENCE ANCHORING BOUNDARY: WHERE VISUAL INTERVENTIONS MATTER AND WHEN THEY HELP IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual interventions provide a training-free way to reduce hallucinations in multimodal large language models. However, an intervention may have little leverage at some layers and may damage a correct answer even when it changes the output. We investigate where visual interventions retain leverage and when they improve predictions. First, we apply the same visual-residual edit at successive decoder layers, matching its RMS strength across depth. Across six checkpoints from three model families, we observe a persistent drop in intervention effect and define its operational onset for this edit as the **Evidence Anchoring Boundary (EAB)**. With boundaries fixed on calibration queries, held-out effects on disjoint evaluation queries before EAB are $13.78\times$–$26.10\times$ larger than those after it. Comparison with intermediate readouts shows that the edit can lose leverage before answer preferences stabilize. Second, we compare counterfactual subtraction across presence, attribute, and relation questions. Subtraction can correct a misleading preference, leave a decision unchanged, or remove a preference supporting the correct answer. Based on these observations, we develop calibrated selective routing, which preserves the original prediction unless calibration supports an intervention for the question type. It improves full-set attribute accuracy by 1.56–1.81 percentage points across three checkpoints, while universal subtraction is neutral or harmful.

1 INTRODUCTION

Multimodal large language models (MLLMs) answer questions about objects, attributes, and relations in an image. However, their answers can include details that the image does not support, resulting in multimodal hallucinations (Rohrbach et al., 2018; Li et al., 2023; Wang et al., 2023; Guan et al., 2024; He et al., 2026). To understand these errors, previous studies examine where models attend, what their intermediate states predict, and how internal interventions affect their outputs (Zhang et al., 2025; Belrose et al., 2023; Shi et al., 2025; Yang et al., 2025). These measurements capture different aspects of visual processing. A layer may attend to an object or support an identifiable answer preference, yet editing its visual representation may have little effect on the final prediction. How that effect changes with decoder depth remains an important question for designing interventions.

Training-free methods address hallucinations by modifying visual inputs, intermediate representations, or decoding behavior (Figure 1, top). They obtain correction signals from alternative visual conditions, intermediate states, attention patterns, or decoding statistics during inference. For example, contrastive methods compare predictions under different visual or instruction conditions, while other methods measure image dependence or adjust attention (Leng et al., 2024; Wang et al., 2024; Favero et al., 2024; Huang et al., 2024; Li et al., 2026a). Layer-selective methods also use internal signals to locate promising intervention sites (Chen et al., 2026a; Shi et al., 2025). For intervention design, identifying such a site leaves two decisions: whether a specified edit still has an appreciable effect there, and whether the resulting change improves the answer for the target question type.

Given the above observations, we investigate two questions: **where can a visual intervention still change the final answer (Q1), and when does that change improve the answer (Q2)?** First, we apply the same visual-residual edit at successive decoder layers, matching the root-mean-square (RMS) strength of every edit (Section 4). This control separates the position of an edit from its magnitude. Across six checkpoints in three model families, including 32B and 34B models, we observe a persistent low-effect region. We call the onset of this low-effect region the *Evidence*

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.

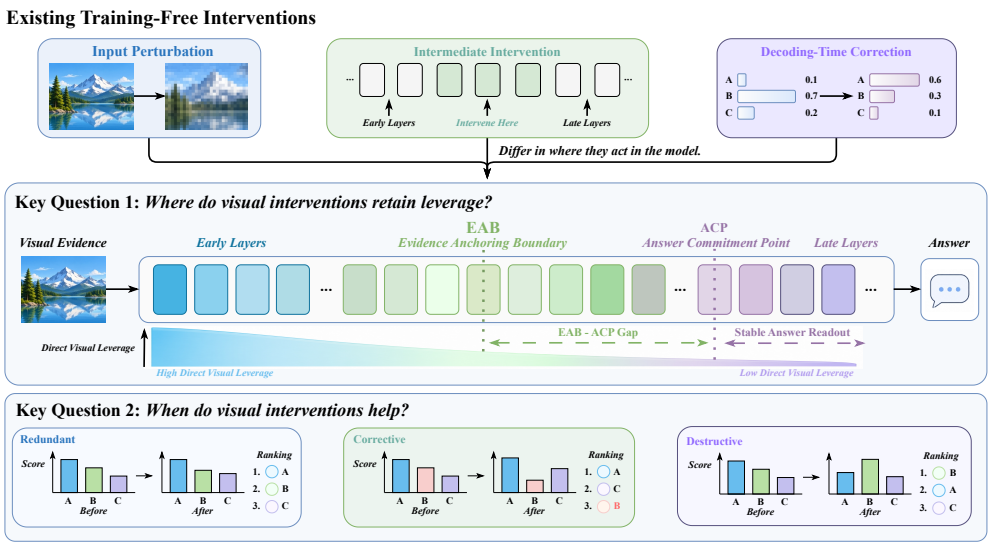

Figure 1: **From intervention leverage to selective visual intervention.** Top: training-free intervention sites. Middle (Q1): the **Evidence Anchoring Boundary (EAB)** marks where the tested visual-residual edit becomes persistently weak under equal-RMS perturbations across depth, potentially before answer stabilization. Bottom (Q2): an edit can be corrective, redundant, or destructive depending on the preference it removes, motivating calibrated selective intervention.

Anchoring Boundary (EAB). We select one operational boundary per checkpoint on calibration queries, freeze it, and evaluate the pre-boundary and post-boundary contrast on disjoint queries: under the frozen rule, pre-EAB effects are $13.78\times$–$26.10\times$ larger than post-EAB effects. Here, EAB characterizes the specified edit and boundary rule.

We then compare EAB with the layer where intermediate answer preferences stabilize, the *Answer Commitment Point (ACP)*. The effect drop can precede readout stabilization: the model continues to develop its answer preference after the tested visual edit has lost most of its influence. A readable intermediate answer and an effective editing site need not occur at the same depth.

Second, we compare counterfactual subtraction across object-presence, attribute, and relation questions (Section 5). We find that the outcome depends on the relative preference removed from competing answers. For a yellow bus, removing a preference for *red* can correct the answer, whereas removing a preference for *yellow* can damage it. Our measurements establish how these cases occur across tasks: subtraction improves misleading-attribute subsets, adds little consistent presence-error information beyond the original confidence, and can hurt direct-contact relations. The same operation can therefore, in the evaluated settings, be corrective, redundant, or destructive.

Based on these findings, we develop a *calibrated selective router* (Figure 4). Specifically, we first use the question type to determine which interventions are eligible. We then use calibration data to decide whether to apply an intervention, preserving the original prediction by default. Experiments show that universal subtraction changes full-set attribute accuracy by $-0.42/-4.92/-0.33$ points across three checkpoints, whereas routing improves it by $1.56/1.81/1.67$ points. Our main contribution is a controlled characterization of intervention depth and value: an equal-strength edit can lose influence before answer readouts stabilize, and subtraction can remove a preference supporting the correct answer. Selective routing shows how these findings guide correction.

2 Related Work

Training-free hallucination mitigation. Previous work reduces hallucinations by changing how visual information contributes to generation during inference (Huang et al., 2026; Yin et al., 2026). VCD compares predictions from original and distorted images, while ICD perturbs instructions (Leng et al., 2024; Wang et al., 2024). Other methods use masking, retrieval, or counterfactual inputs to reduce unsupported preferences (Liu et al., 2024; Chen et al., 2025; Li et al., 2026b). M3ID measures image dependence, DeCo uses intermediate predictions, and OPERA applies decoding penalties and retrospection (Favero et al., 2024; Wang et al., 2025a; Huang et al., 2024; Li et al., 2026a).

Our counterfactual scores follow this training-free setting. We study the conditions under which subtraction improves or damages a prediction across question types and different counterfactual preference structures, and compare selective use with applying the same subtraction everywhere.

Visual localization and layer-selective intervention. Zhang et al. (2025) use attention and gradient maps to guide visual interventions for small details. VIHD identifies decoder windows with high visual attention, masks selected visual tokens, and uses the resulting response distributions to detect hallucinations in medical VQA (Chen et al., 2026a). These studies establish useful intervention sites from spatial or layer-specific signals. We instead measure the final-output effect of a fixed edit at every depth, equalize its RMS strength, and test a calibration-selected low-effect region on independent queries. The distinction is between locating a promising site and measuring how the influence of a specified edit persists across depth for a fixed intervention operator.

Intermediate readouts and causal influence. Logit-lens methods project intermediate hidden states into vocabulary space; tuned lenses learn translators to improve these predictions (Belrose et al., 2023). Causal tracing and activation patching measure how changing internal representations affects the final output (Meng et al., 2022; Yang et al., 2025). Multimodal studies further investigate visual tokens, attention, and information flow across layers (Shi et al., 2025; Kang et al., 2025; Yang et al., 2025; Chen et al., 2026b; Xian et al., 2026). We bring these two measurements together by comparing direct-edit effects with answer-readout stabilization under the same checkpoint and query design. EAB summarizes the former trajectory; ACP summarizes the latter. Their separation shows why intermediate answer availability alone does not determine when a particular visual edit remains effective, even when those readouts continue to evolve downstream.

Multimodal hallucination evaluation. POPE evaluates object presence, AMBER includes attributes and relations, and HallusionBench tests visually confounding questions and prompt presuppositions (Li et al., 2023; Wang et al., 2023; Guan et al., 2024). CHAIR measures object hallucination in generated descriptions (Rohrbach et al., 2018). We use claim-level comparisons to distinguish presence, attribute, and relation effects, and report targeted generation and transfer separately. This separation makes clear which findings concern binary decisions and which extend to generated answers, while keeping task-specific evidence distinct from broader claims.

3 Task and Experimental Setup

Both experiments use binary visual claims. Q1 varies the layer of a fixed edit and measures its final-output effect. Q2 compares original and counterfactual scores to measure whether subtraction improves the decision. The shared format keeps intervention effect and task accuracy distinct.

3.1 Task formulation

We study the model's preference for a binary visual claim before and after a controlled visual intervention. For an image of a yellow bus, for example, "the bus is yellow" and "the bus is red" are competing attribute claims. An object-presence claim can be evaluated on its own; attribute and relation claims can also be compared with alternatives about the same image. We call the claims associated with one image a *claim set*, $\mathcal{C} = \{c_1, \ldots, c_K\}$. Each c_i is tested with the same Yes/No answer format to keep answer scoring directly comparable. For an image x, we compute

$$M(c_i) = \log p_\theta(\text{Yes} \mid x, c_i) - \log p_\theta(\text{No} \mid x, c_i), \tag{1}$$

$$M_{\text{cf}}(c_i) = \log p_\theta(\text{Yes} \mid \bar{x}, c_i) - \log p_\theta(\text{No} \mid \bar{x}, c_i),$$

$$G(c_i) = M(c_i) - M_{\text{cf}}(c_i), \tag{2}$$

Here M is the original Yes/No log-odds: a positive value favors Yes. The counterfactual condition $\bar{x}$ denotes an intervention on visual-token representations, with the prompt and claim held fixed. M_{cf} is the resulting margin, and G measures the change from that condition to the original image. Using G as the answer score therefore subtracts the preference that remains under the counterfactual. That preference can reflect the prompt, model knowledge, and information retained by the chosen ablation; M_{cf} is not assumed to isolate a pure language prior. For competing claims, a larger score indicates a stronger preference. Let $\Delta Q_{ij} = Q(c_i) - Q(c_j)$, for $Q \in \{M, M_{\text{cf}}, G\}$. Then

$$\Delta G_{ij} = \Delta M_{ij} - \Delta M_{\text{cf},ij}. \tag{3}$$

Thus, the relative counterfactual preference $\Delta M_{\text{cf},ij}$ determines how subtraction changes the comparison between two claims. A positive ΔM_{ij} ranks c_i above c_j. For untied scores, subtraction reverses

their ranking when ΔM_{ij} and ΔG_{ij} have opposite signs. The identity specifies the score change, while the experiments determine whether that change is beneficial.

3.2 Models, datasets, and evaluation scope

Our base checkpoints are LLaVA-1.5-7B/13B, Qwen2.5-VL-7B, and InternVL3.5-8B (Liu et al., 2023; Qwen Team, 2025; Wang et al., 2025b). We additionally use LLaVA-NeXT-34B and Qwen2.5-VL-32B to test Q1 at larger scale. Table 1 summarizes the tasks and datasets. Q2 coverage varies by checkpoint; Appendix A.1 gives the evaluated combinations and sample counts.

Table 1: **Main evaluation map.** Q1 uses 120 shared queries per checkpoint (48 calibration, 72 disjoint evaluation); Q2 coverage varies by checkpoint as recorded in Appendix A.1.

Question	Dataset / task	Use in this paper
Q1: layer-wise effect	Shared sampled queries	EAB and answer-readout comparison
Q2: presence	POPE; TVQA-Existence	Confidence-conditioned diagnostics
Q2: attributes	AMBER claim pairs	Outcome analysis and routing
Q2: relations	AMBER claim pairs	Outcome analysis

3.3 Intervention and metrics

For Q1, we edit the visual residual contribution at successive decoder layers and match the root-mean-square strength of the edits across depth. We propagate each edited state through the remaining decoder and measure the change in its final output. The boundary is selected on the calibration split and frozen before measuring the held-out pre-/post-boundary contrast. For Q2, we construct $\bar{x}$ by mean-ablating visual-token representations. This counterfactual supplies the preference removed by subtraction; it is distinct from the equal-RMS depth sweep used for Q1 for comparison.

For presence questions, we compare correct and incorrect Yes predictions at matched original confidence M. Cross-fitted residual $\mathrm{AUC}(G \mid M)$ and incremental log loss measure whether G provides additional information about these errors; residual AUC near 0.5 indicates little additional discrimination. For attributes and relations, we compare pair accuracy under M and G. The hard-attribute subset uses answer labels to explain the role of a misleading counterfactual preference. The router is evaluated on the full attribute set, with no access to this analysis-only label at test time.

3.4 Calibration, comparisons, and uncertainty

Calibration and evaluation use disjoint images, keeping every claim from one image together. We select the attribute routing rule by five-fold image-disjoint cross-validation; each image is scored by a rule selected without it, preventing leakage. Original scoring M, universal subtraction G, and routing are evaluated on the same images. This comparison tests whether selection improves a given intervention, with the matched-token ICD/DeCo audits reported separately in Appendix F.4.

Confidence intervals use paired image-cluster resampling, preserving dependencies among claims about the same image rather than treating claims as independent. For Q1, the held-out intervals quantify the effect ratio conditional on the frozen boundary without reselection. Appendices A.2 and A.3 give the split and reproduction protocols; Appendix F.3 reports the paired tests.

4 Q1: Where Can a Visual Intervention Matter?

We investigate how the effect of a direct visual-residual edit changes across decoder layers under matched intervention strength, then compare this change with answer stabilization.

4.1 A persistent drop in intervention effect

Figure 2a shows the effect of moving the same edit through six checkpoints. We observe larger effects at earlier layers, followed by a persistent decrease. We call the onset of this low-effect region the *Evidence Anchoring Boundary* (EAB). Here, anchoring names the observed loss of influence of this direct visual edit. The boundary is an operational summary of its effect trajectory at a given checkpoint, rather than a universal architectural transition shared across interventions.

Specifically, we edit the state at layer ℓ, pass it through the remaining layers, and measure changes in the final Yes/No margin, answer, and output distribution. All edits have equal root-mean-square strength. For original and intervened output distributions p and $p^{(\ell,\mathrm{int})}$, the distributional effect is

$$D_\ell = D_{\mathrm{KL}}\left(p \,\|\, p^{(\ell,\mathrm{int})}\right), \tag{4}$$

We locate EAB using calibration queries. We select the earliest layer after the effect peak at which both the three-layer local mean and the 75th percentile of the remaining layers fall below 10% of that peak. We then fix the boundary and evaluate it on disjoint queries. Figure 2c shows that effects before EAB are $13.78\times$–$26.10\times$ those after it, with every image-cluster confidence interval excluding one. The larger checkpoints show the same pattern: $26.10\times$ $[22.24, 30.74]$ for LLaVA-NeXT-34B and $18.24\times$ $[14.89, 22.59]$ for Qwen2.5-VL-32B. These ratios establish a held-out contrast between early and late effects under the frozen rule. The reported layer marks the onset of its operational low-effect region. Appendix C.2 discusses location uncertainty and the distinction from a change point.

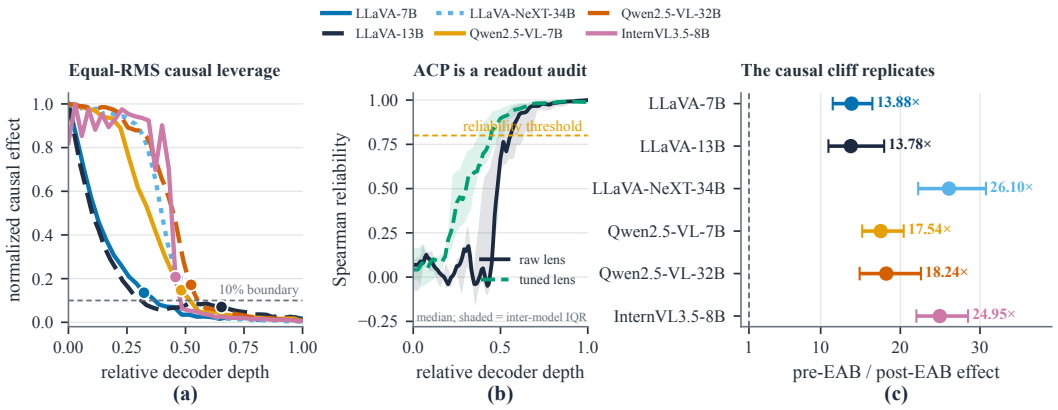

Figure 2: **EAB identifies where the tested direct visual edit loses leverage.** (a) Checkpoint-level boundaries frozen on calibration queries identify a persistent low-effect regime on disjoint evaluation queries under the same equal-RMS strength control. (b) Answer-readout reliability can continue to increase after this measured drop; ACP marks answer-readout stabilization. (c) Fully held-out pre-EAB effects are $13.78\times$–$26.10\times$ post-EAB effects across six checkpoints.

4.2 Separating intervention influence from answer stabilization

To examine whether the drop occurs because the answer has already stabilized, we compare EAB with intermediate answer predictions. We obtain raw readouts by applying the final normalization and language-model head to the answer-position hidden state h_ℓ:

$$z_\ell = W_{\mathrm{LM}}\mathrm{LN}_{\mathrm{final}}(h_\ell), \qquad m_\ell = z_\ell(\mathrm{Yes}) - z_\ell(\mathrm{No}). \tag{5}$$

We use raw and group-held-out binary tuned lenses to identify the *Answer Commitment Point* (ACP), the earliest layer after which decoded margin signs and candidate rankings remain aligned with the final prediction (Belrose et al., 2023; Shi et al., 2025). Figure 2b shows that the readouts can continue to improve after EAB. This result separates two processes: intermediate answer preferences are still developing, while the tested direct visual edit already has little influence on the final output. Post-ACP query edits can also change answers (Appendix B.3), supporting the use of ACP as a readout comparison. Appendix C.1 describes the operator scope.

The first experiment identifies where the tested edit remains effective. We next investigate what makes an intervention useful for a given question beyond causal influence alone.

5 Q2: When Should a Visual Intervention Be Used?

Why does counterfactual subtraction improve some answers but damage others? For the yellow-bus example (Section 3), removing a preference for *red* can correct an error; removing a preference for *yellow* can damage a correct answer. Equal counterfactual scores leave the difference unchanged for

the competing claims considered (Equation 3). Figure 3b illustrates the three ranking outcomes with schematic scores. In the corrective case, $\Delta G = -0.10 - (-0.25) = +0.15$, reversing the preference from red to yellow. Panels (a) and (c) show the aggregate evidence.

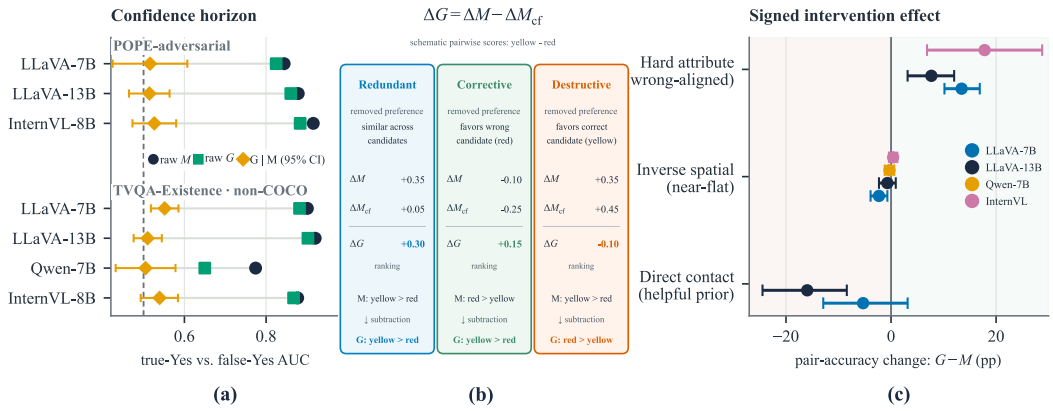

Figure 3: **The removed preference explains three intervention outcomes.** (a) Presence: G adds little consistent discrimination between correct and incorrect Yes predictions after conditioning on M. (b) Schematic yellow-minus-red score differences illustrate $\Delta G = \Delta M - \Delta M_{\text{cf}}$: subtraction preserves, corrects, or damages the ranking according to the removed preference. The values are illustrative, not measured. (c) Observed pair-accuracy changes for hard attributes, inverse-spatial relations, and direct-contact relations; positive values favor G. Error bars show 95% image-cluster confidence intervals, with uncertainty estimated over image-level dependence structures.

5.1 Limited benefit from subtraction

For object presence, we compare correct and incorrect Yes predictions to test whether G adds information beyond the original confidence M. In Figure 3a, residual $\text{AUC}(G \mid M)$ on POPE-adversarial is 0.516, 0.514, and 0.526 for LLaVA-7B, LLaVA-13B, and InternVL. The intervals cover chance, and incremental log-loss gains are unresolved. The non-COCO TVQA-Existence values are $0.552/0.510/0.505/0.539$ across the four base checkpoints. These results show little consistent additional error discrimination after accounting for M across benchmarks.

We also examine inverse-spatial claim pairs. Their counterfactual scores are similar across candidates, so subtraction changes the relative scores little. Accuracy changes range from -2.3 to $+0.4$ points (Figure 3c, middle); the LLaVA-7B decline is statistically resolved (Appendix F.3). Appendices D.1 and G.2 give the related attention diagnostics and examples for completeness.

5.2 When the removed preference is misleading or task-relevant

Using answer labels for analysis only, we identify AMBER attribute pairs whose counterfactual favors the incorrect candidate. On this hard subset, subtraction improves pair accuracy across checkpoints by 13.44, 7.68, and 17.81 points for LLaVA-7B, LLaVA-13B, and InternVL.

However, the full attribute sets contain both corrective and destructive cases under the same intervention. Applying G to every pair changes accuracy by -0.42, -4.92, and -0.33 points. For direct-contact relations, the changes are -5.32 and -15.96 points on the two LLaVA checkpoints. Figure 3c contrasts these effects: subtraction can also remove a task-relevant preference when the original ranking is correct. Table 2 summarizes the results; Appendix G.1 gives examples.

5.3 Calibrated selective routing

The preceding experiments suggest preserving the original prediction unless the question type and calibration results support a correction. We implement this rule with *calibrated selective routing*. First, we assign the question category using dataset metadata or fixed rules over question form. Second, we restrict the eligible interventions to those evaluated for that category. We then use calibration data to select an intervention or retain M before any evaluation image is ever scored. Evaluation labels and correctness are excluded from these decisions to avoid test-time leakage.

Table 2: **Subtraction has different effects across question types for each evaluated checkpoint.** Presence columns report residual $\mathrm{AUC}(G \mid M)$; all other columns report the accuracy change from replacing M with G, in percentage points. (Bracketed values are 95% image-cluster confidence intervals; — indicates that the checkpoint–task pair was not evaluated.)

Checkpoint	Presence: residual $\mathrm{AUC}(G \mid M)$		Paired tasks: accuracy change $G - M$ (percentage points)			
	POPE-adversarial	TVQA non-COCO	Hard attributes (misleading)	Full attributes (mixed)	Direct contact (task-relevant)	Inverse spatial (near-flat)
LLaVA-1.5-7B	.516	.552	+13.44 [10.17, 16.89]	−0.42 [−1.64, 0.84]	−5.32 [−12.90, 3.16]	−2.3
LLaVA-1.5-13B	.514	.510	+7.68 [3.15, 12.01]	−4.92 [−6.39, −3.36]	−15.96 [−24.47, −8.42]	−0.7
Qwen2.5-VL-7B	—	.505	—	—	—	−0.3
InternVL3.5-8B	.526	.539	+17.81 [6.85, 28.77]	−0.33 [−2.83, 2.17]	—	+0.4

Figure 4 summarizes this sequence: category and probes restrict the candidate operators, and calibration determines whether to adjust the score or preserve M by default.

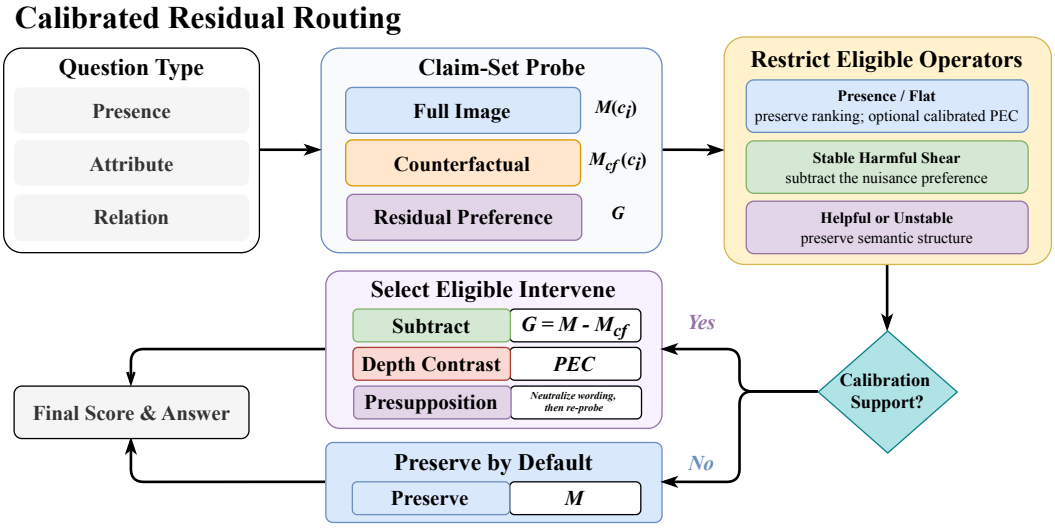

Figure 4: **Calibrated selective routing preserves the original score by default, using only calibration-supported operators for each question category.** Question category and full-image/counterfactual probes restrict eligible operators. Calibration supports subtraction for suitable attribute questions or PEC in its validated presence setting; otherwise the router retains M. For presupposed questions, neutralization is followed by re-probing before a score adjustment is selected. All routing rules are fixed before evaluation, and evaluation labels are excluded from the decision.

For attributes, the router measures how strongly the counterfactual favors one claim over another and uses the score to interpolate smoothly between preservation and subtraction:

$$A(\mathcal{C}) = \max_i M_{\mathrm{cf}}(c_i) - \min_i M_{\mathrm{cf}}(c_i), \tag{6}$$

$$S(c_i) = (1-\beta)M(c_i) + \beta G(c_i), \qquad \beta \in [0,1]. \tag{7}$$

Here $A(\mathcal{C})$ measures the spread of counterfactual scores. Setting $\beta = 0$ retains M, while $\beta = 1$ uses G; intermediate values apply partial subtraction. We select the rule on calibration images using this spread, the fraction of rankings that differ under M and G (DisFrac), and the consistency of the counterfactual-difference sign across equivalent prompts and ablations. Flat, rarely disagreeing, or unstable comparisons retain M. Stable differences can support subtraction when calibration shows a benefit and the sign remains consistent across conditions.

For presence questions, G provides little additional error information, so the default score remains M. In the calibrated POPE-adversarial setting, we additionally allow Prior-Echo Cancellation (PEC):

$$\mathbf{z}_{\mathrm{PEC}} = \mathbf{z}_L + \alpha\,(\mathbf{z}_L - \mathbf{z}_{\ell^\star})\,. \tag{8}$$

This adjustment uses the difference between final logits $\mathbf{z}_L$ and intermediate logits $\mathbf{z}_{\ell^\star}$. We select the layer $\ell^\star$ and coefficient α on calibration data, independently of EAB. EAB identifies effective layers for direct visual-state edits; PEC adjusts output scores. We use PEC only in its validated presence setting. Relation questions retain M unless relation-specific calibration supports subtraction.

For presupposed questions, we neutralize the wording and re-probe before applying the calibrated rule, so the decision uses the neutralized query for scoring (Appendix E.3).

We use five-fold image-disjoint cross-validation for attribute routing. In each fold, we fix the rule before scoring evaluation images, without any evaluation-time adaptation within folds. Appendices E.1–E.3 describe calibration, category assignment, and prompt handling.

6 Evaluation: Benefits, Ablations, and Scope

We compare the router with the original score M and universal subtraction G to test the benefit of selecting when to intervene under fixed routing rules. We then evaluate targeted attribute generation and the transfer of PEC to other tasks. Figure 5 summarizes the results.

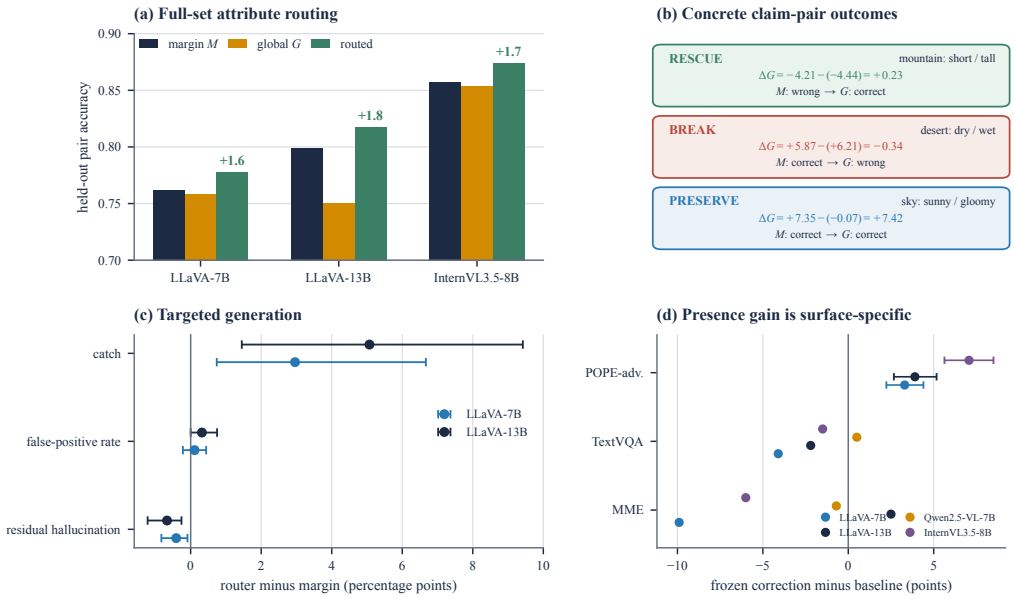

Figure 5: **Selective routing improves attributes; PEC has limited transfer.** (a) Held-out AMBER attribute accuracy for the original margin M, global subtraction G, and the routed score. (b) Cached pairs illustrate rescue, break, and preservation through $\Delta G = \Delta M - \Delta M_{\text{cf}}$. (c) Targeted-generation changes on LLaVA-7B/13B, with 95% image-cluster bootstrap intervals. (d) Frozen PEC effects across POPE-adversarial, TextVQA, and MME: PEC improves its calibrated POPE setting but transfers negatively or inconsistently. Complete results and paired tests are in Appendices F.1–F.3.

6.1 Attribute routing and targeted generation

Universal subtraction changes full-set attribute accuracy by -0.42, -4.92, and -0.33 points on LLaVA-7B, LLaVA-13B, and InternVL. On the same checkpoints, routing improves accuracy by 1.56, 1.81, and 1.67 points (Figure 5a). All three routing gains have image-cluster confidence intervals above zero (Appendix F.3). These results support selecting when to subtract: the full sets contain cases where removing the counterfactual preference damages the answer.

We further apply the frozen routing rules to targeted attribute generation. Hallucination-catch estimates increase by 2.96 and 5.07 points on the two LLaVA checkpoints. False-positive changes are small and unresolved, and residual hallucination decreases (Figure 5c).

6.2 PRESENCE-SPECIFIC INTERVENTION AND TRANSFER FAILURES

Frozen PEC improves POPE-adversarial accuracy by 3.30, 3.90, and 7.07 points on its three evaluated checkpoints. Its effects on TextVQA are $-4.1/-2.2/+0.5/-1.5$ points and on MME are $-9.9/+2.5/-0.7/-6.0$ points across LLaVA-7B/13B, Qwen2.5-VL-7B, and InternVL (Figure 5d). POPE-random changes are only $+0.8/+0.8/-0.3$ points. We therefore restrict PEC to its calibrated POPE-adversarial setting rather than assuming these gains transfer beyond calibration.

6.3 SELECTION, CALIBRATION, AND FAILURE MODES

Table 3: **Evidence for selection and task-specific calibration across the evaluated checkpoints.** Controlled comparisons and observed failure modes. Effects are percentage-point changes unless noted. (Attribute and POPE tests are in Appendix F.3; direct-contact intervals are in Table 2.)

Comparison / design choice	In-domain evidence	Failure / implication
Use G for every attribute (remove selective routing)	*Routed score minus M*	*Global G minus M*
	LLaVA-1.5-7B $+1.56$	LLaVA-1.5-7B -0.42
	LLaVA-1.5-13B $+1.81$	LLaVA-1.5-13B -4.92
	InternVL3.5-8B $+1.67$	InternVL3.5-8B -0.33
Use subtraction across question types	*Presence residual AUC*	*Direct-contact accuracy change*
	LLaVA-1.5-7B .516	LLaVA-1.5-7B -5.32
	LLaVA-1.5-13B .514	LLaVA-1.5-13B -15.96
	InternVL3.5-8B .526	
Apply frozen PEC across tasks	*POPE-adversarial gain*	*Frozen transfer change*
	LLaVA-1.5-7B $+3.30$	Checkpoint POPE-random TextVQA MME
	LLaVA-1.5-13B $+3.90$	LLaVA-1.5-7B $+0.8$ -4.1 -9.9
	InternVL3.5-8B $+7.07$	LLaVA-1.5-13B $+0.8$ -2.2 $+2.5$
		Qwen2.5-VL-7B — $+0.5$ -0.7
		InternVL3.5-8B -0.3 -1.5 -6.0
Retain the stability safeguard	Separate contribution not quantified.	Reject unstable corrections; no isolated accuracy gain is claimed.

Table 3 examines selection and calibration. Applying subtraction to every attribute loses the routing gains, while applying PEC beyond its calibrated task can harm performance. These comparisons support the need to decide where each correction is useful within its calibrated scope. Appendices E and F.4 describe calibration and the scope of the secondary baseline comparisons.

7 LIMITATIONS

EAB describes the tested visual-residual edit at each checkpoint, while ACP measures answer-readout stabilization. The 48/72 calibration/evaluation split supports the reported effect contrast; a larger, more diverse sample is needed to estimate how precisely the boundary location generalizes across settings. The numerical results cover one boundary rule and one primary Q1 operator. New models, prompts, operators, or task categories require recalibration. The routing gains are established mainly for binary attribute claims; broader generation and faithful end-to-end comparisons with external decoding methods remain open. Additional counterfactual passes create a compute cost (Appendix E.4); wall-clock latency and FLOP overhead remain unmeasured.

8 CONCLUSION

We investigate where visual interventions remain effective and when they improve answers. The tested direct visual edit shows a persistent effect drop that can precede answer-readout stabilization. Across question types, subtraction helps or harms according to the preference it removes. Calibrated selective routing improves attribute accuracy where universal subtraction is neutral or harmful. The findings motivate two checks before applying a visual correction: measure the chosen operator's influence at the intended depth, and evaluate the resulting decisions on the target task. Selective routing applies this second check to counterfactual subtraction in calibrated settings.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting, revising, and polishing portions of the manuscript and with writing and debugging a subset of the code used to generate figures. These tools were not treated as sources of evidence. The authors reviewed and edited all AI-assisted text, inspected the AI-assisted plotting code, and verified the resulting figures against the underlying experimental outputs. The authors take full responsibility for the final content of this work, including its text, code, figures, claims, and other artifacts produced with AI assistance.

ETHICS STATEMENT

This work evaluates open-source MLLMs on established public benchmarks and does not involve human-subject experiments or the collection of new personal or sensitive data. We do not release a new dataset or deploy a user-facing system. The studied interventions are intended to reduce unsupported multimodal outputs; however, our transfer results show that they can degrade performance outside their calibrated settings. We therefore restrict our claims to the evaluated models and tasks and emphasize calibrated, default-preserving use. Beyond the general risks associated with MLLM bias, hallucination, and misuse, we are not aware of additional ethical concerns specific to this study.

REPRODUCIBILITY STATEMENT

Section 3 specifies the output scores, interventions, evaluated checkpoints, and claim-specific evaluation scope. Appendix A maps each claim to its evaluated model–task cells and records the experimental protocol and reproduction environment. Appendix B provides the checkpoint-level evidence for Q1, and Appendix C explains the intervention control and boundary interpretation; Appendix D provides the diagnostic checks for Q2; Appendix E specifies the held-out calibration protocol; and Appendix F reports the extended outcomes, transfer tests, and paired significance tests. Appendix G contains mechanism illustrations that are not used for tuning or prevalence claims. Together, these records support the scope, interpretation, and reproduction of the main-text results.

REFERENCES

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

Jiayi Chen, Benteng Ma, Zehui Liao, Winston Chong, Yasmeen George, and Jianfei Cai. VIHD: Visual intervention-based hallucination detection for medical visual question answering. *arXiv preprint arXiv:2605.20772*, 2026a. URL https://arxiv.org/abs/2605.20772. Accepted at MICCAI 2026.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Chen_ICT_Image-Object_Cross-Level_Trusted_Intervention_for_Mitigating_Object_Hallucination_in_CVPR_2025_paper.pdf.

Ting Chen, Geng Li, Guohao Chen, Yu Hu, Guan Huang, Mai Chen, Langsheng Lei, and Jun Du. YARD: Y-architecture register decoding for efficient hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2605.31429*, May 2026b. URL https://arxiv.org/abs/2605.31429.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Favero_Multi-Modal_Hallucination_Control_by_Visual_Information_Grounding_CVPR_2024_paper.html.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Yusheng He, Jizhe Zhou, Xia Du, Zheng Lin, Jun Luo, and Jiancheng Lv. What makes LVLMs hallucinate less? unveiling the architectural factors behind hallucination robustness. *arXiv preprint arXiv:2605.30911*, May 2026. URL https://arxiv.org/abs/2605.30911.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

Xiaoyi Huang, Kejia Zhang, and Zhiming Luo. CHASD: Language increment-calibrated contrastive decoding against hallucination in LVLMs. *arXiv preprint arXiv:2605.23344*, May 2026. URL https://arxiv.org/abs/2605.23344.

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Bohan Li, Haiyang Yu, Yishan Han, Julong Ren, Mengyi Dai, and Shumei Bao. Mitigating multimodal hallucinations through visual attention tracing and origin-point regeneration. *Scientific Reports*, 16:25596, June 2026a. doi: 10.1038/s41598-026-52390-1. URL https://www.nature.com/articles/s41598-026-52390-1.

Xinwei Li, Li Lin, Hui Jiao, Li Yao, Tien-Tsin Wong, and Hanqian Wu. Dive: Decoupling intra-layer visual evidence for mitigating hallucinations in large vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026b. URL https://aclanthology.org/2026.acl-long.1742/.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2310.03744.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024. URL https://arxiv.org/abs/2407.21771.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2202.05262.

Qwen Team. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. URL https://arxiv.org/abs/1809.02156.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision function layer in multimodal llms. *arXiv preprint arXiv:2509.24791*, 2025. URL https://arxiv.org/abs/2509.24791. Accepted at NeurIPS 2025.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2410.11779.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. URL https://arxiv.org/abs/2508.18265.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. URL https://aclanthology.org/2024.findings-acl.937/.

Mingtao Xian, Yifeng Yang, Qinying Gu, Xinbing Wang, and Nanyang Ye. Logit-attention divergence: Mitigating position bias in multi-image retrieval via attention-guided calibration. *arXiv preprint arXiv:2605.11591*, May 2026. URL https://arxiv.org/abs/2605.11591.

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*, 2025. URL https://openreview.net/forum?id=Bjq4W7P2Us.

Xuanyu Yin, Daowan Peng, and Wei Wei. Mitigating hallucinations in vision-language models via contextual entropy calibration. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 2022–2031. Association for Computing Machinery, June 2026. doi: 10.1145/3805622.3810847. URL https://doi.org/10.1145/3805622.3810847.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.17422.

A EXPERIMENTAL COVERAGE AND REPRODUCIBILITY

We give the model and dataset coverage for Section 3, followed by the data splits, intervention settings, and reproduction records needed for independent reproduction.

A.1 CHECKPOINT AND DATASET COVERAGE

Table 4 lists the experiments used in Sections 4–6. Q1 uses the same 120-query design for all six checkpoints. Q2 uses the four base checkpoints, with attribute, relation, and downstream experiments evaluated on the listed subsets. A dash indicates an unevaluated combination.

Table 4: **Claim-specific model and dataset coverage.** Every Q1 checkpoint uses 120 queries (48 calibration/72 evaluation); Q2 and deployment claims are restricted to the populated cells to match their evaluated scope. (POPE counts are adversarial/random; — indicates that the checkpoint–task pair was not evaluated. Statistical audits report their own sample counts in Table 8. LLaVA-NeXT-34B and Qwen2.5-VL-32B are Q1-only scale replications.)

Checkpoint	Q1 queries (EAB/ACP)	Presence claim sets	Attribute pairs	Relation pairs	Downstream evaluation
Base checkpoints used beyond Q1					
LLaVA-1.5-7B	120	POPE: 3,000/3,000 TVQA: 2,000	2,378	contact: 94 spatial: 1,000	attribute router; targeted generation: 1,040; PEC: 3,000; CHAIR: 500
LLaVA-1.5-13B	120	POPE: 3,000/3,000 TVQA: 2,000	2,378	contact: 94 spatial: 1,000	attribute router; targeted generation: 1,077; PEC: 3,000; CHAIR: 500
Qwen2.5-VL-7B	120	POPE: 3,000/3,000 TVQA: 2,000	—	spatial: 1,000	PEC transfer: TextVQA/MME; POPE PEC omitted (two baseline false-Yes cases)
InternVL3.5-8B	120	POPE: 3,000/3,000 TVQA: 2,000	600	spatial: 1,000	attribute router; PEC: 3,000; CHAIR: 500
Q1-only scale replications					
LLaVA-NeXT-34B	120	—	—	—	—
Qwen2.5-VL-32B	120	—	—	—	—

A.2 SPLITS, INTERVENTIONS, AND INFERENCE

EAB uses the same 120 examples for every checkpoint: 48 examples fit the binary tuned lens and determine a single checkpoint-level EAB, which is then frozen. The remaining 72 queries are completely disjoint and are used only to validate post-boundary persistence and estimate the reported pre-/post-EAB effect ratios. The primary layer intervention deletes the visual residual contribution and rescales every edit to equal RMS across depth, separating layer position from perturbation magnitude. No evaluation query is used to select or revise the boundary.

For Q2, the primary counterfactual mean-ablates visual-token representations while holding the prompt and claim fixed. All claims or pairs associated with one image remain in the same split and in the same bootstrap or permutation cluster. PEC is evaluated at the calibration-selected operating point, and PEC presence comparisons use the cached image-cluster McNemar audit. Thus, no evaluation image contributes to threshold selection, and no test-instance label is available to the router. Question category is taken from dataset metadata when available and otherwise assigned by deterministic rules over question form that are fixed before scoring. Unless otherwise noted, image preprocessing, prompt construction, tokenization, and answer extraction follow the released checkpoint defaults and established protocols used in prior multimodal hallucination evaluations (Li et al., 2023; Wang et al., 2023; Guan et al., 2024); deviations specific to our interventions are described in this section.

A.3 REPRODUCIBILITY CHECKLIST

Table 5 connects the frozen choices used by the main claims to the records used for verification. The checklist is organized to distinguish environment control, experimental design, and artifact integrity.

Table 5: **Reproducibility checklist.** Frozen policies and the records used to verify the reported experiments and artifacts, supporting reconstruction of the reported evaluation pipeline.

Item	Frozen value / policy	Reproduction record
Environment and determinism		
Hardware	2× NVIDIA A100-SXM4-80GB; CUDA driver 550.54.14	one checkpoint per GPU; no model sharding for reported base runs
Software	Python; PyTorch 2.8.0, Transformers 4.51.0, NumPy 1.26.4, SciPy 1.17.1	versioned dependency and environment records
Randomness	analysis seed 0; bootstrap seeds 0/1/2 where repeated	deterministic sample designs are cached
Experimental design and inference		
EAB split	120 examples/checkpoint: 48 boundary calibration, 72 completely disjoint evaluation; one checkpoint-level EAB is frozen	cached split assignments and boundary-analysis records
Counterfactual	visual-token mean ablation for the reported Q2 comparisons	fixed prompt and claim
Question category	dataset metadata; deterministic question-form rules otherwise	mapping frozen before scoring; no reference answer or correctness input
Uncertainty	2,000 paired image-cluster bootstrap resamples	resample images with all associated queries/pairs
Hypothesis tests	10,000 cluster sign flips with $(b+1)/(B+1)$ correction; cached McNemar audit for PEC presence	test type specified in Appendix F.3
Attribute router	five-fold image-disjoint cross-validation	cached fold assignments and selection records; labels are never test-time inputs
Artifact regeneration and integrity		
Summary refresh	cached result files only	deterministic aggregation of cached outputs
Paper artifacts	vector PDF/SVG plus 400-dpi PNG	deterministic figure generation and manuscript compilation
Integrity	SHA-256 over result summaries and EAB caches	versioned checksum manifest
Regression suite	cache loaders, statistics, geometry, EAB, and figures	automated validation tests
Baseline scope	matched-next-token audit only for ICD/DeCo	no claim of reproducing full decoding stacks

B Checkpoint-Level Evidence for EAB and ACP

This appendix supplies the checkpoint-level evidence behind Section 4. Section B.1 expands the aggregate EAB result in Figure 2; Section B.2 separates absolute- and relative-depth changes across checkpoint sizes; Section B.3 tests the stronger interpretation of ACP as an irreversibility boundary; and Section B.4 distinguishes spatial localization from causal leverage.

B.1 Per-checkpoint EAB summary

Table 6 reports the checkpoint-level boundaries frozen from calibration queries and the completely held-out effect ratios behind Figure 2. The absolute boundary varies across architectures, while every checkpoint shows a large equal-RMS leverage drop on fully held-out evaluation queries. This table is the numerical basis for the cross-family range reported in Section 4.

Table 6: **The operator-specific EAB pattern replicates across all six checkpoints.** One checkpoint-level boundary is frozen from calibration queries before evaluation; relative EAB is that boundary divided by language-model depth. The resolution column summarizes the cached boundary audit, whereas the pre/post ratios and confidence intervals use the completely disjoint evaluation queries.

Checkpoint	LM depth	Frozen EAB	Relative EAB	Boundary resolution	Equal-RMS pre/post effect	
					Ratio	95% CI
LLaVA-1.5-7B	32	12	.375	99.2%	13.88×	[11.52, 16.49]
LLaVA-1.5-13B	40	13	.325	99.2%	13.78×	[10.98, 17.97]
LLaVA-NeXT-34B	60	33	.550	100%	26.10×	[22.24, 30.74]
Qwen2.5-VL-7B	28	16	.571	100%	17.54×	[15.22, 20.42]
Qwen2.5-VL-32B	64	40	.625	97.5%	18.24×	[14.89, 22.59]
InternVL3.5-8B	36	18	.500	100%	24.95×	[22.02, 28.50]

B.2 Checkpoint-scale comparisons

Table 7 separates absolute layer shifts from relative-depth shifts. The cached paired scale audit gives a $+22.64$-layer shift for Qwen2.5-VL 7B→32B, while its relative EAB shift is unresolved. These audit estimates are distinct from subtracting the frozen checkpoint-level indices in Table 6. The LLaVA-13B→NeXT-34B audit shows a later EAB in relative depth. This comparison also changes model generation and language backbone, so it combines scale and architecture differences.

Table 7: **Checkpoint-scale audit.** (Shifts are larger minus smaller; bracketed values are 95% image-cluster confidence intervals.)

Comparison (depths; resolved)	ΔEAB (layers)	Δ relative EAB	Interpretation
LLaVA-1.5-13B → LLaVA-NeXT-34B 40→60 layers; 119/120 resolved	$+17.72$ $[16.25, 19.12]$	$+.166$ $[.131, .199]$	later EAB in relative depth
Qwen2.5-VL-7B → Qwen2.5-VL-32B 28→64 layers; 117/120 resolved	$+22.64$ $[21.67, 23.66]$	$+.019$ $[-.007, .045]$	relative EAB shift unresolved

B.3 ACP as an answer-readout coordinate

Raw and held-out tuned lenses agree that candidate readouts become more reliable with depth. Query-level interventions can still flip outputs after the readout has stabilized. In the controlled object/attribute/relation audit of 96 triplets per model, the tested windowed edit corrects two LLaVA-7B baseline errors and zero errors for LLaVA-13B, Qwen2.5-VL-7B, and InternVL. The sparse corrections under this controlled audit limit claims about a general correction window. ACP records readout stabilization; intervention tests separately establish causal reach.

B.4 Spatial localization and intervention leverage

The model-level intervention effect and a token-attention map answer different questions. Figure 6 shows one shared POPE example before EAB, near EAB, and after ACP: layers $4/12/24$ for LLaVA-1.5-7B and $4/18/30$ for InternVL3.5-8B. Both models still attend near the queried bowl at late depth even after the aggregate equal-RMS intervention effect has declined. The attention map shows where the model attends, while the intervention-effect measurement in Section 4 tests leverage. This single case illustrates the distinction under these conditions without estimating its prevalence.



Figure 6: **Spatial localization can persist after direct visual-residual leverage fades.** Head-averaged last-query attention on one shared example; green marks the queried-object region.

C EAB Controls and Boundary Interpretation

This appendix explains what the equal-RMS control and frozen-boundary evaluation establish. The numerical evidence is in Table 6; the interpretation depends on both the intervention operator and the boundary rule and does not imply a universal architectural phase transition.

C.1 Matched perturbation strength and operator scope

Equal-RMS edits at every layer control for variation in perturbation norm. Under this control, the primary direct visual-residual edit still has a large held-out pre-/post-EAB effect ratio in every checkpoint. Matching norm does not equate different edit directions, targeted representations, or downstream propagation paths. EAB therefore characterizes the response to the tested operator.

The manuscript does not report a quantitative cross-operator boundary agreement statistic. In particular, the primary Q2 mean-ablation results do not by themselves establish that mean, zero, shuffle, and residual interventions share an EAB. The six-checkpoint replication establishes the depth pattern for the specified Q1 edit, with separate calibration for each checkpoint.

C.2 Boundary location and effect-size uncertainty

The primary rule requires a three-layer local mean and a remaining-suffix 75th percentile below 10% of the checkpoint-level calibration trajectory peak. The selected boundary is then frozen and applied unchanged to the completely disjoint evaluation queries. The confidence intervals in Table 6 resample held-out images while retaining that boundary by construction. They quantify uncertainty in the effect contrast, not in the calibration-selected layer. Likewise, the resolution column records whether the cached boundary audit returns an index; it does not measure the precision of that index.

The reported results do not quantify how the boundary moves under 5%, 15%, or 20% cutoffs, alternative suffix statistics, or persistence lengths. The primary rule gives a reproducible division of the effect trajectory, and the independent evaluation shows that its early and late regions differ substantially. A smooth decay can also satisfy this rule. Distinguishing a change point from smooth decay would require an explicit model comparison; no such comparison is reported here.

D Question-Category Diagnostics for Q2

This appendix supports the interpretation in Section 5 by testing two simpler alternatives: that presence errors can be identified from spatial attention alone, or that prompt form induces a checkpoint-invariant correction. The attention audit evaluates localization; the prompt audit motivates re-probing after neutralization, while keeping their evidential roles explicitly separate in practice.

D.1 Aggregate spatial-attention diagnostic

Figure 7 uses a fixed 300-example subset per base checkpoint. It compares queried-object attention for true-Yes decisions and false-No misses, measures attention to co-occurring objects among false-Yes decisions, and reports PEC suppression for false-Yes and true-Yes cases. The comparison examines whether localization distinguishes correct and incorrect decisions. The observed maps describe where the model attends but yield no stable correction rule for downstream routing across checkpoints and decision types in this audit; we therefore exclude them from router inputs.

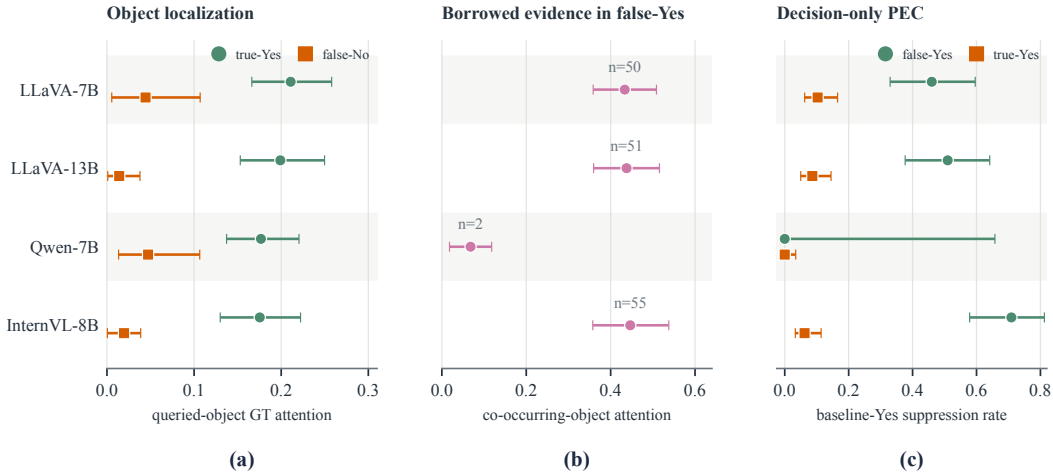

Figure 7: **Aggregate spatial-attention audit across four base checkpoints.** Localization does not by itself identify whether a score intervention will improve the answer.

D.2 Prompt presupposition is checkpoint dependent

Figure 8 uses 150 absent-object examples from each model's frozen POPE subset. Presupposition changes false-Yes rates by $+5.3$ to $+11.3$ points in the two LLaVA checkpoints and Qwen, but by -10.7 points in InternVL; neutralization is likewise strong for some checkpoints and nearly null for Qwen. The effect therefore depends on the checkpoint. We consequently neutralize presupposed wording and then re-probe the claim set before selecting a correction (Appendix E.3).

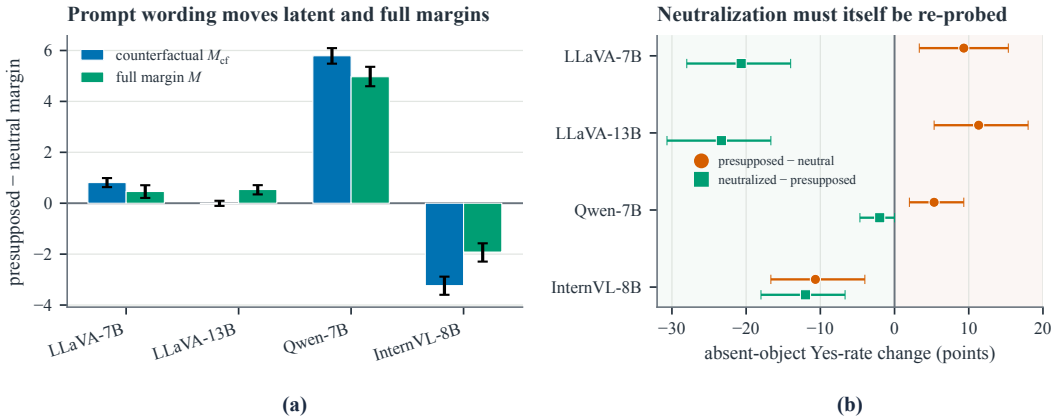

Figure 8: **Prompt presupposition and neutralization are checkpoint-dependent operators.**

E Router Calibration and Routing Protocol

We describe how to select and evaluate the routing rules in Section 5.3 and Figure 4. Calibration uses images separate from those used for evaluation, with no image overlap throughout the routing protocol. The available operators depend on the evaluated model–task combinations in Table 4.

E.1 IMAGE-DISJOINT CALIBRATION

We divide images into five folds, keeping all claims from an image in the same fold. In each rotation, we use four folds to calibrate the attribute router and evaluate it on the remaining fold. We first estimate ranking disagreement, counterfactual score spread, and sign consistency on the calibration data. We then select the threshold and fix the rule before evaluation. Each image receives one prediction from a rule selected without that image. Answer labels define the wrong/helpful alignment groups for aggregate analysis; the router receives no such alignment labels at test time.

For PEC, we select the layer $\ell^\star$ and coefficient α on calibration data. We keep both fixed during evaluation and transfer to other tasks. The transfer results therefore measure how the original configuration behaves on a new task without any task-specific retuning afterward.

E.2 QUESTION-CATEGORY ASSIGNMENT

We use the benchmark's category metadata when available. Otherwise, we apply deterministic rules over question form. We fix the mapping before model scoring and exclude reference answers, predictions, and correctness from its inputs. The resulting category, independent of model output signals, determines which calibrated operators the router can consider.

E.3 PROMPT HANDLING

For presupposed presence questions, we first rewrite the question into a neutral claim and then apply the usual presence probe. We require calibration support for any subsequent score change. Figure 8 shows why this step is needed: the effect of neutralization differs across checkpoints. Re-probing tests the normalized claim for scoring before the router selects an action.

E.4 INFERENCE COMPUTATION

For a set of K claims, the basic subtraction score requires K original-condition scores and K counterfactual-condition scores, compared with K scores for the original model. Additional prompt or stability probes can add evaluations. The routing decision and score blend operate on the resulting scores. Preserving M after probing preserves the prediction but does not recover computation already spent on the counterfactual, even when the original score is preserved.

This count describes the score evaluations implied by the method, not measured latency or FLOPs. Batching, prefix reuse, visual-token count, and prompt re-probing affect the actual runtime. We do not report a wall-clock comparison with the original model or external decoding methods.

F EXTENDED EVALUATION AND STATISTICAL AUDITS

This appendix provides the numerical evidence behind Section 6. It expands the routed attribute and targeted- generation outcomes, reports the calibrated PEC operating range and transfer failures, and gives the image-cluster-aware significance audit. The final subsection states the limited reproduction scope of the secondary baselines without extending beyond the reproduced configurations.

F.1 DETAILED ATTRIBUTE-ROUTING OUTCOMES

Figure 9 expands the attribute and targeted- generation panels summarized in the main-text Figure 5. It shows the absolute held-out pair accuracies behind the reported routing gains and separates hallucination catch, false-positive rate, and residual hallucination for the two targeted-generation checkpoints, while preserving the distinction between these downstream failure modes.

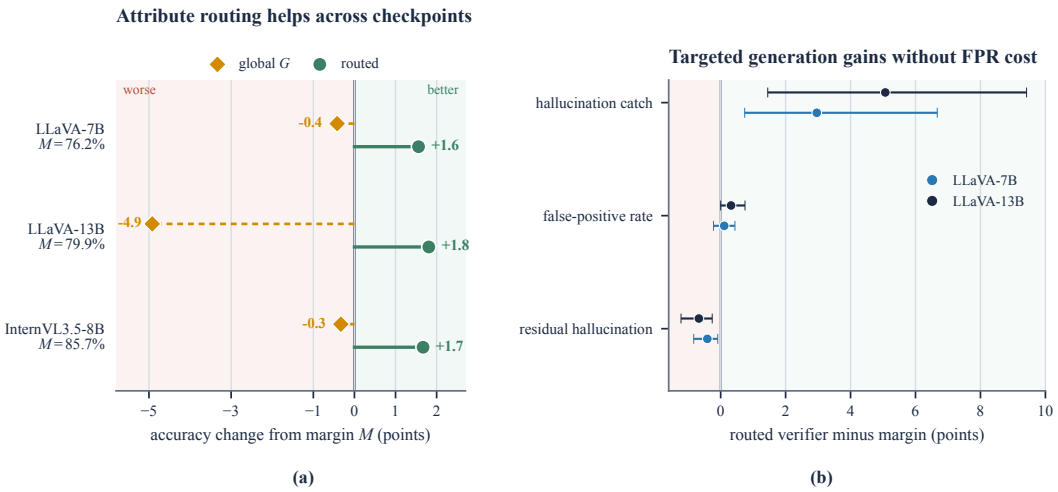

Figure 9: **Detailed attribute-routing and targeted-generation results.** (a) The routed score improves full-set held-out attribute accuracy while global subtraction is neutral or harmful. (b) Point estimates show higher hallucination catch, small false-positive changes, and lower residual hallucination. Horizontal bars are 95% image-cluster intervals; paired tests are in Table 8.

F.2 PEC OPERATING RANGE AND OUT-OF-DOMAIN TRANSFER

Figure 10 shows the α interval over which POPE accuracy is at least the baseline and applies each frozen configuration to TextVQA and MME. Open-ended TextVQA loses accuracy for three of four checkpoints, and full MME changes range from -9.9 to $+2.5$ points. Each transfer test uses the original frozen configuration without any task-specific retuning. On the lower-headroom POPE-random split, the same correction changes accuracy by $+0.8$, $+0.8$, and -0.3 points for LLaVA-7B, LLaVA-13B, and InternVL, further limiting the scope of the presence result.

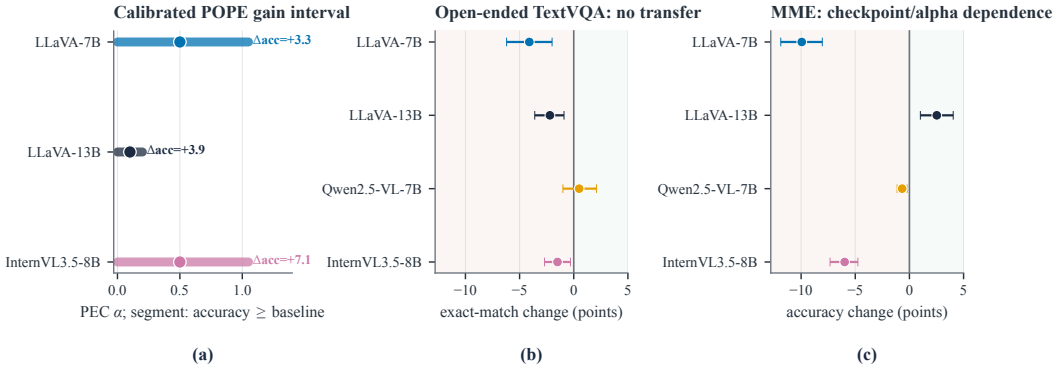

Figure 10: **PEC has a calibrated presence operating range but does not transfer universally.** Segments mark POPE gain regions; transfer whiskers are 95% paired intervals.

Figure 11 decomposes the frozen MME scores. LLaVA-7B loses most on OCR, counting, celebrity, and poster recognition while improving numerical calculation and position; LLaVA-13B improves several categories but loses on posters and landmarks. Qwen changes little in aggregate under its mild configuration, whereas InternVL has substantial celebrity and code-reasoning losses.

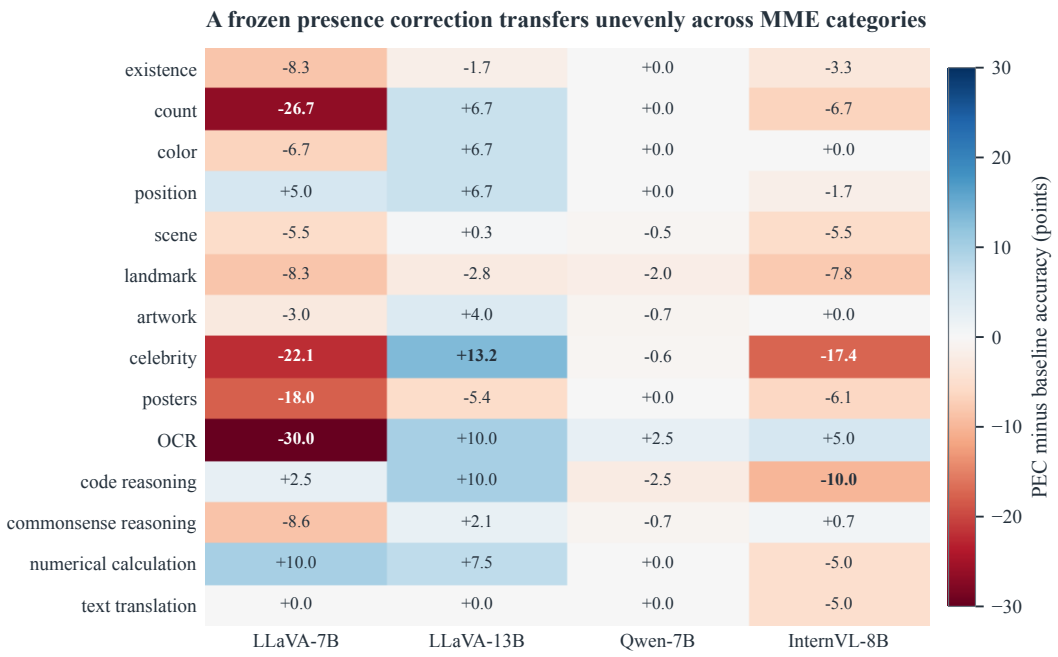

Figure 11: **Frozen PEC transfer is heterogeneous across MME categories and checkpoints.**

F.3 Paired significance audit

For row-level accuracy contrasts, let $d_i = A_i - B_i \in \{-1, 0, 1\}$. The reported effect is $100\bar{d}$ in percentage points. A 95% percentile interval uses 2,000 paired bootstrap resamples of image clusters. For PEC presence comparisons, p-values use the cached McNemar audit (Appendix A.2). For the other accuracy contrasts, the two-sided permutation test flips all d_i from the same image together, uses 10,000 Monte Carlo draws, and applies the finite-sample correction $(b+1)/(B+1)$. Targeted-generation residual hallucination is a ratio, so only its cluster-bootstrap interval is reported. Table 8 reports these tests for attribute routing, presence and inverse-spatial subtraction, targeted generation, and calibrated PEC presence correction. Direct-contact confidence intervals are in Table 2. The LLaVA-7B hallucination-catch increase has a positive bootstrap interval but remains unresolved by the paired test ($p = .1314$); we report both summaries without treating them as interchangeable.

Table 8: **Paired significance audit.** Effects are $A - B$ percentage points; inference is image-cluster aware. (A dash in the p column indicates a ratio outcome for which only a cluster-bootstrap interval is reported.)

Outcome	Checkpoint	n	Effect (pp)	95% CI	p
Attribute subtraction ($G - M$)					
All attributes	LLaVA-1.5-7B	2,378	-0.42	$[-1.64, +0.84]$	.5491
Hard attributes	LLaVA-1.5-7B	521	$+13.44$	$[+10.17, +16.89]$	$< 10^{-4}$
All attributes	LLaVA-1.5-13B	2,378	-4.92	$[-6.39, -3.36]$	$< 10^{-4}$
Hard attributes	LLaVA-1.5-13B	508	$+7.68$	$[+3.15, +12.01]$	.0010
All attributes	InternVL3.5-8B	600	-0.33	$[-2.83, +2.17]$	.8946
Hard attributes	InternVL3.5-8B	73	$+17.81$	$[+6.85, +28.77]$	.0044
Selective attribute routing (router-M)					
Full attributes	LLaVA-1.5-7B	2,378	$+1.56$	$[+0.97, +2.19]$	$< 10^{-4}$
Full attributes	LLaVA-1.5-13B	2,378	$+1.81$	$[+1.18, +2.44]$	$< 10^{-4}$
Full attributes	InternVL3.5-8B	600	$+1.67$	$[+0.67, +2.83]$	.0058

Continued on next page

Table 8 (continued)

Outcome	Checkpoint	n	Effect (pp)	95% CI	p
Presence and relation subtraction ($G - M$)					
TVQA presence	LLaVA-1.5-7B	1,500	-0.80	$[-1.60, 0.00]$	.0656
TVQA presence	LLaVA-1.5-13B	1,500	-0.67	$[-1.27, -0.07]$	.0508
TVQA presence	Qwen2.5-VL-7B	1,500	-1.93	$[-2.87, -1.07]$	.0002
TVQA presence	InternVL3.5-8B	1,500	$+0.07$	$[-0.53, +0.73]$	1.0000
Inverse spatial	LLaVA-1.5-7B	1,000	-2.30	$[-3.90, -0.70]$	.0070
Inverse spatial	LLaVA-1.5-13B	1,000	-0.70	$[-2.30, +0.90]$	.4749
Inverse spatial	Qwen2.5-VL-7B	1,000	-0.30	$[-1.20, +0.60]$	.6434
Inverse spatial	InternVL3.5-8B	1,000	$+0.40$	$[-0.40, +1.20]$	.4507
Targeted generation (router-M)					
Hallucination catch	LLaVA-1.5-7B	135	$+2.96$	$[+0.74, +6.67]$	.1314
False-positive rate	LLaVA-1.5-7B	905	$+0.11$	$[-0.22, +0.44]$	1.0000
Hallucination catch	LLaVA-1.5-13B	138	$+5.07$	$[+1.45, +9.42]$	.0168
False-positive rate	LLaVA-1.5-13B	939	$+0.32$	$[0.00, +0.75]$	.2449
Residual hallucination	LLaVA-1.5-7B	1,040	-0.41	$[-0.83, -0.09]$	—
Residual hallucination	LLaVA-1.5-13B	1,077	-0.67	$[-1.22, -0.26]$	—
Presence correction (PEC-baseline)					
POPE accuracy	LLaVA-1.5-7B	3,000	$+3.30$	$[+2.23, +4.40]$	8.47×10^{-9}
POPE accuracy	LLaVA-1.5-13B	3,000	$+3.90$	$[+2.67, +5.17]$	1.26×10^{-9}
POPE accuracy	InternVL3.5-8B	3,000	$+7.07$	$[+5.63, +8.50]$	4.84×10^{-23}

F.4 Secondary baseline scope

The ICD and DeCo audits evaluate their released scoring formulas through a matched one-token interface. Full decoding additionally involves method-specific control flow and stopping rules. These audits check score behavior under shared conditions; the main comparison evaluates M, universal G, and routing. End-to-end competitiveness with the original decoding methods remains untested.

G Qualitative Examples of Intervention Outcomes

This appendix illustrates the mechanisms summarized in Sections 5 and 6. The selected examples explain decision changes and localization errors; they are not used to estimate prevalence or calibrate the router, while keeping their role illustrative rather than evaluative.

G.1 Attribute rescue, damage, and preservation

Figure 12 instantiates Eq. 3 on three cached LLaVA-7B AMBER pairs, complementing the schematic cases in Figure 3b. The mountain pair is rescued by subtraction, the desert pair loses a correct decision, and the sky pair preserves its ranking. The removed preference explains each outcome.

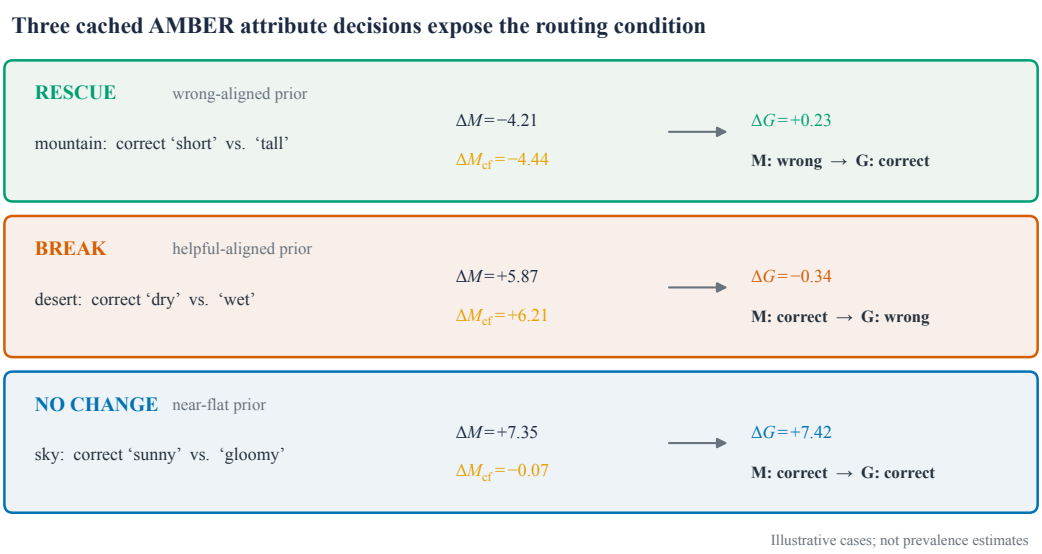

Figure 12: **Concrete attribute decisions expose the routing condition.** Counterfactual subtraction rescues, damages, or preserves a ranking according to the relative preference it removes. These examples illustrate the mechanism and are not prevalence estimates.

G.2 Presence localization cases

Figure 13 shows grounded, borrowed-evidence, and localization-miss cases. Solid green boxes mark the queried object; dashed amber boxes mark present co-occurring objects. The cases clarify why localization is not itself an intervention rule: attention may align with the queried object, borrow support from a co-occurring object, or miss the relevant region. PEC changes the logit-space decision, not the displayed attention map, and these examples are not used as PEC features.

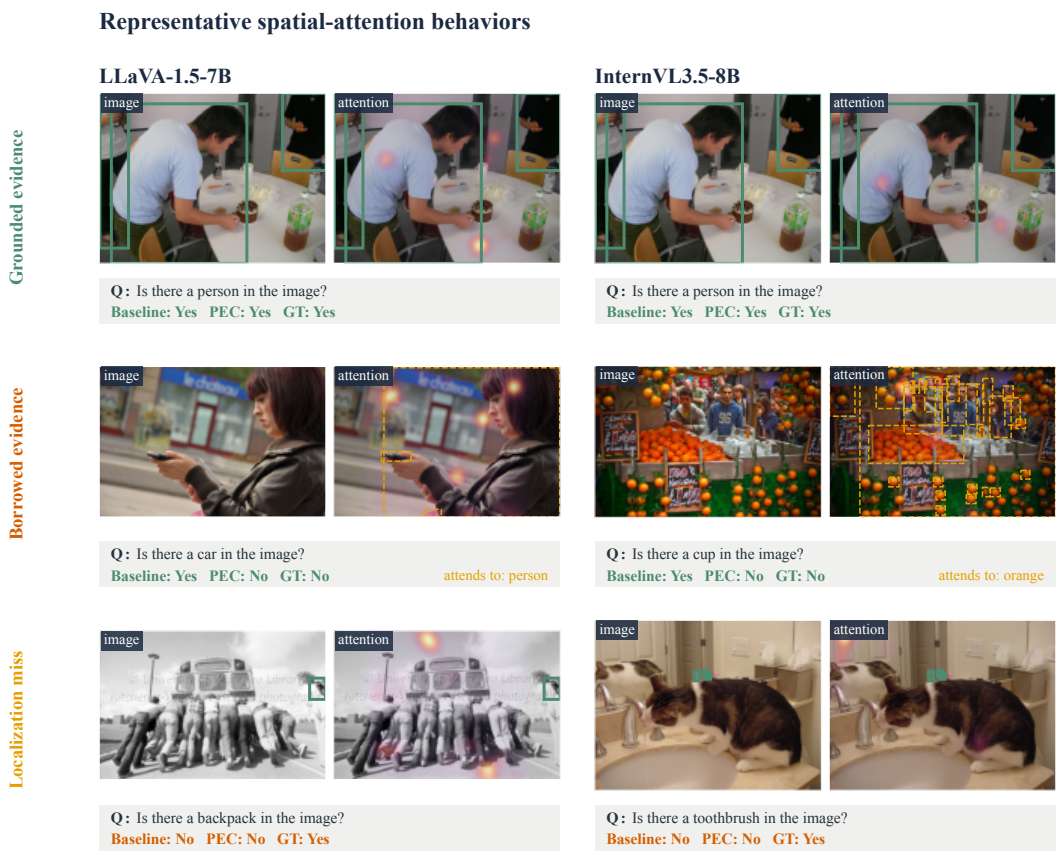

Figure 13: **Representative grounded, borrowed-evidence, and localization-miss cases.**