DOES LOCALIZATION INFORM ADAPTATION FOR VISUAL GROUNDING IN VISION-LANGUAGE MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Correct answers in visual question answering can arise from language priors or irrelevant visual context, so improving visual grounding requires changing how models use evidence. Parameter adaptation offers a route to this change, but leaves open which layers should be updated. Functional localization suggests adapting the layers whose states most strongly influence an answer. This choice assumes that strong state-level influence identifies effective adaptation sites. We test this assumption using a counterfactual VQA dataset that pairs every edit with an operation-matched control and evaluates evidence dependence (Removal), response updates after evidence replacement (Replacement), and invariance to background changes (Context). Across five multimodal large language models (MLLMs; 4B–13B, three architecture families), models respond to relevant evidence but also to irrelevant background changes. Localization and original-state restoration are strongest in early layers, yet similar localization profiles can accompany different single-layer adaptation gains and accuracy costs. On Qwen3.5-9B, localization-ranked placement offers no consistent advantage over random placement at a fixed parameter budget. Joint training with behavior-aligned supervision improves the evaluated grounding metrics, with external abstention gains alongside zero-answer counting and consistency costs. Localization tells us where visual evidence acts on the answer, but not, by itself, where to adapt.

1 INTRODUCTION

Visual grounding in visual question answering (VQA) concerns a model's selective reliance on question-relevant visual evidence when producing an answer (Reich et al., 2023). A correct response from a multimodal large language model (MLLM) does not establish such reliance (Reich et al., 2023; Guan et al., 2024), as it may also result from language priors (Agrawal et al., 2018) or question-irrelevant visual context (Gupta et al., 2022). Improving grounding therefore means changing how a model uses visual evidence so that it relies more on what is relevant and less on what is not.

Parameter adaptation offers a natural route to such change. Semantic-edit training and image-conditioned preference optimization specify how answers should depend on visual input (Agarwal et al., 2020; Wang et al., 2024). A training objective, however, specifies what should change, not where in the model to change it. Functional localization suggests where: intervening on the image-token states at each decoder layer identifies the layers whose states most strongly influence the answer (Palit et al., 2023; Golovanevsky et al., 2025b). Such localization already directs adaptation, from attribution-guided head tuning (Yang et al., 2025) to function-specific LoRA placement (Shi et al., 2025). The underlying hypothesis is that layers with strong functional influence are also the most effective sites to adapt.

Two observations cast doubt on this hypothesis. Localization of this kind is measured in a frozen model: it records how a prediction responds to changes in internal states on a given input, whereas adaptation learns persistent changes to the model's computation across inputs. In language-model factual editing, Hase et al. (2023) found that causal-tracing strength offered little predictive value for edit success beyond layer identity. These observations suggest a *localization–adaptation gap*:

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.

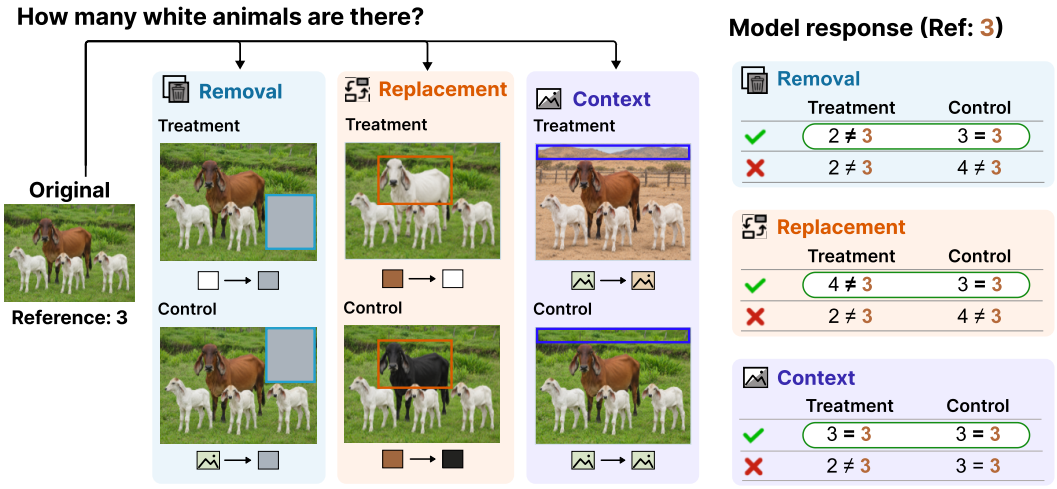


Figure 1: **Paired visual edits and response patterns.** Left: each column pairs a visual edit with its operation-matched control to isolate one behavior. Right: each treatment effect is measured against its paired control. Without the control, an answer change could reflect either use of the evidence or mere disturbance by the edit; the paired difference separates the two. Section 3 defines the construction and scoring.

knowing which layers matter for a behavior need not reveal where adaptation is most effective. We ask whether this gap appears in visual grounding: are the layers that localization singles out also the layers that adapt best, and does using localization to decide where to adapt lead to better adaptation?

Testing whether localization guides adaptation requires that diagnosis and training concern the same evidence-use behavior. We therefore operationalize selective evidence use through three controlled tests, extending counterfactual grounding evaluation (Agarwal et al., 2020; Gupta et al., 2022; Reich et al., 2023) with an operation-matched control for every edit, so that each test separates the model's use of the evidence from its reaction to the edit itself (Figure 1). The tests ask whether answers depend on relevant evidence (**Removal**), follow the new evidence after replacement (**Replacement**), and remain stable under background changes (**Context**). These behaviors also guide training (Appendix C.1).

Localization measures how strongly the image-token states at each decoder layer influence each behavior, and original-state restoration validates this influence on decoded answers; both use activation patching. We test the adaptation value of this signal in two ways: whether localization strength tracks single-layer adaptation gains, and whether the resulting ranking improves adaptation over random placement at a fixed parameter budget. Joint training and external evaluation further test whether the dataset and its behavior-aligned supervision support model adaptation and transfer.

Across five MLLMs, the models respond to the evidence a question is about, and background swaps change their answers as well, relative to matched reconstructions. Localization and restoration are strongest in early layers across these models. In the single-layer scans, Removal gains peak farther into the decoder than localization does (Appendix B.2). In our fixed-budget experiments, adapting the layers selected by the ranking offers no consistent advantage over random placement. Joint training improves the evaluated grounding behaviors and yields gains on related tasks in public benchmarks, supporting the dataset and its behavior-aligned supervision as a basis for model adaptation. Our contributions are summarized as:

- We make visual grounding a behavior that can be diagnosed and trained in the same terms: our counterfactual VQA dataset and evaluation protocol pair every edit with an operation-matched control, and the same behaviors guide diagnosis and training.
- We show across five MLLMs (4B–13B, three architecture families) that these models respond to question-relevant visual evidence and also to irrelevant background changes, and that localization and restoration are strongest in early layers.

- We document a localization–adaptation gap in visual grounding: across behaviors and backbones, the strength of this localization signal need not track single-layer gains, and as a layer-selection rule under a matched parameter budget it yields no consistent advantage over random placement.

2 Related Work

2.1 Evaluating Visual Grounding and Counterfactual Robustness

Counterfactual protocols for visual grounding have converged on paired conditions: answers should change when question-relevant evidence is edited and hold when irrelevant content changes. Question-prior benchmarks target language-prior reliance (Agrawal et al., 2018; Cadene et al., 2019; Chen et al., 2020; Niu et al., 2021). On the image side, Agarwal et al. (2020) test invariance under irrelevant-object removal and covariance under counted-instance removal, then reuse edited images for training. Subsequent protocols swap features of irrelevant objects (Gupta et al., 2022), require per-question necessity–sufficiency judgments (Reich et al., 2023), track answer changes across removed and retained objects (Moon et al., 2024), and measure co-occurrence bias for contextual and absent objects (Xu et al., 2026). In these image-side protocols, the non-evidence edit is read as a separate test or as part of a joint criterion. We instead pair every edit with an operation-matched control and read each test as a within-item treatment-minus-control effect (Section 3).

2.2 Functional Localization and Parameter Adaptation

Probes and attention can show that visual information is present in a model's representations, but presence does not entail use (Hewitt & Liang, 2019; Elazar et al., 2021; Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Activation patching tests use directly by intervening on states: Palit et al. (2023) adapt causal tracing to BLIP, NOTICE (Golovanevsky et al., 2025b) replaces Gaussian corruption with semantically contrasting image pairs, and, in language models, corruption and metric choices can change which components appear important (Zhang & Nanda, 2024). State interventions leave parameters unchanged: PvP (Golovanevsky et al., 2025a) steers outputs toward visual evidence, Li et al. (2026) inject representations located by causal tracing into later layers, and Nooralahzadeh et al. (2026) run diagnosis, patching, and steering on visual–prior conflicts, finding steering more effective at earlier layers than at the diagnostic crossover. Where a behavior is diagnosed and where it is best changed can therefore differ even within state interventions. We compute patching scores for each behavior against its operation-matched control and use original-state restoration to test whether the influence reaches decoded answers (Section 4).

Parameter adaptation involves choices of supervision and update location. Image-conditioned preference optimization supplies supervision that rewards evidence use (Wang et al., 2024; Xie et al., 2024). Localization has guided the update location in three forms: editing existing weights at located layers (Basu et al., 2024; Chen et al., 2025; Zhou et al., 2025), learning per-head offsets with pretrained weights frozen (Yin et al., 2024), and fine-tuning attribution-selected attention heads, compared against an equal-count random set and full-parameter tuning (Yang et al., 2025). Shi et al. (2025) select layers where swapping vision tokens changes counting predictions and compare LoRA on those layers with full-depth LoRA of larger budget and with LoRA on unselected layers.

Whether the localization signal itself indicates where adaptation is most effective nonetheless remains open. Hase et al. (2023) asked this question for factual editing in language models: once layer identity is known, causal-tracing strength adds little to predicting edit success, so a localization signal has to be evaluated at the level where it is used. We carry the question to visual grounding at two levels: layer by layer, testing whether localization strength tracks single-layer adaptation gains (Section 4.3), and as a placement rule compared with random and full-depth placement under a fixed trainable-parameter budget (Section 5.2).

3 Problem Setup and Controlled Evaluation

3.1 Problem Formulation

We study controlled VQA questions about object color and count, for which the question-relevant evidence is a known region of the image. A model is visually grounded when its decoded answer

depends on that evidence rather than on irrelevant content. We test this through three behaviors, each defined by a pair of edited inputs (Section 3.2): whether the answer changes when the evidence is occluded (**Removal**), follows the new content when the evidence is replaced (**Replacement**), and stays fixed when only irrelevant context changes (**Context**).

Adaptation starts from a frozen base model M_0 and trains LoRA adapters (Hu et al., 2022) at a set S of decoder layers. Functional localization assigns each layer a behavior-specific score $\Gamma_a(\ell)$ from state interventions (Section 4.1).

We ask whether this score predicts gains from adapting individual layers, and whether using it to select layer sets improves adaptation over random or full-depth placement at a fixed total parameter budget. We evaluate both questions with the behavioral measures of Section 3.3.

3.2 Counterfactual Data and Matched Controls

Each behavior is defined against an alternative explanation. An answer that changes under evidence occlusion may reflect dependence on the evidence or disturbance by the edit; an answer that changes under replacement need not follow the new content; and an answer that changes under a background swap may reflect the reconstruction rather than the new background. For each original image–question pair we therefore construct a treatment edit and a control edit that applies the same operation without changing the evidence–answer relation (Figure 1). Removal occludes the evidence region (treatment) or an equally sized non-evidence region (control). Replacement changes the evidence to support a validated new answer y_i' (treatment) or applies a matched edit to a non-evidence region, which preserves the original answer y_i^* (control). Context swaps the surrounding background for external content (treatment) or reconstructs it through the same pipeline (control). Every edited image and every y_i' is validated without consulting the evaluated models. Dataset examples appear in Appendix D.

Images come from COCO (Lin et al., 2014). Items that share an original image form an image group, the unit for splitting and for bootstrap resampling. All adapters use the same 1,000 training records. Each record is anchored to one original image–question pair, and its behavior-specific treatment and control views share that original image. All in-domain experiments use the same evaluation dataset, with 400 items for each of Removal, Replacement, and Context. Results are reported at the behavior level, averaging equally over the two question types (Appendix A.1); gain interpretation is discussed in Appendix B.2. Evaluation images are disjoint from training images (Appendix A.1).

3.3 Behavioral Measures and Baseline Diagnosis

Let $x_i(c)$ be the input of item i under condition $c \in \{\mathrm{tr}, \mathrm{ct}\}$ of behavior a, and $r_{i,a}^M(c) \in \{0, 1\}$ the scored response of model M on it. For Removal and Context, the score is one when the edited response differs from the model's own answer on the original input, correct or not; for Replacement, it is one when the response matches y_i', with the same target for treatment and control (answer matching: Appendix A.2). We subtract the control score from the treatment score within each item and average over the eligible set $\mathcal{P}_a$:

$$\delta_{i,a}[h] := h_{i,a}(\mathrm{tr}) - h_{i,a}(\mathrm{ct}),$$
$$\Delta_a(M) := \mathbb{E}_{i \in \mathcal{P}_a}\left[\delta_{i,a}[r^M]\right], \tag{1}$$

where h is any per-condition readout; for the base model we write Δ_a.

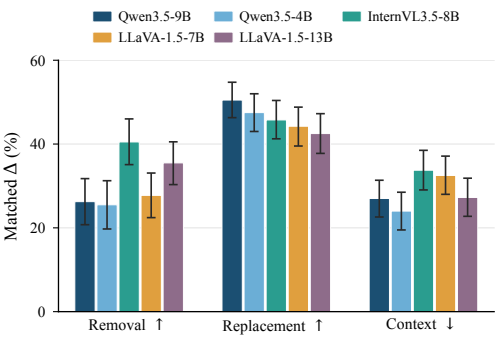


Figure 2: **Five MLLMs respond to question-relevant evidence and also to background changes.** Bars are matched effects (treatment minus control, %) with 95% confidence intervals; arrows mark the desired direction. Exact estimates and sample sizes are in Appendix Table A.1.

Positive Removal and Replacement effects are desired: occlusion changes answers, and replacement produces y_i', more often than their controls do. A positive Context effect indicates unwanted sensitivity to background changes (Figure 1, right). Matched effects score whether responses change or follow the edit, not whether they are

correct: a Removal flip only records that the answer moved, and a stable Context answer can come from ignoring the image. We check original-image accuracy when interpreting adaptation effects (Appendix B). We apply these measures to five MLLMs from three architecture families: Qwen3.5-4B and Qwen3.5-9B (Qwen Team, 2026), InternVL3.5-8B (Wang et al., 2025), and LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2024).

All five models share one pattern (Figure 2): they respond to the evidence the question is about, and a new background also changes the answer more often than the reconstructed original does. The pattern holds across architectures and scales, although its magnitude varies between models. Responsiveness to relevant evidence thus coexists with sensitivity to irrelevant context: the models use visual evidence, but not selectively. These behavioral tests specify what adaptation should improve.

4 Localization and Layerwise Adaptation

We compare localization, restoration, and single-layer adaptation to test whether the layers most strongly influencing an answer are also effective adaptation sites. Figures 4 and 5 detail Qwen3.5-9B; Appendix B presents the cross-model results.

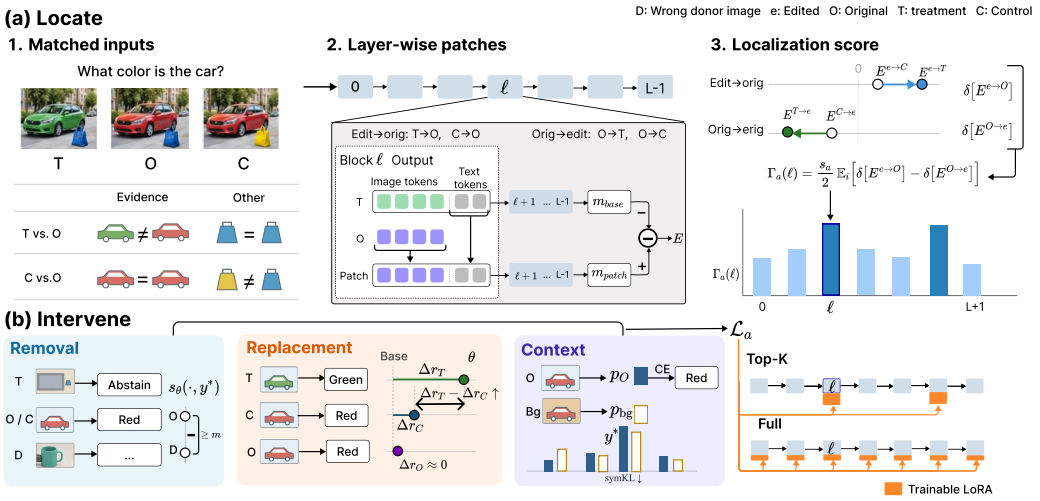


Figure 3: **Locate and Intervene (schematic).** (a) Bidirectional image-token patching in the frozen model yields the localization score $\Gamma_a(\ell)$ (Equation 2). (b) Orange adapters are trained separately for each behavior and allocation. D is a wrong donor image, scored against O for the same question and answer; s_θ is mean teacher-forced answer-token log probability. Δr is the change from frozen Base in new-versus-original answer preference. Context compares original (solid blue) and background-edited (ochre outline) distributions; CE supervises only the original. Specialist objectives are summarized in Equation 3. Localization rankings select Top-k layers; Full uses all layers at lower rank under the same parameter budget. Displayed scores, distributions, and layer slots are schematic.

4.1 Behavior-Specific Layer Localization

Following activation patching (Meng et al., 2022; Zhang & Nanda, 2024), we exchange aligned image-token states between original and edited inputs at decoder-block outputs, retain text-token states, and recompute the remaining network (Figure 3a). We patch both directions for treatment and control. Let $E^d_{i,a}(c;\ell)$ denote the patched-minus-unpatched answer-token margin at layer ℓ in direction d, using the reference answer for Removal and Context and y'_i for Replacement. Applying the matched operator of Equation 1, the signed localization score is

$$\Gamma_a(\ell) = \frac{s_a}{2}\mathbb{E}_{i\in\mathcal{P}^{\mathrm{loc}}_a}\Big[\delta_{i,a}[E^{\mathrm{edit}\to\mathrm{orig}}(\ell)] - \delta_{i,a}[E^{\mathrm{orig}\to\mathrm{edit}}(\ell)]\Big], \tag{2}$$

where $s_a = +1$ for Replacement and -1 for Removal and Context. Larger positive $\Gamma_a(\ell)$ indicates greater matched influence of the layer's image-token states on the answer margin. Implementation details are in Appendix B.1.

Across all five models, localization is strongest in early layers and declines with depth, with an earlier decline in the pooled LLaVA profiles (Figure B.1a). The detailed scan in Figure 4a shows steps at full-attention blocks. Localization rankings closely follow shallow-to-deep layer order (Table B.1).

4.2 Validating Influence on Decoded Answers

Γ_a is measured on the answer-token margin, which can shift without changing the decoded answer. To test its influence on decoded behavior, we restore original-image states into both edited inputs at layer ℓ. Let $R_a(\ell)$ be the restored-minus-unpatched matched behavioral effect over the restoration-eligible items. We report reversal as $-100R_a(\ell)$, so positive values indicate reversal of the edited-input effect (Appendix B.1).

Across the five models, restoration changes decoded answers most strongly in early layers, with reversal declining toward zero at later layers (Figures 4a and B.1b).

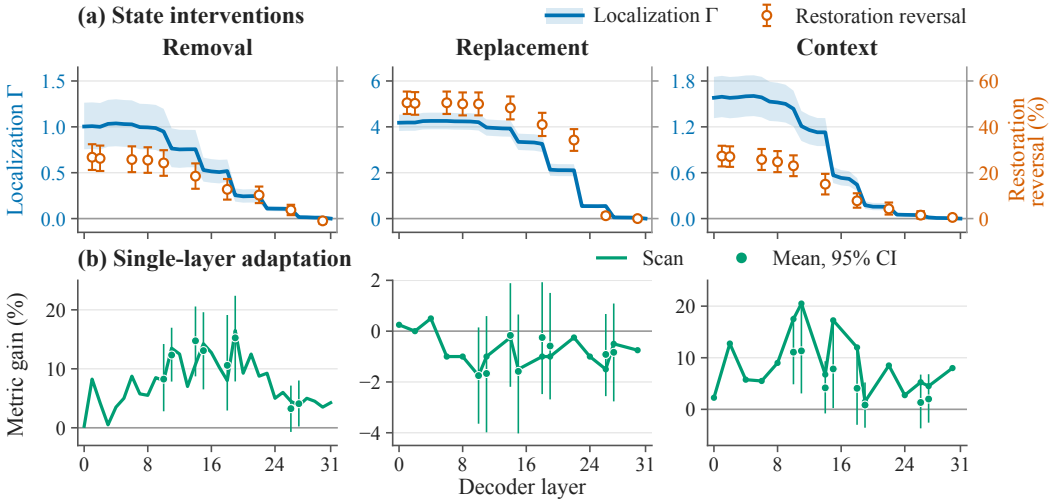


Figure 4: **State interventions and single-layer adaptation in Qwen3.5-9B.** (a) Localization Γ_a (blue; separate left scales) and restoration reversal (orange; shared right scale). (b) Metric gains (green), averaged equally over Color and Count. Lines connect measured layers; error bars indicate pointwise 95% intervals where available. Sampling and uncertainty estimates are described in Appendix B.2. Gain scales differ; layers are zero-indexed.

4.3 Layerwise Localization–Adaptation Correspondence

For each behavior, we adapt one decoder block at a time with LoRA (Hu et al., 2022), holding the specialist training protocol fixed within each backbone and approximately matching parameter counts across layers (Appendix B.2). We use the shared training and evaluation data of Section 3.2.

Training objectives. The same specialist objectives apply to single-layer and fixed-budget adaptation. With CE_θ denoting mean teacher-forced answer-token cross-entropy and x_i^0 the original input, they are

$$\begin{aligned} \mathcal{L}_{\text{removal}}(i) &= \mathcal{L}_{\text{wCE}}(i) + 0.2\mathcal{L}_{\text{rank}}(i), \\ \mathcal{L}_{\text{replacement}}(i) &= \mathcal{L}_{\text{pair}}(i) + 0.5\mathcal{L}_{\text{sign}}(i) \\ &\quad + 0.5\mathcal{L}_{\text{keep}}(i) + 0.25\mathcal{L}_{\text{leak}}(i), \\ \mathcal{L}_{\text{context}}(i) &= \mathrm{CE}_\theta(x_i^0, y_i^*) + 0.5\,\mathrm{SymKL}_i(O, S). \end{aligned} \tag{3}$$

Removal combines view-weighted CE with a ranking term that favors the reference answer on the original image over a wrong-image donor. Color evidence-removal and no-image views target ab-

stention; Count-removal views target the edited count, including zero. Replacement rewards a larger treatment–control difference in new-versus-original answer preference changes from frozen Base (pair), favors y_i' on treatment and y_i^* on control (sign), retains y_i^* on original and control views (keep), and penalizes original-view preference drift (leak). Context supervises the original answer and aligns full-vocabulary distributions for original (O) and background-edited (S) inputs through symmetric KL along the same teacher-forced original-answer sequence. Appendix C.1 gives the subterms, view weights, and target policy.

We measure adaptation gain $G_a(M)$ as the change from Base in the matched behavioral effect Δ_a, reversing the sign for Context. Positive gains therefore follow each metric's preferred direction (Appendix B.2).

Removal gains peak in middle layers, where localization is already declining, while Replacement changes remain small (Figure 4b).

Figure 5 directly compares localization strength with single-layer metric gains: Removal and Replacement show little association, while Context shows a positive association. The additional InternVL3.5-8B and LLaVA-1.5-7B scans show different adaptation profiles despite similar early-layer localization (Appendix B.2, Figure B.1c). In several of these cases the localization–gain correlations also closely track depth (Table B.1), and favorable Context gains accompany reduced original-image accuracy.

Single-layer gains need not add when several layers are trained jointly: updates at one layer can change the inputs to later layers. Section 5 directly tests whether localization-ranked layer sets improve adaptation under a fixed total parameter budget.

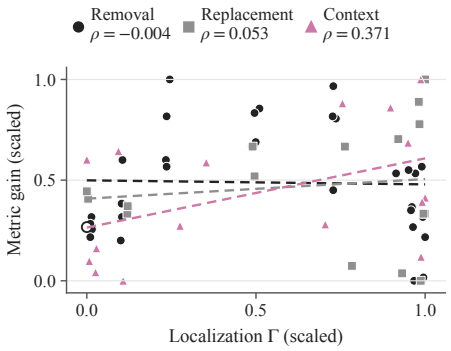


Figure 5: **Localization versus adaptation gain.** Points denote layers; axes are min–max scaled per behavior. Dashed lines: least-squares fits; ρ: Spearman correlation. Original scales and markers: Appendix B.2.

Finding 1. Localization strength need not track single-layer adaptation gains.

Localization peaks in early layers, while Removal gains peak farther into the decoder and favorable Context gains can come with reduced original-image accuracy; the correspondence varies by behavior and model. Patching probes the influence of states that already exist, whereas adaptation changes how they are produced, so a layer's influence on an answer and its value as an adaptation site have to be measured separately.

5 Evaluating Localization-Guided Adaptation

This section asks whether the localization ranking is useful as a placement rule, and what behavior-aligned supervision achieves at full layer coverage.

5.1 Fixed-Budget Adaptation Protocol

The fixed-budget and Joint experiments in this section use Qwen3.5-9B. For each behavior, we compare three ways of placing a fixed trainable-parameter budget under the specialist objectives in Equation 3. Top-k adapts the k layers with the largest $\Gamma_a(\ell)$, and Random-k adapts k randomly drawn layers using sampling stratified by block type, under the same parameter budget. We compare both placements at $k \in \{4, 8, 16\}$, with Random-k averaging five sampled layer sets at each width. Full adapts all 32 decoder layers and serves as the $k = 32$ endpoint that both placements share. All placements use approximately 43.3M trainable parameters, with LoRA ranks 128, 64, and 32 for $k = 4, 8, 16$ and rank 16 for Full (Figure 3b). Removal Top-16 uses $\{0, \ldots, 14, 18\}$. Throughout,

ranking refers to the localization order and *rank* to the LoRA rank. Within each behavior, the training objective and schedule are fixed across placements.

We evaluate on the shared dataset defined in Section 3.2. Because the same items both define the ranking and score adaptation, the comparison favors Top-k: any selection advantage should be easiest to detect here, so its absence is a conservative reading.

5.2 Layer Selection and Budget Allocation

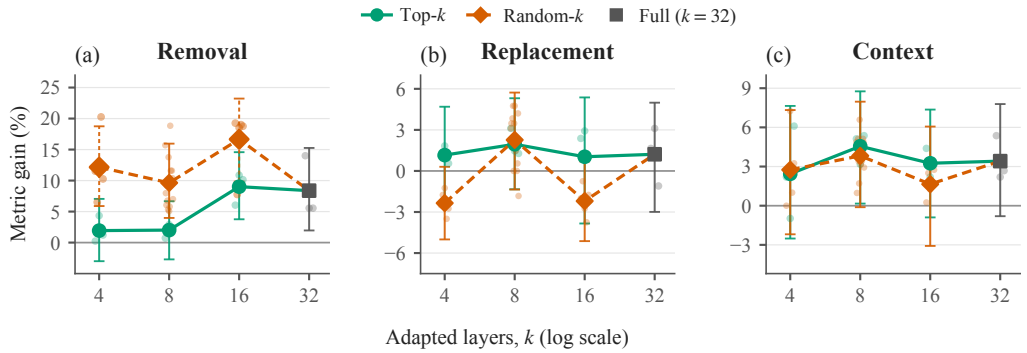


Figure 6: **Ranked placement offers no consistent advantage, and underperforms on Removal.** Qwen3.5-9B specialist metric gains at a matched total parameter budget. Circles show Top-k, diamonds show Random-k, and squares mark the shared Full endpoint ($k = 32$). Lines connect the tested widths. Panels use independent vertical scales. Error bars are 95% bootstrap intervals over image groups, seeds, and sampled layer sets.

For Removal, Top-k yields lower mean gains than Random-k at every tested width ($k = 4, 8, 16$; Figure 6). Expanding Top-k coverage to 16 layers brings its mean close to Full, but still below Random-16. This absence of a same-width advantage is consistent with the single-layer mismatch in Section 4.3: localization ranks early layers highest, while Removal gains peak farther into the decoder.

For Replacement, Top-k yields higher mean gains than Random-k at $k = 4, 16$ but a slightly lower mean at $k = 8$, so the ordering changes with width. For Context, the two placements differ little at every width, and gains do not increase monotonically with coverage.

Finding 2. Localization-ranked placement offers no consistent advantage.
At a matched total parameter budget, Top-k shows no consistent mean advantage over Random-k at matched widths; broader Top-k coverage can raise Removal gains without establishing an advantage for the ranking itself. As a placement rule, the signal therefore shows the localization–adaptation gap of Section 1, on this backbone and for the tested objectives, widths, and layer sets.

5.3 Joint Adaptation and Transfer

The placement comparisons concern where to adapt, not whether these behaviors can be improved at all. We therefore check that the supervision itself is effective: were it not, the absence of a placement advantage would be uninformative. We also ask which gains coexist in a single adapter and how they transfer beyond the training distribution. Joint uses the same training and evaluation data and trains a Full-scope adapter with a separate multi-view CE objective. For record i, let $\overline{\mathcal{L}}_{i,b}$ be the mean CE over its available views in category b, and let $\mathcal{B}_i$ contain its nonempty categories. We optimize

$$\mathcal{L}_{\text{joint}} = \frac{1}{2} \sum_{t \in \{\text{Color,Count}\}} \mathbb{E}_{i|t} \left[\frac{\sum_{b \in \mathcal{B}_i} w_b \overline{\mathcal{L}}_{i,b}}{\sum_{b \in \mathcal{B}_i} w_b} \right]. \tag{4}$$

Views are averaged within category before weighting, normalization uses only categories present in each record, and Color and Count receive equal weight. Joint contains only these CE terms; the specialist regularizers do not enter it. Category weights and view assignments are given in Appendix C.1.

The objective explicitly supervises abstention on unanswerable training views, including Color evidence-removal and no-image inputs. Count-removal views retain the edited count as the target, including zero after full removal. This supervision encourages evidence-dependent answering: providing an answer when the visual evidence supports one and abstaining when it is insufficient.

Joint improves the aggregate metric for each grounding behavior, with the largest gain on Removal (Table 1). Original-image accuracy stays within 1% of Base.

Table 1: **Base versus Joint on Qwen3.5-9B.** IID: matched effects $100\Delta_a$ (%), averaged equally over Color and Count. External (%): VizWiz-Abst. and VizWiz-Acc. are abstention on unanswerable and accuracy on answerable VizWiz questions (Removal); CV-VQA$_{\geq 1}$ (edited images, Replacement) and TallyQA$_{\geq 1}$ (natural questions) are counting accuracy on items whose answer is at least one; IV-VQA is the answer-flip rate after an irrelevant object is removed (Context).

	IID			External				
Model	Removal ↑	Replacement ↑	Context ↓	VizWiz-Abst. ↑	VizWiz-Acc. ↑	CV-VQA$_{\geq 1}$ ↑	TallyQA$_{\geq 1}$ ↑	IV-VQA ↓
Base	26.2	50.5	27.0	1.6	55.1	77.8	81.7	3.1
Joint	37.8	52.2	25.3	27.5	54.8	80.3	83.2	3.7

We evaluate whether this supervised abstention behavior transfers to unanswerable questions in VizWiz (Gurari et al., 2018). Correct abstention rises from 1.6% to 27.5%, with little change in accuracy on answerable questions (Table 1). Transfer to counting is mixed: accuracy improves on the positive-count subsets of CV-VQA (Agarwal et al., 2020) and TallyQA (Acharya et al., 2019), but falls by 21.0% and 15.2% on their zero-answer subsets, respectively. Overall TallyQA accuracy decreases slightly, and IV-VQA consistency also worsens.

Finding 3. Joint supervision improves grounding metrics, with selective transfer.
The dataset and its behavior-aligned supervision support adaptation beyond the training examples. This result concerns the effectiveness and transfer of the supervision; it does not establish a preferred layer placement or uniformly better performance across tasks.

6 Conclusion

We investigated whether functional localization can guide adaptation for visual grounding. Localization and restoration consistently highlight early layers, yet similar localization profiles can accompany different adaptation gains and accuracy costs. Under a matched parameter budget, localization-ranked placement offers no consistent advantage over random placement. Joint supervision improves the evaluated grounding behaviors and supports selective transfer beyond the training distribution, including abstention on unanswerable questions. For improving visual grounding, placement decisions should be evaluated through their effects on evidence use and answer accuracy. Localization tells us where visual evidence acts on the answer, but not, by itself, where to adapt.

Our study covers controlled color/count VQA with layer-level image-token patching and LoRA adaptation. Fixed-budget and Joint experiments use one backbone; broader validation across architectures, grounding tasks, and adaptation methods remains open. Mixed external transfer also warrants further study. A key direction is to identify layers whose updates improve visual evidence use while preserving answer accuracy. This could combine localization with learning signals from gradients or short adaptation runs, and account for interactions among jointly adapted layers. Matched-budget comparisons could then test whether these criteria yield better placement rules.

AI USE STATEMENT

We used AI tools to assist with writing and language polishing, including improvements to grammar, clarity, and readability, and to support literature retrieval and the discovery of relevant related work. The authors take full responsibility for the final manuscript, including the accuracy and integrity of its text, claims, and references.

ETHICS STATEMENT

This work studies visual grounding in existing vision-language models using publicly available datasets. To the best of our knowledge, this study raises no specific ethical concerns.

REPRODUCIBILITY STATEMENT

Code for reproducing our experiments is available in the anonymous repository:

`https://anonymous.4open.science/r/Does-Localization-Inform-Adaptation-for-VisualGrounding-in-Vision-Language-Models-CF5B/`

Data construction and evaluation are detailed in Appendix A; localization and adaptation protocols, training objectives, and optimization settings are provided in Appendices B–C.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL `https://papers.nips.cc/paper_files/paper/2024/hash/0dfe31d6e703e138d46a7d2fced38b7c-Abstract-Conference.html`.

Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Qizhou Chen, Taolin Zhang, Chengyu Wang, Xiaofeng He, Dakan Wang, and Tingting Liu. Attribution analysis meets model editing: Advancing knowledge correction in vision language models with VisEdit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL `https://ojs.aaai.org/index.php/AAAI/article/view/32215`.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL `https://aclanthology.org/2021.tacl-1.10/`.

Michal Golovanevsky, William Rudman, Michael Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025a.

Michal Golovanevsky, William Rudman, Vedant Palit, Ritambhara Singh, and Carsten Eickhoff. What do VLMs NOTICE? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025b. URL https://aclanthology.org/2025.naacl-long.571/.

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 (CVPR)*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. SwapMix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3927bbdcf0e8d1fa8aa23c26f358a281-Abstract-Conference.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

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*, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/40431.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ye-Bin Moon, Hyeon-Woo Nam, Wonseok Choi, and Tae-Hyun Oh. BEAF: Observing BEfore-AFter changes to evaluate hallucination in vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv preprint arXiv:2604.09364*, 2026.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2856–2861, 2023. URL https://openaccess.thecvf.com/content/ICCV2023W/CLVL/html/Palit_Towards_Vision-Language_Mechanistic_Interpretability_A_Causal_Tracing_Tool_for_BLIP_ICCVW_2023_paper.html.

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

Daniel Reich, Felix Putze, and Tanja Schultz. Measuring faithful and plausible visual grounding in VQA. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision function layer in multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 27510–27537, 2025. doi: 10.52202/085713-0926. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/27c4e15d9af120d7fef04432c7db577f-Abstract-Conference.html.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://aclanthology.org/2024.emnlp-main.460/.

Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://aclanthology.org/D19-1002/.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

Zhe Xu, Zhicai Wang, Junkang Wu, Jinda Lu, and Xiang Wang. Causal-HalBench: Uncovering LVLMs object hallucinations through causal intervention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 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 (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8001c3568152d134d821cd46d4d84768-Abstract-Conference.html.

Fangcong Yin, Xi Ye, and Greg Durrett. LoFiT: Localized fine-tuning on LLM representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/122ea6470232ee5e79a2649243348005-Abstract-Conference.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Yang Zhou, Pengfei Cao, Yubo Chen, Qingbin Liu, Dianbo Sui, Xi Chen, Kang Liu, and Jun Zhao. M2Edit: Locate and edit multi-granularity knowledge in multimodal large language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://aclanthology.org/2025.emnlp-main.1478/.

A Data and Evaluation

A.1 Construction and splits

Each original image–question pair anchors treatment and operation-matched control views. Every view and construction variant retains the original image's group identity. An item enters a behavior's analysis only when both arms pass checks defined without consulting model responses; Replacement also requires a validated counterfactual answer y'_i. Eligibility is determined separately for each behavior, and incomplete pairs are excluded from its matched comparison.

Removal treatment and control share the source image, fill implementation, codec, and edited area; the fill is placed in the evidence region or an area-matched non-evidence region. Replacement uses the same validated attribute change and compositing settings at an evidence or control location. For Context, external and identity background donors pass through the same resize, feather, composite, and codec chain. The identity-donor arm serves as the reconstruction control; a separately re-encoded original provides a reference for the codec alone.

Evidence regions and control placements must preserve the declared target content of their comparison. Two reviewers independently assess Replacement candidates for the intended attribute change and visual acceptability, with disagreements adjudicated under a fixed protocol. Automatic color, shape, or naturalness scores may prioritize review but do not determine eligibility. Removal populations are kept separate by fill and task.

All in-domain experiments follow the shared data protocol in Section 3.2. All views of an original image remain in its split. Each evaluation item is an image–question pair with its treatment and matched-control views; views are not counted as separate evaluation items. The 400 items per behavior span 400 image groups and comprise 200 object-color and 200 object-count questions, which receive equal weight in every behavior-level average.

A.2 Measurement and baseline diagnosis

An answer normalizer ν defines equivalence as $\mathrm{eq}(a, b) = \mathbb{1}[\nu(a) = \nu(b)]$. The target-hit reader gr compares a decoded response with the validated answer y'_i, and an abstention indicator ab matches fixed refusal expressions in the normalized response. Scoring rules are fixed across the two arms of every comparison.

For Removal and Context, the arm response indicates a change from the model's own decoded answer on the original image. For Replacement, it indicates a match to y'_i, using the same target in treatment and control. Point estimates average variants within each condition before taking within-item treatment–control differences and averaging across items (Eq. 1). Uncertainty resamples original-image groups with replacement, retaining every question, condition, and variant of a selected group. Removal intervals are computed separately for each fill–task stratum.

Table A.1: Diagnostic matched effects $100\Delta_a(M)$ in %, with 95% image-group bootstrap intervals, as plotted in Figure 2. All five models use the shared evaluation dataset (400 items per behavior; 200 Color and 200 Count). Effects are compared within a behavior.

Model	Removal	Replacement	Context
Qwen3.5-9B	26.25 [20.75, 31.75]	50.50 [46.30, 54.75]	27.00 [22.59, 31.36]
Qwen3.5-4B	25.50 [19.73, 31.25]	47.50 [43.00, 52.00]	24.00 [19.50, 28.50]
InternVL3.5-8B	40.50 [35.09, 46.00]	45.75 [41.25, 50.40]	33.75 [29.05, 38.50]
LLaVA-1.5-7B	27.75 [22.44, 33.08]	44.25 [39.53, 48.81]	32.50 [28.00, 37.12]
LLaVA-1.5-13B	35.50 [30.32, 40.53]	42.50 [37.77, 47.25]	27.25 [22.75, 31.85]

B Localization and Layerwise Adaptation

B.1 Patching and restoration

For each behavior, patching requires the treatment, operation-matched control, both donor–recipient directions, and aligned active image tokens. Filtering retains the same image groups for all four arm–direction cells. At decoder-block output ℓ, the donor states replace the recipient states at every aligned active image-token position; text tokens and non-active image positions remain unchanged, and the remaining network is recomputed.

For arm $u \in \{\mathrm{tr}, \mathrm{ct}\}$, edited-to-original patching uses $x^{\mathrm{don}}_{i,a,u,d} = x_i(c^u_a)$ and $x^{\mathrm{rec}}_{i,a,u,d} = x^0_i$; original-to-edited patching reverses these roles. The readout target is the original answer for Removal and Context and the validated counterfactual answer for Replacement. Under each model's prompt template, $t_{i,a}$ is the first non-whitespace token of the answer continuation. Any preceding whitespace-only token is appended to all paired inputs; x below includes this prefix when present. The target-token margin and raw patching effect are

$$\phi_{i,a}(x) = \mathrm{logit}_{M_0,t_{i,a}}(x) - \max_{v \neq t_{i,a}} \mathrm{logit}_{M_0,v}(x), \tag{5}$$

$$E^d_{i,a,u}(\ell) = \phi_{i,a}\big(x^{\mathrm{rec}}_{i,a,u,d} \leftarrow_\ell x^{\mathrm{don}}_{i,a,u,d}\big) - \phi_{i,a}\big(x^{\mathrm{rec}}_{i,a,u,d}\big)\,. \tag{6}$$

Within each direction, we subtract the control effect from the treatment effect. Equation 2 subtracts the original-to-edited contrast from the edited-to-original contrast, divides by two, and applies sign -1, $+1$, and -1 for Removal, Replacement, and Context, respectively. Averaging over the eligible items gives $\Gamma_a(\ell)$; uncertainty resamples whole image groups. All four cells are required for this average.

The Qwen3.5-9B profiles remain on an early plateau through layer 10 and then fall at full-attention blocks. The zero at layer 31 is structural: after the final block output, no subsequent cross-token computation carries a patched image-token state to the answer position.

Five-backbone localization. Across the five backbones, pooled localization is strongest at shallow sampled depths and declines toward deeper layers (Figure B.1a). The pooled LLaVA profiles decline earlier in relative depth than the Qwen and InternVL profiles. We use the same estimator on 400 items per behavior, equally split between Color and Count, at ten approximately matched relative depths. For a decoder with N blocks, the zero-based layers are $\lfloor jN/32 + 1/2 \rfloor$ for $j \in \{1, 2, 6, 8, 10, 14, 18, 22, 26, 30\}$. Qwen3.5-9B uses the corresponding subset of its complete layer scan. Means are question-weighted, with pointwise intervals from 10,000 image-group bootstrap draws. Complete task profiles are provided in the reproduction companion.

For this localization comparison, images have a maximum side length of 512 pixels before model-specific processing. InternVL uses one 448×448 image; LLaVA uses CLIP-mean square padding before 336×336 processing. The whitespace-prefix readout applies to LLaVA Count items. Exact grids and processing details are retained with the numerical summaries in the reproduction companion.

Restoration. The original image is the common donor for the treatment and matched-control recipients. Each item requires all three inputs and their aligned active image tokens. At each sampled depth, the original-image block outputs replace those tokens in both recipients; the restored

treatment–control effect minus the unpatched effect gives $R_a(\ell)$. With r^{rest} the restored score and $\mathcal{P}_a^{\mathrm{rest}}$ the eligible items, we define

$$R_a(\ell) = \mathbb{E}_{i \in \mathcal{P}_a^{\mathrm{rest}}} \Big[\delta_{i,a}[r^{\mathrm{rest}}(\ell)] - \delta_{i,a}[r^{M_0}] \Big]. \tag{7}$$

An original-to-original patch at the same depth and token scope is the exact sham and reproduces the unpatched output under deterministic evaluation.

For Qwen3.5-9B in Figure 4a, restoration samples the layers $\{1, 2, 6, 8, 10, 14, 18, 22, 26, 30\}$ and uses the shared behavior-specific evaluation sets in Appendix A.1. Plots report reversal $-100R_a$ in %, with the interval endpoints reversed by the sign transformation. The five-model profiles in Figure B.1b use the shared evaluation dataset: 400 items per behavior, comprising 200 Color and 200 Count items. These profiles retain the same absolute ten-layer grid for every model; the combined figure maps each observed layer to $\ell/(N-1)$. Both readouts show broad early-to-late attenuation, with transition depth varying across behaviors and models.

B.2 Layerwise adaptation

For Qwen3.5-9B, within each behavior the single-layer scan changes only the adapted decoder block. It shares the training records, objective, target labels, sampling, optimizer, and checkpoint rule with the fixed-budget experiments (Appendix C.1); training settings are summarized in Table C.2. Rank 16 in linear-attention blocks and rank 17 in full-attention blocks give 1,377,280 and 1,357,824 trainable parameters, respectively, a 1.4% difference.

Metric gain. With $\Delta_a(M)$ the matched behavioral effect defined in Section 3.3, the metric gain from Base is

$$G_a(M) = q_a \big[\Delta_a(M) - \Delta_a(M_0) \big], \qquad q_{\mathrm{Removal}} = q_{\mathrm{Replacement}} = 1, \quad q_{\mathrm{Context}} = -1, \tag{8}$$

so that positive G_a denotes a change in the metric's preferred direction. The two sign conventions have different references: s_a follows the direction in which the edit should move the answer margin, while q_a follows the metric's preferred direction. The signs of Γ_a and G_a are therefore not directly comparable.

Localization–gain correspondence. The analysis pairs $\Gamma_a(\ell)$ with the metric gain averaged equally over Color and Count. Spearman coefficients use average ranks after rounding to resolve numerical ties. Table B.1 compares localization with negative layer index as a depth baseline for Qwen3.5-9B, InternVL3.5-8B, and LLaVA-1.5-7B. Figure 5 overlays the Qwen3.5-9B observations after min–max scaling each coordinate within each behavior. Figure B.2 retains the original scales and shows the non-monotonic Removal profile.

Table B.1: Layerwise Spearman correlations, with n adapted layers. The three backbones retain their respective scan grids and training settings. Coefficients use average ranks to resolve ties. Context gains require the accuracy checks discussed below.

Model	Behavior	n	$\rho(\Gamma_a, G_a)$	$\rho(-\ell, G_a)$	$\rho(\Gamma_a, -\ell)$
Qwen3.5-9B	Removal	32	-0.004	-0.018	$+0.980$
	Replacement	16	$+0.053$	$+0.233$	$+0.924$
	Context	16	$+0.371$	$+0.388$	$+0.976$
InternVL3.5-8B	Removal	10	-0.550	-0.661	$+0.952$
	Replacement	10	-0.241	-0.241	$+1.000$
	Context	10	$+0.195$	$+0.195$	$+1.000$
LLaVA-1.5-7B	Removal	10	$+0.445$	$+0.415$	$+0.988$
	Replacement	10	$+0.433$	$+0.335$	$+0.988$
	Context	10	$+0.650$	$+0.620$	$+0.976$

Interpreting metric gains. Qwen3.5-9B's Removal response peaks in middle layers, whereas Replacement changes remain small. For Context, Color gains are modest, while Count shows larger but more variable gains. Original-image accuracy and answer-distribution diagnostics are retained in the reproduction companion, together with the detailed task profiles.

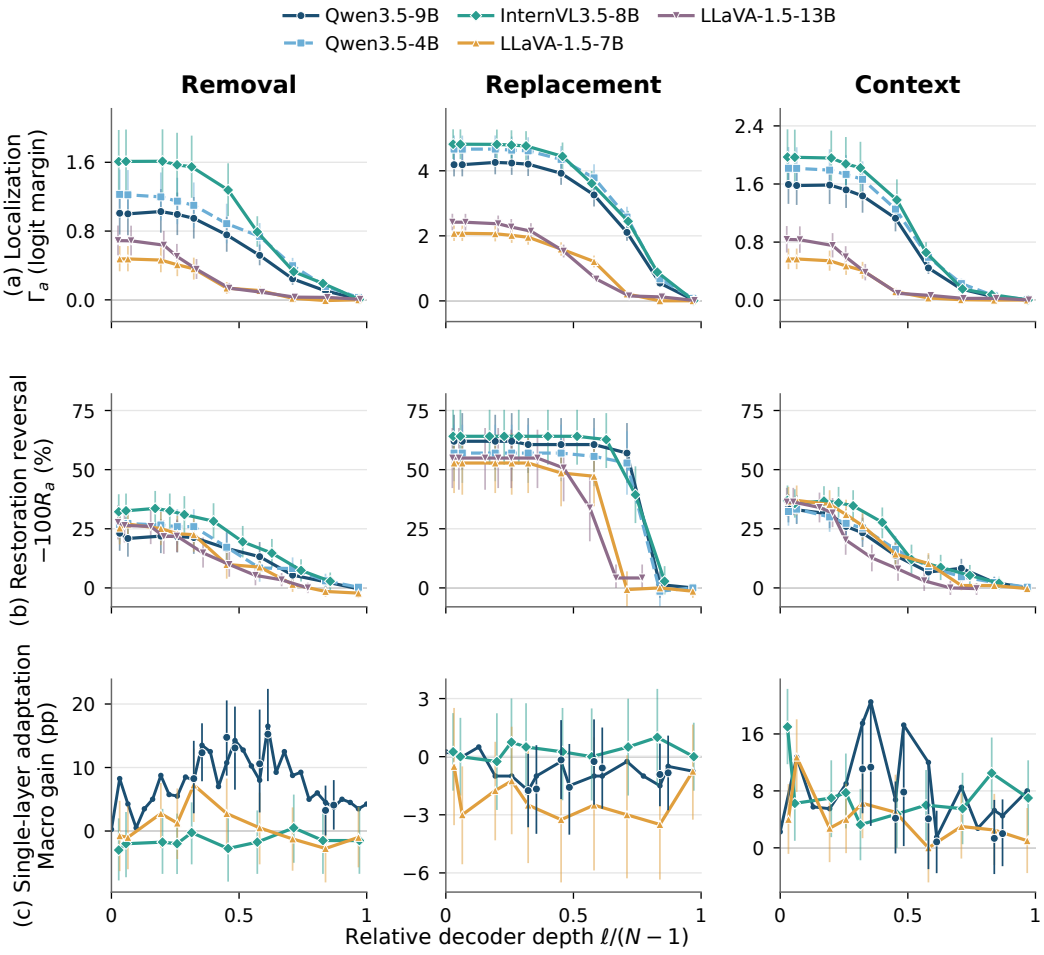


Figure B.1: **Localization, restoration, and single-layer adaptation.** Columns show Removal, Replacement, and Context. (a) Pooled image-token localization Γ_a on the original logit-margin scale. (b) Restoration reversal $-100R_a$ in % across the same five models. (c) Single-layer macro gains (%) for Qwen3.5-9B, InternVL3.5-8B, and LLaVA-1.5-7B; larger Qwen markers show repeated-layer means. Colors and markers identify the same models across panels; lines connect the observed layers at relative decoder depth $\ell/(N-1)$. Error bars are pointwise 95% intervals. Qwen adaptation intervals include repeated training measurements; the other two adaptation scans use image-group intervals conditional on the trained adapter. The state interventions use the samples and layer grids specified above; the adaptation row uses the protocols in Appendix C.2.

Cross-backbone comparison. Figure B.1c places all three backbones on the same relative-depth axis. LLaVA's Removal gain peaks at layer 10 ($+7.25\%$, 95% interval $[+1.42, +12.98]$), resembling Qwen's middle-layer response, but with an 8.5% loss in macro original-image accuracy. InternVL's Removal intervals all include zero. Replacement gains remain small in Qwen and InternVL and are negative at every sampled LLaVA layer. Context gains in both additional backbones accompany answer concentration and lower original-image accuracy: macro accuracy falls by 2.5–19.5% in InternVL and 1.75–17.0% in LLaVA. These metric gains therefore need to be interpreted alongside accuracy and answer-distribution changes.

Table B.1 compares the corresponding layer rankings. InternVL's Replacement and Context localization ranks are identical to negative depth; LLaVA's localization–gain correlations are close to their depth-baseline counterparts. Similar early-layer localization thus coexists with different adaptation profiles and accuracy costs. These scans do not establish localization strength as a reliable beneficial placement rule beyond depth, or determine the benefit of jointly adapting several layers.

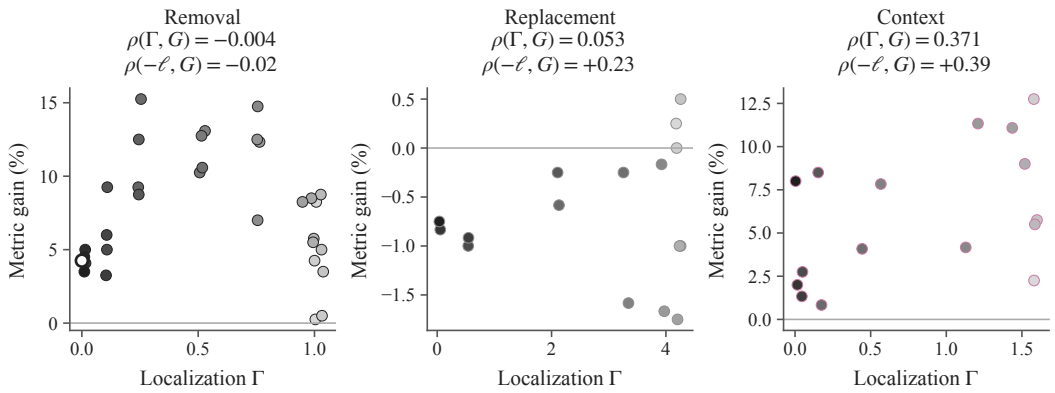


Figure B.2: **Qwen3.5-9B layerwise localization and single-layer adaptation gains.** Each point is a measured layer, shaded by depth; the open marker identifies the final Removal layer.

C Training Details

C.1 Views, targets, and objectives

We expand the specialist objectives in Equation 3 and the separate Joint objective in Equation 4. Each specialist objective applies to both the single-layer and fixed-budget experiments. For a supervised answer sequence y, the mean teacher-forced answer-token log probability and cross-entropy are

$$s_\theta(x, y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log p_\theta(y_t \mid x, y_{<t}), \qquad \mathrm{CE}_\theta(x, y) = -s_\theta(x, y).$$

Here $x_i^0 = x_i^O$ denotes the original image–question input, and $[z]_+ = \max(0, z)$.

Table C.1 gives each view's target and loss role. The fixed abstention target is u, and c_i^{edit} is the declared count after editing, including zero for full removal. These label rules are shared across placements. The reconstruction control `spurious_recon` is evaluation-only for all objectives, including Joint. Synthetic partial Color-removal views that preserve the answer use the removal-control target and weight. A record's original, treatment, and matched-control views share one original-image identity; the ranking donor comes from another record.

Removal. The five-condition set $\mathcal{C}^{\mathrm{train}}$ contains original, context-edited, removal-control, evidence-removal, and no-image views, with weights 1.0, 0.4, 0.2, 0.2, and 0.02, respectively, so $W = 1.82$. The training policy rebinds cross-sample donor $j(i)$ deterministically for each seed and epoch, with donor input $x_i^\times = (I_{j(i)}, Q_i)$. Using the targets in Table C.1, the loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{wCE}}(i) &= W^{-1} \sum_{c \in \mathcal{C}^{\mathrm{train}}} w_c \, \mathrm{CE}_\theta\big(x_i(c), y_{i,c}^{\mathrm{train}}\big), \\ \mathcal{L}_{\mathrm{rank}}(i) &= \big[m - \big\{s_\theta(x_i^0, y_i^*) - s_\theta(x_i^\times, y_i^*)\big\}\big]_+, \\ \mathcal{L}_{\mathrm{removal}}(i) &= \mathcal{L}_{\mathrm{wCE}}(i) + \lambda_{\mathrm{rank}} \mathcal{L}_{\mathrm{rank}}(i). \end{aligned} \tag{9}$$

Here $m = 0.5$ and $\lambda_{\mathrm{rank}} = 0.2$; the ranking coefficient is outside the CE normalization by W.

Table C.1: Training views, targets, and loss roles. The same specialist view policy applies in Sections 4.3 and 5.1; Joint views use their declared category assignments.

Objective	View	Train?	Target or reference; loss role
Removal	Original, context edit, removal control	Yes	y_i^*; weighted CE. The original also anchors donor ranking.
	Evidence removal	Yes	Color: u; Count: c_i^{edit}; weighted CE
	No image	Yes	u; weighted CE
	Wrong-image donor	Yes	Score y_i^* with the original question; donor ranking
Replacement	Original O	Yes	Candidates y_i^*, y_i' and frozen Base scores; keep and leak
	Replacement T	Yes	Candidates y_i^*, y_i' and frozen Base scores; pair and sign
	`null_replace` C	Yes	Candidates y_i^*, y_i' and frozen Base scores; pair, sign, and keep
Context	Original O	Yes	y_i^*; CE and symmetric KL
	Background edit S	Yes	Distributions along the same y_i^* sequence; symmetric KL
	`spurious_recon`	No	Matched evaluation control
Joint	Available direct-label views	Yes	Original, answer-preserving edits, and matched controls: y_i^*; directed replacement: y_i'; evidence removal and no image: the labels above. Weighted multi-view CE.

Replacement. The specialist uses directed replacements with a unique verified answer y_i'. Write O, T, and C for the original, replacement, and `null_replace` views. For each view X, define the edited-answer preference and its change from the frozen Base model as

$$r_X^\theta = s_\theta(x_i^X, y_i') - s_\theta(x_i^X, y_i^*), \qquad \Delta r_X = r_X^\theta - r_X^0.$$

With $\operatorname{softplus}(z) = \log(1 + \exp z)$, the objective is

$$\begin{aligned} \mathcal{L}_{\text{pair}}(i) &= \operatorname{softplus}(\ln 2 - (\Delta r_T - \Delta r_C)), \\ \mathcal{L}_{\text{sign}}(i) &= \operatorname{softplus}(-r_T^\theta) + \operatorname{softplus}(r_C^\theta), \\ \mathcal{L}_{\text{keep}}(i) &= -s_\theta(x_i^O, y_i^*) - s_\theta(x_i^C, y_i^*), \\ \mathcal{L}_{\text{leak}}(i) &= (\Delta r_O)^2, \\ \mathcal{L}_{\text{replacement}}(i) &= \mathcal{L}_{\text{pair}}(i) + 0.5\mathcal{L}_{\text{sign}}(i) \\ &\quad + 0.5\mathcal{L}_{\text{keep}}(i) + 0.25\mathcal{L}_{\text{leak}}(i). \end{aligned} \tag{10}$$

The pair term compares the treatment and control changes with margin $\ln 2$; both sign margins are zero. The keep term retains the original answer on O and C, and the leak term penalizes preference drift on O.

Context. Let p_{it}^O and p_{it}^S be the model's full-vocabulary distributions at answer position t for the original and background-edited views, teacher-forced on the same original-answer sequence of length T_i. The symmetric KL averages both directions over aligned answer positions:

$$\begin{aligned} \operatorname{SymKL}_i(O, S) &= \frac{1}{2T_i} \sum_{t=1}^{T_i} \left[\operatorname{KL}(p_{it}^O \| p_{it}^S) + \operatorname{KL}(p_{it}^S \| p_{it}^O) \right], \\ \mathcal{L}_{\text{context}}(i) &= \operatorname{CE}_\theta(x_i^0, y_i^*) + 0.5 \operatorname{SymKL}_i(O, S). \end{aligned} \tag{11}$$

Joint. For record i, let $\mathcal{V}_{i,b}$ contain its available views assigned to category b, and let $\mathcal{B}_i$ contain only categories with at least one view. The category weights for original, removal, replacement, context, and no image are 1, 1, 1, 0.5, and 0.02, respectively. First average CE within each category, then normalize the weighted sum over the categories present in that record:

$$\begin{aligned} \overline{\mathcal{L}}_{i,b} &= \frac{1}{|\mathcal{V}_{i,b}|} \sum_{v \in \mathcal{V}_{i,b}} \mathrm{CE}_\theta(x_i^v, y_{i,v}^{\mathrm{train}}), \\ \mathcal{L}_{\mathrm{views}}(i) &= \frac{\sum_{b \in \mathcal{B}_i} w_b \overline{\mathcal{L}}_{i,b}}{\sum_{b \in \mathcal{B}_i} w_b}, \\ \mathcal{L}_{\mathrm{joint}} &= \frac{1}{2} \sum_{t \in \{\mathrm{Color}, \mathrm{Count}\}} \mathbb{E}_{i|t}[\mathcal{L}_{\mathrm{views}}(i)] \,. \end{aligned} \tag{12}$$

The expectation follows the training sampling within each task; Color and Count each receive half the weight. Joint includes only these CE terms.

C.2 Optimization settings

For Qwen3.5-9B, within each behavior single-layer and fixed-budget experiments share the training records, objective, labels, sampling, optimizer, and checkpoint rule (Table C.2). Their layer sets and ranks follow the respective designs in Sections 4.3 and 5.1.

Table C.2: Qwen3.5-9B specialist training settings for the single-layer and fixed-budget experiments. Layer coverage and rank allocation follow each placement design.

Setting	Specification
Optimization	AdamW; constant learning rate 2×10^{-5}; weight decay 0; gradient clipping at 1.0
Batch construction	Four image groups per update. Context processes one group per forward pass; Replacement processes four and uses gradient checkpointing.
Checkpoint	Final update
LoRA	All attention and MLP projections; scaling $2r$; dropout 0.05
Sampling	Within a seed, data order is shared across placements; Removal also shares wrong-image donor bindings.

Joint trains a single adapter across all 32 decoder layers at rank 16, scaling 32, and dropout 0.05.

Pointwise 95% intervals for Qwen3.5-9B layerwise adaptation are reported at the eight repeated layers $\{10, 11, 14, 15, 18, 19, 26, 27\}$. The scan covers 32 Removal layers and 16 Replacement/Context layers; only the repeated layers receive intervals in Figure B.1c. Detailed experimental settings are provided in the reproduction companion.

Additional backbones. InternVL3.5-8B and LLaVA-1.5-7B adapt the ten layers used by their Locate grids, with rank 16 on the seven attention and MLP projections. They retain the shared objectives, labels, training inputs, sampling, and update counts. LLaVA uses AdamW at a constant 2×10^{-5} learning rate. InternVL uses a development-selected schedule: peak learning rate 1.2×10^{-4}, 16 linear warmup steps, and cosine decay to one tenth of the peak, without selecting on the evaluation set. Images follow the backbone-specific processing in Appendix B.1.

D DATASET EXAMPLES

Figure D.1 presents six examples to illustrate the dataset's questions, visual content, and edited views.

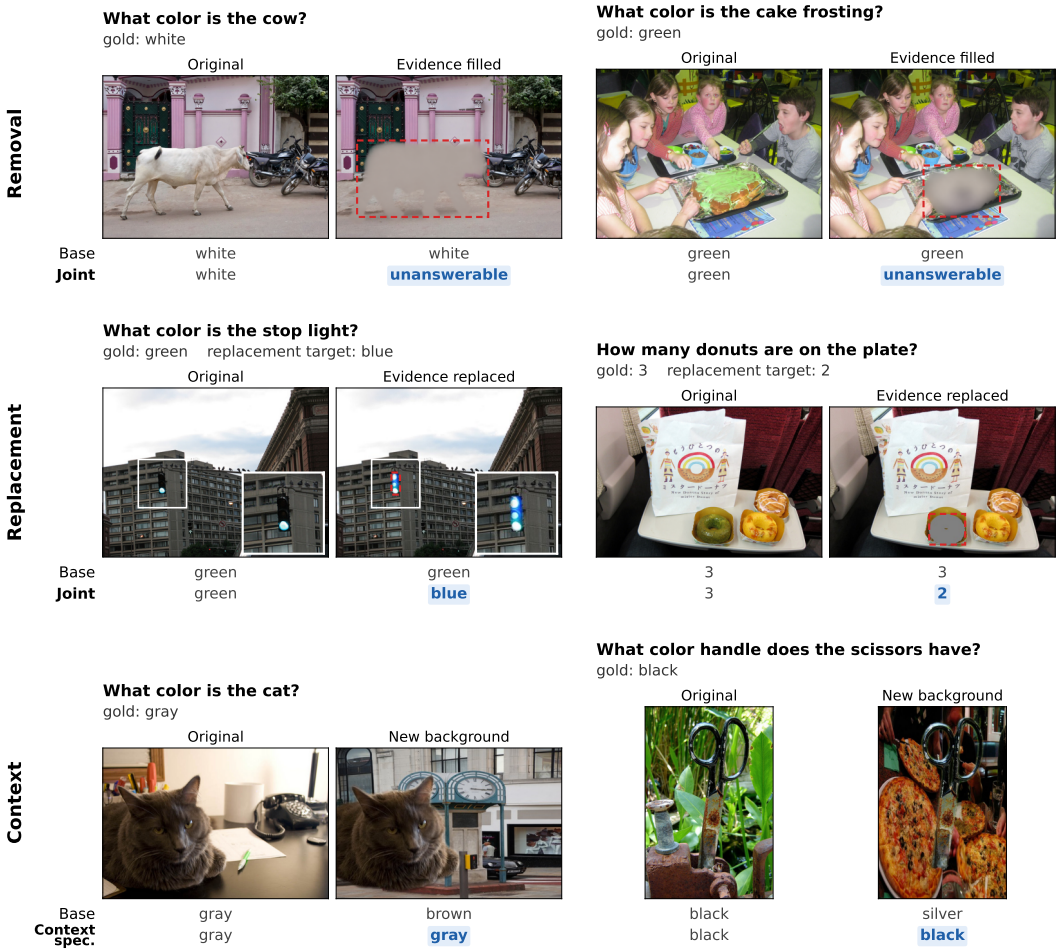


Figure D.1: **Examples of the three visual edits.** Removal obscures question-relevant evidence; Replacement changes the target color or count; Context changes the background while preserving the target. Each case pairs original and edited images. The responses underneath are Qwen3.5-9B outputs on these inputs, comparing Base with Joint in the top two rows and with the Context specialist in the bottom row.