DOES RL POST-TRAINING GROUND VISION-LANGUAGE MODELS? A CAUSAL MEASUREMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) post-training raises the benchmark accuracy of vision-language models (VLMs), but it is unclear whether the gains come from better use of visual evidence or from stronger language priors. Accuracy cannot settle this, since many benchmark items are answerable without the image, and neither can attention, which need not reflect how the image is used. We introduce CARVE, a programmatic benchmark of solver-certified counterfactual edits: each edit changes one piece of visual evidence, certifies the new answer, and records the image region it occupies. We test whether answers follow the edit, whether they follow it when a printed value stays the same but moves to a different element, and whether the edited region's hidden states alone redirect the answer. Across four Qwen-VL backbones trained with GRPO, VPPO, and PRPO, post-training improves edit following modestly on three backbones, yet models that follow a numeral edit keep their original answer after 10.9–39.5% of edits that only move a label, against at most 1.7% of numeral edits, in all 16 configurations; post-training does not remove this failure. The gain is not reliably carried by the evidence: the edited region reproduces 43% of the behavioral gain in our 3B models (95% CI $[18, 67]$) but 96% in our 7B models ($[66, 139]$), and the 3B pattern persists with a frozen vision tower, locating the change in the decoder. Answer-changing evidence stays at visual positions rather than moving into the question, and attention to the image rises after training even where the causal role of visual states falls. Grounding claims for VLMs should therefore be tested against evidence whose effect on the answer is known, not inferred from accuracy or attention.

1 INTRODUCTION

Vision-language models (VLMs) have become increasingly capable at tasks that require joint visual perception and language-based reasoning (Liu et al., 2023; Bai et al., 2025), but their reliability depends on using the visual evidence relevant to an answer, not only on producing it. Fu et al. (2025) show that task-relevant visual information can be well represented by a VLM's vision encoder yet poorly utilized by the full model, and that adapting the language model often improves vision-centric performance more than modifying the encoder: the bottleneck may lie less in what a model can see than in how it uses what it sees. Reinforcement learning (RL) post-training has become an effective approach for improving multimodal reasoning (Wang et al., 2025; Huang et al., 2026), yet higher accuracy need not imply better use of visual evidence: multimodal reasoning models are more prone to visual hallucinations than non-reasoning models, which Xu et al. (2025) attribute to less attention to visual inputs as reasoning chains grow longer. When RL post-training improves a VLM, does the model become better at reasoning *from the image*, or simply better at producing answers through language-side reasoning and learned priors?

Existing evaluations cannot effectively answer this. On standard multimodal benchmarks, much of the post-training gain is retained even without the image. For the two Qwen2.5-VL backbones we study, the gain without the image is 77–98% of the gain with the image on the ViRL39K held-out split and 84–96% on MMK12 (Section 6.1). Image perturbations (Huang et al., 2026) show that changing an image can alter the output, but not whether the model responds to the relevant evidence. Attention is similarly inconclusive: image tokens may receive little attention (Liu et al., 2024), while highly attended tokens can still be dispensable (Kang et al., 2025). Internal intervention methods (Basu et al., 2024; Neo et al., 2025; Salazar et al., 2026), meanwhile, do not pair a known answer

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.

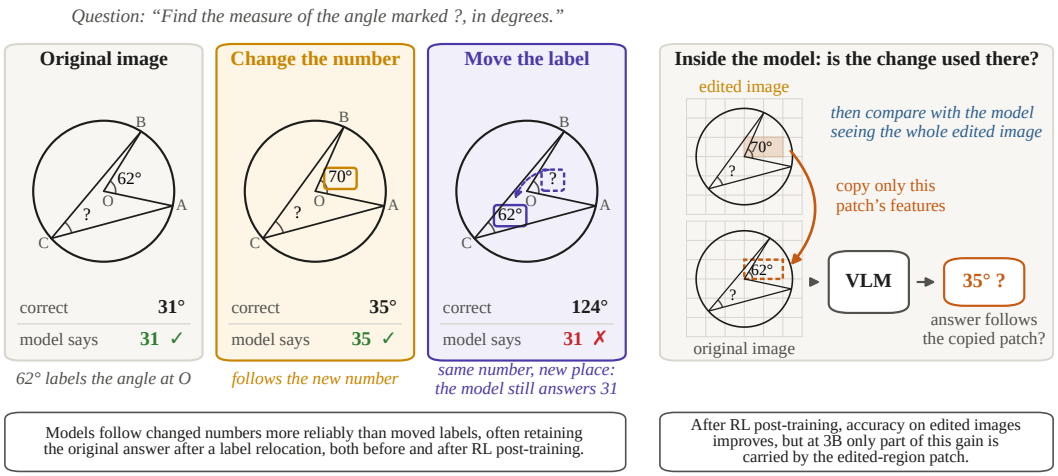


Figure 1: **Does the model answer from what the image shows?** *Left:* changing the printed number ($62° \to 70°$) changes the certified answer to $35°$; moving the same number to another angle changes it to $124°$, but the model still answers 31, the answer to the original image. *Right:* localized transfer copies only the edited patch's internal states from the edited-image run into the original-image run and checks whether the answer becomes the new one. Illustrative example; results in Section 6.

change with the localized image region responsible for it. A direct test therefore requires controlled changes for which we know what evidence changed, what the answer should become, and where the change lies in the image.

To enable this analysis, we introduce CARVE (Section 3), which generates scenes whose answers a symbolic solver certifies, together with localized edits that change either a printed value or the element it labels (Figure 1), and filters out items answerable without the image. We then evaluate visual evidence use at three levels: answerability with and without the task image, behavioral responses to certified edits, and causal responses to hidden-state interventions, for which we introduce causal evidence following (CEF). Behavioral evaluation asks whether the prediction changes to the certified new answer. CEF transfers hidden states from the edited region in the edited-image run into the original-image run and asks whether they alone redirect the answer, while pathway transfers test whether answer-changing evidence resides at visual or question positions.

Our analysis separates four aspects of visual evidence use that are indistinguishable from accuracy alone. *The image is needed:* on CARVE, blind accuracy of the Qwen2.5-VL models stays below 6% while sighted accuracy is 38–49%. *Answers follow the evidence, modestly:* post-training raises conditional edit success on three of four backbones, significantly on the items all configurations of a backbone solve for 10 of 12 trained configurations. *Values are not bound to their elements:* models keep the original answer after 10.9–39.5% of edits that move a label, against at most 1.7% of numeral edits; on the items all configurations solve, post-training rarely changes this rate, and GRPO on Qwen3-VL-4B more than doubles it. *The gain is not reliably carried at the evidence:* the edited region carries 43% of the gain in edited-image accuracy in our 3B models but 96% in our 7B models (the intervals are wide), and frozen-tower controls locate the change in the decoder. Answer-changing evidence stays at visual positions, and attention to the image rises after training whether or not the causal role of visual states falls.

Our contributions are:

1. CARVE, a benchmark of solver-certified counterfactual edits with recorded masks that separates value from referent evidence (Section 3).
2. Localized causal evidence following, with outside-region and frozen-tower controls and certified pathway transfer (Section 4).
3. A comparison of three RL objectives on four backbones showing that behavioral gains, binding failures, and the location of evidence use dissociate, and that attention does not track causal use (Section 6).

2 RELATED WORK

RL post-training and visual use. RL post-training improves VLMs on multimodal reasoning benchmarks (Wang et al., 2025; Huang et al., 2026), yet reasoning-oriented models hallucinate more and attend less to the image over long chains (Xu et al., 2025). Recent objectives target visual use directly: VPPO reweights advantages by visual dependency and restricts updates to perceptually relevant tokens (Huang et al., 2026); PRPO reshapes token-level advantages with a visual-dependency metric (Li et al., 2026a); Vision-SR1 rewards perceptions that support answering without the image (Li et al., 2026c); INTUITOR replaces the answer reward with the model's own certainty (Zhao et al., 2026). These objectives are evaluated by accuracy, so whether they change how evidence is used remains untested.

Visual information flow. Where visual information goes is contested: it can flow directly from the visual-token positions where the object appears (Neo et al., 2025), be summarized first in a subset of late visual tokens (Basu et al., 2024), or be compressed into the question-token representations, from which a VLM can describe an image with no direct access to the image tokens at all (Zhang et al., 2025; Kaduri et al., 2025). The accounts conflict directly: Neo et al. (2025) find that blocking attention into those late visual tokens barely affects object identification. Salazar et al. (2026) reconcile the picture as direct and text-mediated pathways, show the choice is task-dependent, and find that models fall back to text mediation when the usual route is ablated. Causal tracing extends these tools to VLMs (Basu et al., 2024; Li et al., 2026b), binding analyses ask how features attach to the right object (Assouel et al., 2026), and the stability of activation restoration across latent-noise counterfactuals has been proposed as a hallucination signal (Gupta et al., 2026).

Counterfactual evaluation. Paired images are increasingly used to test evidence use. Programmatic suites resample problems for robustness (Zou et al., 2025), move diagram content into the question (Zhang et al., 2024), or change one semantic property of a rendered scene (Bansal & Gangwani, 2026); generative edits violate category priors (Sun et al., 2026), remove objects to test explanation faithfulness (Menzat et al., 2026), or target concepts a model cites in its own explanation (Ding et al., 2026); and blurring a detected region measures output sensitivity (Zhang, 2026). Most score the edit behaviorally, with the post-edit answer left undefined or judged by a model, and none separates reading a value from binding it to an element or compares models before and after RL. Certified counterfactual answers exist for patching studies (Salazar et al., 2026), but without a per-item record of which region changed, an intervention cannot be aimed at the evidence itself.

3 A COUNTERFACTUAL BENCHMARK WITH CERTIFIED EDITS

To test whether an answer uses a specific piece of visual evidence, we must change that evidence and know what the answer should become, which scene element was changed, and where it lies in the image, so that the response can be scored and an intervention aimed at that region. We introduce CARVE (**C**ertified **A**nswer-changing, **R**egion-localized **V**isual **E**dits), a benchmark that provides all three by construction.

Certified pairs. Each item is generated from a template: we sample scene parameters θ (for example, point coordinates, printed values, and label positions) from template-specific ranges, render the image $x = \mathrm{Render}(\theta)$, and compute the answer to question q with a symbolic solver, $y = \mathrm{Solve}(\theta, q)$. Because θ is known by construction, the answer is exact rather than estimated from the image. An edit changes θ to θ', giving $x' = \mathrm{Render}(\theta')$ on the same canvas and the certified new answer $y' = \mathrm{Solve}(\theta', q)$; edits with $y' = y$ are discarded. The edit mask M is the pixel difference, verified to be empty outside M up to anti-aliasing. Each evaluation unit is therefore

$$(x,\ x',\ q,\ y,\ y',\ M), \qquad y \neq y', \tag{1}$$

providing a fixed question, a certified answer change, and a known region of visual change. Section 4 maps M to the corresponding visual-token positions $S(M)$.

Design choices. Certification requires scenes that provably match their pixels, so CARVE is synthetic: a measurement instrument, not a proxy for real-world accuracy. Each edit changes one element under a fixed question; global transformations, which alter every patch, are excluded.

Value and referent edits. *Value edits* change the quantity the answer depends on, by rewriting a printed numeral and redrawing the geometry it annotates (*numeral edits*) or by moving an element so

that a counted or compared fact changes (*structural edits*). *Referent edits* keep every printed value and change the element it belongs to, by swapping or moving labels or relocating a congruence or right-angle mark, so following them also requires binding values to elements. In Figure 1, a model that halves any visible numeral answers 31, 35, and 31, and only the referent edit exposes it. Behavioral evaluation uses both families; localized transfer (Section 4.3) uses numeral edits only, because a referent edit changes a relation carried largely by geometry outside M. Operators are listed in Appendix A.2.

Ensuring the image is necessary. Questions specify the task and the relevant referents but not the quantities that determine the answer. Because a model can still exploit template-specific answer distributions or linguistic priors, we balance answer distributions within templates, drop items that Qwen2.5-VL-3B answers correctly without the image on at least two of eight samples, and remove edits whose new answer can be guessed from the question alone. No filter uses an evaluated model's predictions with the image to avoid selection bias across configurations (Appendix A.4).

Contents. CARVE has 50 templates spanning geometry, counting, charts, tables, graphs, puzzles, and instrument readings, rendered in four styles. The evidence is an annotation such as a label or tick mark in 27 templates (MODE-L) and a spatial property such as a length comparison in 23 (MODE-G). The evaluation set CARVE-MIXED caps each template at 60 items: 2,962 items and 6,137 edits. Removing 57 edits whose new answer is guessable from the question and the 120 edits of one template with a rendering error leaves 5,960 pairs (Appendix A.5).

4 EVALUATING VISUAL EVIDENCE FOLLOWING

We evaluate visual evidence use at three levels: answerability with and without the task image (Section 4.1), behavioral responses to certified edits (Section 4.2), and responses to interventions on hidden states (Sections 4.3 and 4.4). Behavioral and causal evaluation both use CARVE pairs (x, x', q, y, y', M) (Eq. 1). We write $\hat{y} = f(x, q)$ and $\hat{y}' = f(x', q)$ for the model's predictions on the original and edited images.

4.1 SIGHTED AND BLIND ACCURACY

Blind accuracy scores the model with the task image removed, using the same questions, answers, and decoding; a response without a final answer is scored as incorrect. It measures answerability from the question and learned priors.

4.2 BEHAVIORAL EVIDENCE FOLLOWING

Our behavioral evidence-following metric (BEF) asks whether a model that solves the original item follows the edit. We consider pairs whose original prediction is correct and measure how often the edited-image prediction matches the certified new answer:

$$\mathrm{BEF} = \Pr\big(\hat{y}' = y' \mid \hat{y} = y\big). \tag{2}$$

A failed edit on a solved pair either keeps the original answer or gives another one, including no final answer.

4.3 CAUSAL EVIDENCE FOLLOWING

Behavioral evidence following shows whether a model responds correctly to an edited image, but not whether the internal representations of the edited evidence can redirect its prediction. Causal evidence following (CEF) tests this through localized cross-image activation patching. For each pair, we run the original image (x, q) as the *recipient* and the edited image (x', q) as the *donor*. Let $H_j^\ell(x, q)$ denote the hidden state at token position j at decoder layer ℓ, and let $S(M)$ denote the visual-token positions that overlap the edited region M. We replace the recipient states at these positions with the donor states,

$$\widetilde{H}_j^\ell = \begin{cases} H_j^\ell(x', q), & j \in S(M), \\ H_j^\ell(x, q), & j \notin S(M), \end{cases} \tag{3}$$

continue the forward pass, and denote the resulting prediction by $\hat{y}_\ell^{\mathrm{CEF}}$. Following the behavioral metric, we define

$$\mathrm{CEF}^\ell = \Pr\big(\hat{y}_\ell^{\mathrm{CEF}} = y' \mid \hat{y} = y\big). \tag{4}$$

CEF therefore measures how often transferring the edited-region representations is sufficient to redirect an originally correct prediction to the certified counterfactual answer.

Comparing transfer with behavior. At the entry to the first decoder layer, text states are identical in both runs and transferring all visual states reproduces edited-image inference, so accuracy on the edited image is the ceiling for localized transfer. We therefore compare, on the same pairs, how often the model gives the certified new answer when it sees the edited image and when only the edited-region states are transferred, without conditioning on the original answer, so that both rates share one denominator. The *gain share* is the change in transfer accuracy after post-training divided by the change in edited-image accuracy, pooled over objectives: the part of the behavioral gain that the edited region carries (Appendix B.2).

Interpretation. Because x and x' differ only within M, successful transfer shows that the states at the edited positions suffice to produce the certified answer. Failed transfer does not show that the edit was ignored, since the vision encoder spreads part of the edit's effect beyond $S(M)$; we therefore report an outside-region control of equal size, measure how much of the edit's change falls inside $S(M)$, and treat CEF as a lower bound on evidence use at the edit.

4.4 Pathway transfer and attention

Prior work finds that visual information can stay at visual positions (Neo et al., 2025) or move into question tokens (Zhang et al., 2025; Kaduri et al., 2025), and that both routes can carry it to the answer (Salazar et al., 2026). To ask which route carries answer-changing evidence, we transfer donor states at a middle decoder layer ℓ over three broader position sets, all visual positions V, question positions Q, and all prompt positions, and score $T_P^\ell = \Pr(\hat{y}^{P,\ell} = y' \mid \hat{y} = y)$, where $\hat{y}^{P,\ell}$ is the prediction after replacing the recipient states at positions P, written T_V, T_Q, and T_{all}. Layers below ℓ keep the recipient's cache, so T_{all} is the ceiling at that depth. Because V covers every visual position, this test also applies to referent edits. Following restoration and corruption in causal tracing (Basu et al., 2024), we also use a noise image as the donor and score whether the model's original prediction survives when clean visual states are restored into a noise-image run (*visual-state recovery*) or question states are replaced with noise-run states (*question-state corruption*). Finally, attention to visual tokens is widely read as a measure of visual reliance (Liu et al., 2024; Xu et al., 2025), although attention sinks make its magnitude an unreliable guide (Kang et al., 2025). We therefore report the attention share from answer tokens to visual tokens and test whether it moves with these causal measures.

5 Experimental Setup

Models. The main comparison covers four backbones, Qwen2.5-VL-3B, Qwen2.5-VL-7B (Bai et al., 2025), Qwen3-VL-2B, and Qwen3-VL-4B (Qwen Team, 2025), each as its public base model and after post-training with GRPO (Shao et al., 2024), VPPO (Huang et al., 2026), or PRPO (Li et al., 2026a). All trained configurations update the vision tower and share their backbone's public initialization. To separate encoder from decoder effects, we also train GRPO, VPPO, and PRPO with a frozen vision tower on Qwen2.5-VL-3B, and GRPO with a frozen tower on Qwen2.5-VL-7B.

Training. All objectives share one recipe: ViRL39K prompts (Wang et al., 2025) (31,629 after cleaning, with 1,000 held out), AdamW at a constant learning rate of 10^{-6}, 384 prompts per rollout step with eight responses each, split into optimizer updates of 128 prompts, asymmetric DAPO clipping (Yu et al., 2025) ($\epsilon_{\mathrm{low}} = 0.2$, $\epsilon_{\mathrm{high}} = 0.28$), no KL penalty, and two epochs (162 steps for Qwen2.5-VL, 164 for Qwen3-VL). We evaluate the final checkpoint, and each configuration uses a single training seed (Appendix B.1).

Evaluation. Behavioral metrics use all 5,960 CARVE-Mixed pairs. CEF uses the 1,546 numeral edits of a subset with one edit per item, at early, middle, and late decoder depths, the early depth being the entry to the first decoder layer; noise-image interventions use 1,000 items; and certified pathway transfer uses 2,011 pairs from 980 items (Appendix B.2). Blind evaluation omits the image, except for Qwen3-VL on MMK12, which uses a blank image. Decoding is greedy, with the same

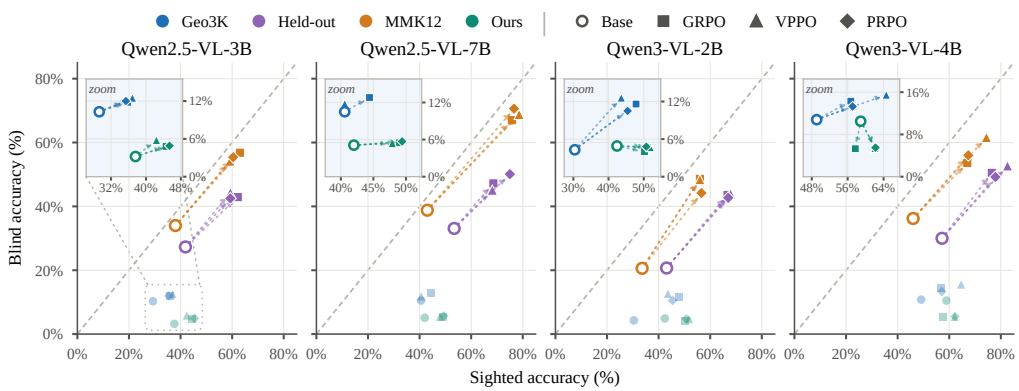


Figure 2: Benchmark accuracy with and without the task image.

prompt, answer extraction, and generation limit per backbone family. Base Qwen3-VL models often give no final answer within the limit, mostly through repetition loops or unfinished reasoning; among solved pairs they give none on 9–12% of edited images, against at most 2% after training.

Statistics. Intervals are 95% source-item bootstrap intervals (10,000 resamples), which resample each item with all its edits; tests within a family use Holm correction. With a single training seed, these intervals reflect evaluation sampling only, and we do not rank objectives (Appendix B.3).

6 Results

We organize the results by the question each measurement tests: whether the image is necessary (Section 6.1), whether answers follow certified evidence (Section 6.2), whether values are bound to their correct referents (Section 6.2.1), whether the effect is localized to the relevant evidence (Section 6.3), and whether attention to the image tracks these changes (Section 6.4).

6.1 Accuracy with and without the image

On standard benchmarks, much of the performance survives without the image (Figure 2): for the Qwen2.5-VL-3B base model, blind accuracy on MMK12 (Meng et al., 2025) nearly matches sighted accuracy (34.0% vs. 38.0%). For the Qwen2.5-VL backbones, the blind gain from post-training is 77–98% of the sighted gain on the held-out split and MMK12, but less on Geo3K (Lu et al., 2021), so much of the gain on the first two does not require the image. On CARVE, blind accuracy stays below 6% for every trained configuration and rises by at most 2.6 points while sighted accuracy improves. Low blind accuracy shows that the image is necessary, not that the model follows the evidence that determines the answer, which the certified edits test next.

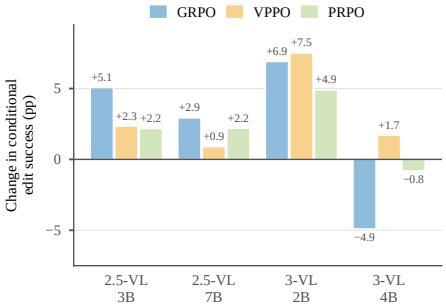


Figure 3: Change in conditional edit success from base (pp).

6.2 Behavioral evidence following

We evaluate whether post-training improves responses to answer-changing edits with conditional edit success (BEF) on all 5,960 pairs (Figure 3). All three objectives raise it on Qwen2.5-VL-3B, Qwen2.5-VL-7B, and Qwen3-VL-2B, by 2.2–5.1, 0.9–2.9, and 4.9–7.5 points, whereas on Qwen3-VL-4B GRPO lowers it by 4.9 points and VPPO and PRPO change it by less than 2. Because each rate conditions on the pairs a configuration solves, we test the changes on the items every configuration of a backbone solves. On this fixed set, all six Qwen2.5-VL increases are significant after Holm correction (3B: 5.1–7.6 points; 7B: 2.0–3.9), as are the three Qwen3-VL-2B increases, the Qwen3-VL-4B VPPO increase, and its GRPO decrease (Appendix C). The benefit of post-training is real but modest and backbone-dependent; with one seed per configuration, we do not rank objectives.

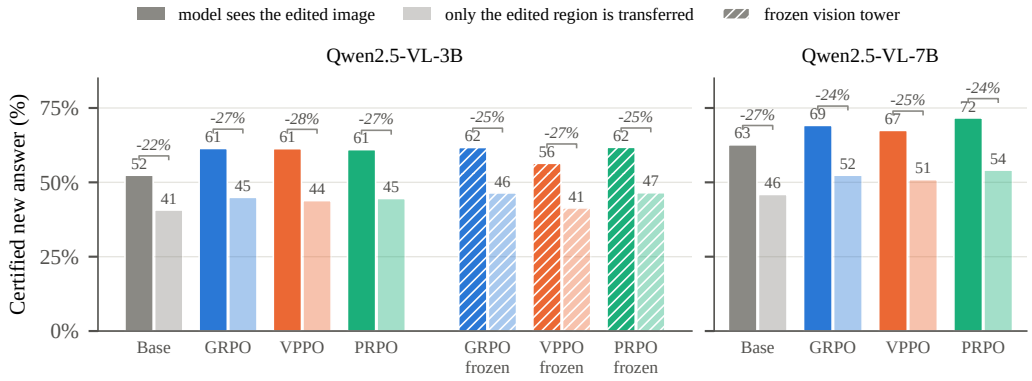


Figure 4: Certified answers: edited image vs. edited-region transfer (numeral edits, layer $\ell = 0$).

6.2.1 Prior lock: following a number but not its referent

Edit success pools edits that differ in what they require. Numeral edits change a printed value and structural edits the geometry carrying it; we group them, since both change the quantity the answer depends on. A referent edit instead keeps every value and moves it to another element (Figure 1). We measure how often a solved pair keeps its original answer after the edit; on referent edits, this is prior lock.

The two edit families dissociate in every configuration. In all 16 configurations, models keep the original answer after at most 1.7% of numeral edits but after 10.9–39.5% of referent edits (Figure 5). Within scenes that carry both edit types, referent retention exceeds numeral retention by 25–39 points in every Qwen2.5-VL configuration. When a referent edit fails, the model returns the original answer in 60–74% of cases and another printed value in only 1–8%, so the failure is a lock on the old answer rather than confusion between labels. Nor is it a failure to see the moved label: in every Qwen2.5-VL configuration except the 3B base model, most locked edits (58–87%) pass both a reading probe and a probe asking which element carries the moved value (Appendix C). Accuracy on the original image, blind evaluation, and numeral edits all miss this failure; only a certified edit that leaves every printed value unchanged exposes it.

Post-training does not remove it. On each configuration's solved pairs, referent retention falls at Qwen2.5-VL-3B (from 39.5% to 31.7–35.2%) and rises elsewhere, but these comparisons mix training effects with changes in which items qualify. On commonly solved items, 8 of the 12 changes are not significant after Holm correction, even though edit success rises significantly on those items for 10 of 12. The exceptions are small decreases under GRPO and VPPO on Qwen3-VL-2B (-3.4 and -3.6 points) and increases under PRPO ($+5.4$) and GRPO ($+13.1$) on Qwen3-VL-4B (Appendix C). On Qwen3-VL-4B, GRPO's loss in edit following is prior lock: on referent edits, edit success falls from 78.2% to 67.5% while retention rises from 10.9% to 23.3%, whereas on numeral-edit, edit success barely moves (91.4% to 90.4%).

6.3 Is the behavioral gain carried by the evidence location?

If post-training improves edited-image accuracy because the model uses the changed evidence where it lies, transfer from the edited region should rise by a similar amount; if the gain comes from elsewhere in the image or from the decoder's reasoning, it should not. We test this on numeral edits and the two Qwen2.5-VL backbones, where every configuration answers nearly all items (Figure 4).

At 3B, the edited region carries only part of the gain. Post-training raises edited-image accuracy from 52% to 61% but transfer accuracy only from 41% to 44–45%, a gain share of 43% (95% CI $[18, 67]$). **At 7B, it carries nearly all of it.** Edited-image accuracy rises from 63% to 67–72% and transfer from 46% to 51–54%, a gain share of 96% ($[66, 139]$), 54 points above 3B ($[16, 101]$). With two model sizes from one family, we do not read it as a trend with scale. Conditioning on solved originals (Eq. 4) gives the same directions (Appendix D).

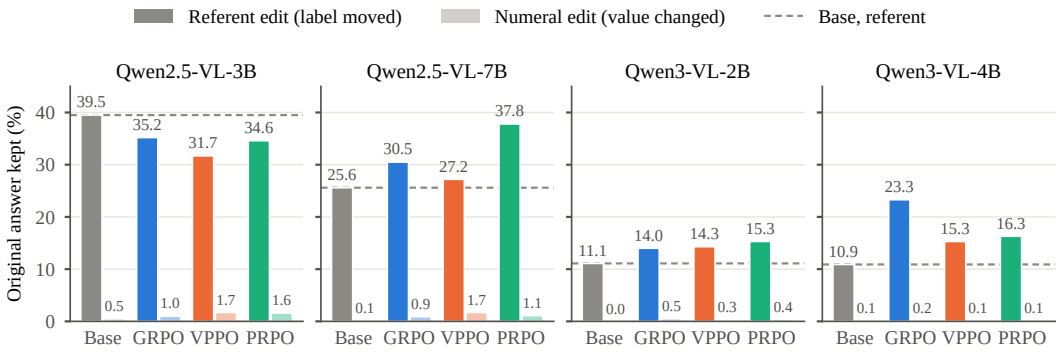


Figure 5: Share of solved pairs on which the model keeps the original answer after the edit.

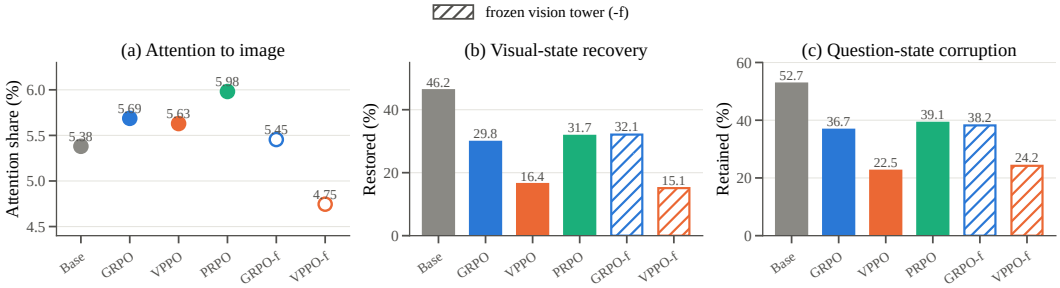


Figure 6: Model: Qwen2.5-VL-3B (a) Attention from answer tokens to visual tokens, layers 12–23. (b) Original answer restored by clean visual states in a noise-image run, layer 12. (c) Original answer kept after question states are replaced by noise-run states, layer 12.

The change arises in how the decoder uses the image, not in how the encoder sees it. Training does not change how the encoder spreads the edit: the edited region holds the same share of the edit's effect at the decoder input before and after training (Appendix D). Freezing the vision tower, so that the edited region's states stay exactly the base model's, does not remove the pattern either: at 3B the frozen models show the same gap between seeing the edit and transferring it as the trainable ones (Figure 4, hatched bars), and at 7B frozen GRPO carries as much of its gain at the edit as trainable GRPO (Appendix D). Post-training therefore alters how the decoder uses the evidence (Fu et al., 2025), and can raise accuracy on edited images without relying more on the evidence location.

Controls and bounds. Transferring an equal number of positions outside the edit yields the new answer on at most 2.5% of solved pairs, so the edited positions carry edit-specific information. They hold 26–29% of the edit's layer-0 change while covering 9% of visual tokens (Appendix D), so localized transfer is a lower bound on evidence use at the edit.

6.4 Does more attention to the image mean more use of it?

Better use of visual information is often read from more attention to image tokens. We test this on 1,000 items of Qwen2.5-VL-3B (Figure 6) and 7B (Appendix E) by exchanging states at a middle decoder layer between the clean run and a run whose image is replaced by noise, with the question unchanged, and scoring whether the original answer survives.

At 3B, the causal role of visual states falls after training. *Visual states restored* into the noise-image run recover the original answer on 46.2% of base items but 16.4–31.7% after training: the same states support the answer less often once trained. *Question states corrupted* cost trained models more (answer kept on 52.7% for base, 22.5–39.1% after training): those positions are harder to do without. *Question states restored* alone, however, recover it no more often than the noise image alone keeps it (2.6–3.1% vs. 2.7–3.8%), so they are not a self-contained summary of the image. At 7B the fall appears only under VPPO; GRPO and PRPO stay within 6 points of base (Appendix E).

Attention does not track this change. Middle-layer attention to visual tokens rises after training at 3B (5.38% to 5.63–5.98%) while recovery falls; with a frozen tower the causal pattern is unchanged but attention does not rise (5.45% and 4.75%), and at 7B attention rises under all three objectives

although only VPPO shows the causal drop. Attention share therefore does not indicate reliance on visual states at either size.

Answer-changing evidence stays at visual positions. The exchanges above ask whether an answer survives; this one asks whether a *new* one is produced. We copy edited-image states into the original-image run at the middle layer and score the certified new answer (Figure 7). At both 3B and 7B, visual positions reproduce the depth-matched ceiling in every configuration ($T_V/T_{\text{all}} = 0.96$–$1.02$): transferring them alone is as effective as transferring everything. Question positions yield the new answer on at most 1.9% of value and 7.9% of referent pairs; post-training raises the referent rate by up to 7 points and leaves the value rate unchanged (Appendix E). After training, question states are therefore more *necessary* for the answer but not *sufficient* to change it: the route passes through them more fragilely, while the evidence that decides the answer is still read from visual positions.

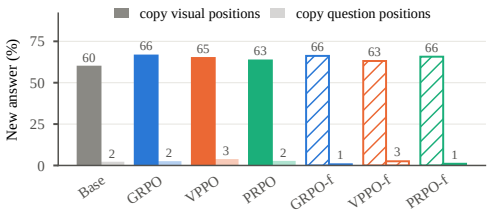


Figure 7: Share of solved pairs that switch to the certified new answer when edited-image states are copied at visual or question positions (Qwen2.5-VL-3B, layer 12).

7 Discussion

Our results suggest that improved multimodal performance after RL post-training does not necessarily imply stronger use of the evidence that should determine the answer. On CARVE, RL improves responses to certified edits on three of four backbones, yet binding errors persist and the gain is not uniformly carried by the edited evidence: the edited-region patch recovers 43% of the behavioral gain in our 3B models but 96% in our 7B models, while answer-changing information stays at visual rather than question positions. Better answers on edited images can thus arise from stronger use of the relevant evidence or from greater reliance on other visual information; whether the 3B–7B difference reflects scale will require more sizes and seeds. Prior work has often assessed visual grounding through tests such as whole-image perturbations, image dependence, or attention magnitude. Our results caution against relying on such tests in isolation. VPPO and PRPO, despite explicitly emphasizing visually dependent tokens, do not consistently separate from GRPO under our behavioral or causal evaluations, while attention increases after training at both scales even when the causal contribution of visual states does not. These observations suggest that such proxies can indicate whether vision matters broadly, but not necessarily whether the model follows the particular visual evidence that should determine its answer.

CARVE and CEF provide a complementary way to test such claims. CARVE supplies counterfactuals for which both the required answer change and its image location are known, while CEF asks whether the internal representation of that localized evidence is sufficient to redirect the model's prediction. Together, they turn grounding from an aggregate performance claim into a testable question about whether the right evidence actually controls the answer. This distinction is directly relevant to visual hallucination and, more broadly, to AI safety in multimodal systems, where reliable behavior depends not only on producing plausible or correct outputs but on remaining tied to the evidence that should determine them.

More broadly, CEF provides a way to move beyond observational proxies and directly test whether localized evidence causally influences a model's decision. Certified pairs can also provide ground truth for circuit and representation analyses and support stronger intervention-based evaluations of evidence use. We hope CARVE and CEF encourage evaluations that ask not only whether a model produces the correct answer, but whether the evidence that should determine it actually does so.

Limitations. Each configuration uses one training seed, so we do not rank objectives, and the 3B–7B contrast rests on two sizes from one family. CARVE is synthetic and excludes natural images. Because post-training changes answer completion in Qwen3-VL, the causal analyses use Qwen2.5-VL, where every configuration answers at least 97% of items. Localized transfer is a lower bound and covers only numeral edits, which also redraw the geometry they annotate, a cue referent edits lack, although readout probes show that most locked edits are read and located correctly. We evaluate only answer-changing edits.

AI USE STATEMENT

We used large language model assistants to edit the manuscript and to draft portions of its text, which the authors revised; to help find and summarize related work, with every reference checked by the authors against the original publication; and to write and debug code for the scene generators, training and intervention harnesses, analysis scripts, and figures. We did not use generative AI to produce benchmark content: CARVE images are rendered programmatically from scene parameters, and every answer is computed by a symbolic solver. The only model used in dataset construction is the blind-solvability filter (Appendix A.4), which queries Qwen2.5-VL-3B with the image withheld. All AI-assisted code was reviewed and run by the authors, every reported number is regenerated from logged outputs, and we take full responsibility for the content of this work

REPRODUCIBILITY STATEMENT

All numbers regenerate from versioned scripts against logged run artifacts. The benchmark generator, intervention harnesses (localized patching and pathway transfer), per-item outputs, and provenance records (git SHA, model path, seed) will be released. Training and evaluation settings are in Appendices B.1 and B.2; intervals are source-item bootstrap intervals, and tests use bootstrap p-values with Holm correction (Appendix B.3).

REFERENCES

Rim Assouel, Declan Iain Campbell, Yoshua Bengio, and Taylor Whittington Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3RQ863cRbx.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Aayam Bansal and Ishaan Gangwani. Counterbench: A controllable counterfactual testbed reveals systematic reasoning failures in vision-language models. In *The 1st CogVL: Cognitive Foundations for Multimodal Models Workshop at CVPR 2026*, 2026. URL https://openreview.net/forum?id=U9LQ19atDB.

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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=s63dtq0mwA.

Sihao Ding, Santosh Vasa, Aditi Ramadwar, and Thomas Monninger. Edct-bench: Uncovering faithfulness gaps in vlms via explanation-driven counterfactual testing, 2026. URL https://arxiv.org/abs/2609.17953.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations, 2025. URL https://arxiv.org/abs/2506.08008.

Abhivansh Gupta, Simardeep Singh, Advika Sinha, Shreyansh Modi, and Akshat Tomar. How many counterfactuals does it take? probing vlm hallucinations through circuits and causal effects, 2026. URL https://arxiv.org/abs/2606.08777.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=bRA4lVWJVQ.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558, June 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7uDI7w5RQA.

Qiming Li, Tianlun Li, Xiaolong Cheng, Hangyu Li, Ruiyan Gong, Kangning Niu, Kaitao Jiang, and Mu Xu. Prpo: Perception-reinforced policy optimization via token-level dynamic advantage reshaping, 2026a. URL https://arxiv.org/abs/2606.08708.

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 *AAAI*, pp. 31645–31653, 2026b. URL https://doi.org/10.1609/aaai.v40i37.40431.

Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Lee Boyd-Graber, Haitao Mi, and Dong Yu. Vision-SR1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=C1M4ETatgM.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

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.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.528. URL https://aclanthology.org/2021.acl-long.528/.

Fanqing Meng et al. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

Bayar Menzat, Maximilian Süss, Ruizhi Wang, Benno Steinegger, Thomas Lukasiewicz, and Oana-Maria Camburu. Counterfactual tests for measuring chain-of-thought faithfulness in visual language models, 2026. URL https://arxiv.org/abs/2609.06704.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chanJGoa7f.

Qwen Team. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Israfel Salazar, Stella Frank, Dan Oneata, Desmond Elliott, and Constanza Fierro. Pathways of visual information flow in vision-language models, 2026. URL https://arxiv.org/abs/2607.03358.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. URL https://doi.org/10.48550/arXiv.2402.03300.

Jingyu Sun, Jiachen Tu, Yuyang Xue, Yaoxin Jiang, Guoyi Xu, Zhengtao Yao, Rui Qian, Yizheng Sun, Hongpeng Zhou, Jingyuan Sun, and Yan Lin. Seeing what is actually there: Prive-bench and prive-tools for counterfactual evaluation of agentic visual evidence in vlms, 2026. URL https://arxiv.org/abs/2607.16311.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4oYxzssbVg.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KzU33wR875.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2a36EMSSTp.

Genpei Zhang. The visual insensitivity gap: Diagnosing when vision-language models fail to use visual evidence, 2026. URL https://arxiv.org/abs/2609.00868.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII*, pp. 169–186, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73241-6. doi: 10.1007/978-3-031-73242-3_10. URL https://doi.org/10.1007/978-3-031-73242-3_10.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791, 2025. doi: 10.1109/CVPR52734.2025.01842.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OU9nFEYR2M.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VOAMTA8jKu.

A BENCHMARK CONSTRUCTION DETAILS

A.1 TEMPLATES AND EVIDENCE MODES

CARVE has 50 templates spanning geometry, counting, charts, tables, graphs, puzzles, and instrument readings, in four rendering styles. MODE-L templates (27) ask for exact values carried by annotations such as labels or tick marks. MODE-G templates (23) ask only questions robust to visual estimation and sample parameters with a decidability margin so that a careful human scores near 100%: length comparisons only when the ratio exceeds 1.25, and angle categories only when the angle is at least 10° from 90°; marks are separated by at least 6 pixels. The templates render in four style families (10 in the style of Geo3K, 13 MathVision, 15 DynaMath, 12 MathVerse); 35 have numeric and 15 categorical answers. CARVE-MIXED caps each template at 60 items; 33 templates reach the cap and the rest contribute 50–59. Reading a value is necessary but not sufficient on MODE-L templates, since referent edits test binding the value to its element. Table 1 lists every template with its style, mode, answer type, and edit counts.

A.2 EDIT TAXONOMY

Value edits comprise value rewrites (MODE-L; the label and the geometry it annotates always change together, since a label contradicting the drawing would be a broken item) and geometry edits (MODE-G; a point moves or a radius changes so that the compared or counted fact flips). *Referent edits* comprise referent swaps (two element labels exchanged so that the question's referent changes), label-role moves, and congruence-mark moves (a mark relocated, changing the derivation). Of the 6,137 edits in MIXED, 2,470 are value rewrites, 2,756 structural edits (2,696 geometry edits and 60 part-count changes), and 911 referent edits (819 swaps, 60 label-role moves, 32 mark moves). Global transformations (rotation, translation, scaling) are excluded: they touch every patch, which breaks the locality the interventions require, and rotation silently changes answers for orientation-dependent templates. Numeral masks are larger than referent masks (median 32 vs. 5 visual tokens), because a value rewrite also redraws the geometry it annotates. The generator also produces edits that should leave the answer unchanged (style edits, distractors, edits to values the answer does not depend on, and renaming of unreferenced points); they are not evaluated in this paper.

A.3 CONSTRUCTION AND LOCALIZATION

Each item begins as a parameter vector θ for a template. The solver computes y from θ and q before rendering, so the answer never depends on the rendered pixels. An edit operator maps θ to θ', the solver recomputes y', and the scene is re-rendered with the same canvas, style, and random seed, so that only the edited elements differ. The parameter difference identifies the affected elements, whose rendered footprint gives the mask M. The visibility check (F2 below) confirms that x and x' differ inside M and are identical outside it up to anti-aliasing.

A.4 FILTERING PIPELINE

Filters run from cheap to expensive, and each logs a pass or fail per item or edit. **F0, render lint:** no overlapping labels, all elements on canvas with a margin, minimum font size and mark separation. **F1, solver certification:** each edit must change the re-solved answer by more than three times the answer tolerance. Referent swaps change labels rather than parameters, so their new answer comes from the template's solver on the relabeled scene rather than from an independent second solve. **F2, visibility:** the pixel difference must change at least one full visual token inside M and nothing outside it, allowing an anti-aliasing tolerance of 12 intensity levels per channel and a 3-pixel dilation of M. **F3, deduplication:** by template and quantized parameter vector, and by perceptual hash of the original image. **F4, answer-prior balance:** a template-level check that the most frequent answer does not exceed a balanced share by more than 0.22, and that edits raise the answer in 25–75% of cases, so that "guess higher" or "guess lower" is not a strategy; templates failing it are regenerated rather than items dropped. **F5, blind solvability:** $k = 8$ samples at temperature 0.7 from Qwen2.5-VL-3B-Instruct with the image withheld; an item is dropped if it is answered correctly on at least $2/8$ samples. Of 6,700 generated items, 81 (1.2%) are removed. **F5b, answer-side audit:** on a 20%

Table 1: The 50 CARVE templates in CARVE-MIXED. Style: rendering style family. Mode: MODE-L (exact values from annotations) or MODE-G (estimation-robust questions). Answer: numeric or categorical. Items: items in MIXED. Value, Struct., Ref.: generated value rewrites, structural edits, and referent edits. Excl.: edits removed because the new answer is guessable from the question, or (†) because of T25's rendering error. Pairs: evaluated pairs.

ID	Template	Style	Mode	Answer	Items	Value	Struct.	Ref.	Excl.	Pairs
T1	triangle angle sum	Geo3K	L	num.	60	60	0	60	1	119
T2	parallel transversal	Geo3K	L	num.	60	120	0	0	1	119
T3	pythagoras	Geo3K	L	num.	60	120	0	0	–	120
T6	circle intersection	Geo3K	G	cat.	57	0	114	0	–	114
T7	exterior angle	Geo3K	L	num.	59	118	0	59	–	177
T8	isosceles angle	Geo3K	L	num.	60	60	0	77	2	135
T9	angle division	Geo3K	L	num.	60	60	60	0	8	112
T10	similar triangles	Geo3K	L	num.	60	138	0	0	–	138
T11	inscribed angle	Geo3K	L	num.	60	60	0	60	–	120
T12	composite area	Geo3K	L	num.	60	179	0	0	–	179
T13	brick wall	MathVision	G	cat.	60	0	120	0	–	120
T14	shaded cells	MathVision	G	cat.	59	0	118	0	–	118
T15	dice hidden pips	MathVision	G	cat.	60	0	120	0	–	120
T16	lattice area	MathVision	G	cat.	59	0	118	0	–	118
T17	lattice perimeter	MathVision	G	cat.	59	0	118	0	–	118
T19	points inside	MathVision	G	cat.	60	0	120	0	–	120
T20	stack map	MathVision	L	num.	60	120	0	0	–	120
T21	table total	DynaMath	L	num.	60	60	0	60	–	120
T22	bars above line	DynaMath	G	cat.	59	0	118	0	–	118
T23	function value	DynaMath	G	num.	58	0	116	0	–	116
T24	node degree	DynaMath	G	cat.	56	0	56	56	12	100
T25	bisector transversal	MathVerse	L	num.	60	60	0	60	120†	0
T26	histogram range	DynaMath	G	num.	59	0	118	0	7	111
T27	venn counts	DynaMath	L	num.	60	120	0	60	–	180
T28	shortest path	DynaMath	L	num.	60	60	0	60	–	120
T29	starred addition	DynaMath	L	cat.	60	120	0	0	–	120
T31	tangent secant	MathVerse	L	num.	60	60	0	60	–	120
T33	intersecting bisector	MathVerse	L	num.	60	60	0	60	–	120
T34	perpendicular mark	MathVerse	L	num.	60	60	0	60	–	120
T35	sector area	MathVerse	L	num.	60	120	0	0	–	120
T36	prism volume	MathVerse	L	num.	58	116	0	0	–	116
T37	slope on grid	MathVerse	G	num.	60	0	120	0	1	119
T38	curve crossings	MathVerse	G	cat.	56	0	112	0	–	112
T39	circle on grid	MathVerse	G	num.	59	0	118	0	–	118
T40	magic triangle	MathVision	L	num.	60	60	60	0	–	120
T43	dot plot median	MathVision	G	num.	60	0	120	0	–	120
T44	seven segment	MathVision	G	num.	60	0	120	0	–	120
T45	nested circles	MathVision	G	cat.	59	0	118	0	–	118
T48	edge sum puzzle	DynaMath	L	num.	60	60	0	60	–	120
T49	pie slice angle	DynaMath	L	num.	59	59	0	59	–	118
T50	trapezoid area	MathVerse	L	num.	60	120	0	0	–	120
T51	polygon interior angle	DynaMath	G	num.	50	0	100	0	23	77
T52	inner square area	DynaMath	L	num.	60	120	0	0	–	120
T53	edge count	DynaMath	G	cat.	60	0	120	0	–	120
T54	gauge needle	DynaMath	G	num.	60	0	120	0	–	120
T55	number train	MathVision	L	num.	60	60	0	60	2	118
T58	balance scale	MathVision	G	cat.	59	0	118	0	–	118
T59	sinusoid period	MathVerse	G	num.	57	0	114	0	–	114
T60	hottest month	DynaMath	G	cat.	60	0	120	0	–	120
T61	isosceles bisector	MathVerse	L	num.	60	120	0	0	–	120
Total (50 templates)					2,962	2,470	2,756	911	177	5,960

sample of edits, the same check is run against the new answer y', catching edit distributions whose outcomes are guessable from text. **F6, manual pass:** 120 items, three from each of 40 templates, are reviewed against a checklist (figure legible, question unambiguous, edit visible, new answer correct by hand), with every failure root-caused before the benchmark is frozen.

Model correctness is deliberately not a filter. Dropping items a model fails would tune the benchmark to that model and bias every cross-configuration comparison. Model behavior is used only diagnostically at the template level, and any parameter-range adjustment is logged rather than applied as a silent per-item drop.

A.5 From generated edits to evaluated pairs

CARVE-Mixed contains 2,962 items and 6,137 generated edits. We remove 57 edits whose new answer can be guessed from the question text alone, leaving 6,080 (they concentrate in T51, T24, and T9, with 23, 12, and 8 edits; Table 1), and the 120 edits of one template (T25) after a rendering error was found in it, leaving the 5,960 pairs used for all behavioral metrics. The localized-transfer set likewise excludes T25 (2,902 items). Value-free question text therefore does not by itself prevent leakage through the prior over answers, and counterfactual benchmarks need answer-side checks in addition to question-side filters.

B Experimental details

B.1 Training configurations

Table 2: **Training configurations.** All main configurations update the vision tower; frozen-tower configurations are used only for the encoder–decoder decomposition of CEF. A step is one rollout step (three optimizer updates).

Backbone	Objectives	Tower	Prompts	Steps	Prompts/step	Hardware
Qwen2.5-VL-3B	GRPO, VPPO, PRPO	trainable	31,629	162	384	2×H100
Qwen2.5-VL-3B	GRPO, VPPO, PRPO	frozen	31,629	162	384	2×H100
Qwen2.5-VL-7B	GRPO, VPPO, PRPO	trainable	31,629	162	384	4×GH200
Qwen2.5-VL-7B	GRPO	frozen	31,629	162	384	4×GH200
Qwen3-VL-2B	GRPO, VPPO, PRPO	trainable	31,629	164	384	2×H100
Qwen3-VL-4B	GRPO, VPPO, PRPO	trainable	31,629	164	384	2×H100

Shared recipe. Training uses ViRL39K prompts. From 38,870 raw prompts we drop 2,289 multi-image and 851 malformed prompts, remove 3,101 prompts that overlap the Geo3K test set (perceptual image hash distance at most 5 or fuzzy question match of at least 92), and hold out 1,000 prompts stratified by category, leaving 31,629 for training; prompts longer than the context limit (about 1%) are dropped when loading. Each rollout step samples 384 prompts with eight responses each (3,072 responses) and splits them into three optimizer updates of 128 prompts (1,024 responses). We use AdamW with a constant learning rate of 10^{-6} and no warmup, bfloat16, and the DAPO policy loss with asymmetric clipping ($\epsilon_{\text{low}} = 0.2$, $\epsilon_{\text{high}} = 0.28$) and no KL penalty. The reward is answer correctness under the ViRL39K scorer, with no format reward; the maximum response length is 2,048 tokens, and rollouts use temperature 1.0 and top-p 0.99. Training covers two epochs (162 rollout steps, 486 optimizer updates), and we evaluate the final checkpoint of every configuration, not the one with the best validation reward. Qwen2.5-VL-7B uses the same hyperparameters. Qwen3-VL uses the same recipe and launcher; after the length filter its loader has 82 batches of 384 prompts per epoch rather than 81, so its two epochs end at rollout step 164; a run takes about 40 (2B) and 48 (4B) GPU-hours on H100s.

Objective-specific settings. GRPO uses group-normalized advantages without further reweighting. VPPO (Huang et al., 2026) reweights trajectory advantages by visual dependence and restricts updates to perceptually relevant tokens (perception-token fraction 0.4, minimum advantage scale 0.9, entropy penalty 0.06). PRPO (Li et al., 2026a) reshapes token-level advantages with its visual-dependency metric (perturbation black-patch probability 0.6 with 14×14 patches, $\sigma = 0.01$,

$\eta = 0.001$, advantage-reshaping bounds $m = 0.1$, $M = 4.0$, $k = 2.0$). The VPPO configurations use both of its components, gradient filtering to the most visually dependent tokens and trajectory-level advantage scaling, with the published 7B hyperparameters.

B.2 Evaluation protocol

Evaluation sets. Behavioral metrics use the 5,960 audited pairs of CARVE-Mixed (Appendix A.5). The CEF set takes one edit per item (the first answer-changing edit in generator order, almost always a value rewrite; T25 excluded, Appendix A.5), giving 2,902 pairs, of which the 1,546 numeral edits are used. Noise-image interventions use a fixed 1,000-item subset chosen by a deterministic per-template ordering. Certified pathway transfer uses 2,011 audited pairs from a 980-item subset (1,728 value and 283 referent edits), plus 561 referent edits from other items for power.

Referent readout. For each referent edit, the model answers two probes on the original and the edited image. A *reading* probe asks it to list every number written in the figure, with no element named (two templates instead ask about the edited attribute: which slice is shaded, and whether a named segment carries a tick mark); it is scored against the generator's printed set, and we also record whether the moved value is present. A *binding* probe names a value and asks which element it labels ("Which angle is labeled $48°$?"), scored against the generator's record of the referent. The reading probe is available for 855 referent edits in 14 templates and the binding probe for 376 in 6 templates.

Decoding and scoring. All evaluations use greedy decoding with the same prompt template and answer extraction across configurations and conditions. Numeric answers are matched within each template's tolerance and categorical answers by normalized exact match. A response without a final answer is scored as incorrect; in the failure split it counts as neither the new nor the original answer. The generation limit is 512 tokens for Qwen2.5-VL. For Qwen3-VL, responses are generated with a 1,024-token limit, and every response without a final answer is regenerated at 2,048 tokens, in every configuration. Because decoding is greedy, this changes only responses that hit the first limit; it applies to 1–2% of trained-model responses and a third of base-model responses, and recovers 5% (2B) and 21% (4B) of the latter.

Intervention settings. CEF and pathway interventions run in float32 with batch size 9; precision and batching are fixed within each evaluation family because they can change patched predictions. Patching is applied during the prefill pass, after which generation proceeds normally from the patched cache. Conditional CEF metrics condition on the clean prediction of the CEF run itself: clean predictions from the behavioral run (bfloat16, batch 32) agree in correctness on 89–92% of items, so importing them would mix two denominators. Table 3 gives the decoder layers used.

Table 3: **Intervention depths.** The early depth is the entry to the first decoder layer, where transferring all visual states reproduces edited-image inference.

Backbone	Decoder layers	Early	Middle	Late
Qwen2.5-VL-3B	36	0	12	24
Qwen2.5-VL-7B	28	0	9	18

Mask-to-token mapping. $S(M)$ contains every visual token whose merged patch footprint overlaps M; the 532×532 canvas gives 14-pixel patches merged 2×2 into a 19×19 grid of 361 visual tokens. Outside-region positions are sampled by a deterministic stride over the merged token grid that excludes changed and evidence tokens, with the same count as $S(M)$ and no buffer around M.

Transfer metrics. Let $\hat{y}_\ell^{\mathrm{CEF}}$ be the prediction after the transfer in Eq. 3 at layer ℓ. Edited-image accuracy is $\mathrm{Acc}_{\mathrm{cf}} = \Pr(\hat{y}' = y')$ and transfer accuracy is $\mathrm{Acc}_{\mathrm{CEF}} = \Pr(\hat{y}_0^{\mathrm{CEF}} = y')$, both over all numeral pairs. For trained configurations f of a backbone with base f_0, the gain share is

$$\rho = \frac{\overline{\mathrm{Acc}}_{\mathrm{CEF}}(f) - \mathrm{Acc}_{\mathrm{CEF}}(f_0)}{\overline{\mathrm{Acc}}_{\mathrm{cf}}(f) - \mathrm{Acc}_{\mathrm{cf}}(f_0)},$$

where bars denote averages over the trained objectives (a ratio of means). The localization share $\mathrm{LS} = \Pr(\hat{y}_0^{\mathrm{CEF}} = y' \mid \hat{y}' = y',\ \hat{y} = y)$ is the fraction of a configuration's successful edit following that transfer reproduces.

B.3 Statistical analysis

Intervals. Intervals resample source items with replacement, keeping all edits of a sampled item, and recompute each metric on the resample (10,000 resamples, seed 0, percentile intervals). A difference between two configurations is computed on the same resampled items for both. Conditional metrics are recomputed within each resample on that configuration's own solved pairs; fixed-set comparisons restrict every configuration of a backbone to the items all of them solve. Ratios such as the gain share are computed within each resample; their denominators were positive in every resample.

Tests. Trained-minus-base differences are tested with two-sided bootstrap p-values from the same resamples, with Holm correction over the six comparisons (three objectives, two backbones) within each backbone family and each metric. Results for pathway transfer and attention are descriptive.

C Behavioral controls

Table 4: **Change in conditional edit success from base (points).** Own: each configuration's solved pairs, with 95% interval. Fixed: items every configuration of the backbone solves (Qwen2.5-VL-3B: 588 items, 1,216 pairs; 7B: 793 items, 1,680 pairs; Qwen3-VL-2B: 779 items, 1,655 pairs; 4B: 1,126 items, 2,354 pairs). p: Holm-adjusted.

		Own solved pairs			Fixed set	
Backbone	Method	Δ	95% CI	p	Δ	p
Qwen2.5-VL-3B	GRPO	$+5.1$	$[+2.5, +7.6]$	<0.001	$+7.6$	<0.001
	VPPO	$+2.3$	$[-0.1, +4.8]$	0.199	$+5.1$	<0.001
	PRPO	$+2.2$	$[-0.4, +4.8]$	0.199	$+5.8$	<0.001
Qwen2.5-VL-7B	GRPO	$+2.9$	$[+0.8, +5.1]$	0.037	$+3.9$	<0.001
	VPPO	$+0.9$	$[-1.2, +3.0]$	0.406	$+2.7$	0.004
	PRPO	$+2.2$	$[+0.1, +4.3]$	0.160	$+2.0$	0.028
Qwen3-VL-2B	GRPO	$+6.9$	$[+4.6, +9.2]$	<0.001	$+9.0$	<0.001
	VPPO	$+7.5$	$[+5.3, +9.6]$	<0.001	$+8.3$	<0.001
	PRPO	$+4.9$	$[+2.6, +7.2]$	<0.001	$+5.9$	<0.001
Qwen3-VL-4B	GRPO	-4.9	$[-6.7, -3.1]$	<0.001	-11.5	<0.001
	VPPO	$+1.7$	$[+0.1, +3.2]$	0.072	$+2.2$	0.047
	PRPO	-0.8	$[-2.5, +1.0]$	0.381	-0.8	0.381

Responses without a final answer. Base Qwen3-VL gives no final answer on 9–12% of solved pairs against at most 2% after training, mostly through repetition loops or unfinished reasoning, and never by declining to answer. Excluding these responses would bias the two metrics in opposite directions: it raises base edit success, shrinking post-training's gain, and raises base prior lock, shrinking post-training's increase. We therefore count them as failures and report the exclusion bound as a sensitivity check. On the fixed set, which already removes most such responses, excluding them moves base prior lock only from 7.7% to 8.2% (2B) and from 10.5% to 11.0% (4B), so none of the fixed-set conclusions depends on this choice. On each configuration's own solved pairs, prior lock appears to rise under every Qwen3-VL objective, but at 2B this reflects which items qualify: on the fixed set, GRPO and VPPO lower it slightly.

Within-scene contrast. On scenes that carry both edit types, referent retention exceeds numeral retention in every configuration (Table 6), so the dissociation is not a difference between the scenes that receive each edit type.

Table 5: **Original answer kept after an answer-changing edit (%),** on each configuration's solved pairs. Ref.: referent edits (prior lock); Num.: numeral edits. NA: referent pairs with no final answer (Qwen3-VL; below 3% for every Qwen2.5-VL configuration). Δ: change in prior lock from base on the fixed set, with Holm-adjusted p.

Backbone	Method	Ref.	Num.	NA	Δ	p
Qwen2.5-VL-3B	Base	39.5	0.5	–	–	–
	GRPO	35.2	1.0	–	-1.9	1.000
	VPPO	31.7	1.7	–	-2.2	1.000
	PRPO	34.6	1.6	–	-5.6	0.122
Qwen2.5-VL-7B	Base	25.6	0.1	–	–	–
	GRPO	30.5	0.9	–	-1.2	1.000
	VPPO	27.2	1.7	–	$+2.9$	0.549
	PRPO	37.8	1.1	–	$+1.7$	1.000
Qwen3-VL-2B	Base	11.1	0.0	9.9	–	–
	GRPO	14.0	0.5	0.3	-3.4	0.039
	VPPO	14.3	0.3	0.1	-3.6	0.039
	PRPO	15.3	0.4	0.3	$+0.9$	0.269
Qwen3-VL-4B	Base	10.9	0.1	6.6	–	–
	GRPO	23.3	0.2	0.7	$+13.1$	<0.001
	VPPO	15.3	0.1	0.3	$+1.2$	0.225
	PRPO	16.3	0.1	4.1	$+5.4$	<0.001

Table 6: **Within-scene contrast (%).** Scenes with both a referent and a numeral edit, solved on the original image. All contrasts have $p < 0.001$.

Backbone	Method	Scenes	Ref.	Num.	Contrast	95% CI
Qwen2.5-VL-3B	Base	431	40.6	1.6	39.0	[34.2, 43.8]
	GRPO	563	37.2	1.2	36.0	[32.0, 40.2]
	VPPO	514	35.8	2.9	32.9	[28.6, 37.4]
	PRPO	568	35.6	1.6	34.0	[29.8, 38.1]
Qwen2.5-VL-7B	Base	522	25.1	0.2	24.9	[21.2, 28.7]
	GRPO	611	27.4	0.8	26.6	[23.0, 30.2]
	VPPO	574	32.9	3.2	29.8	[25.8, 33.7]
	PRPO	630	33.2	0.5	32.6	[29.0, 36.4]

Composition of referent failures. When a referent edit on a solved pair is not followed, the model most often returns the original answer (Table 7); it gives another value printed in the figure in only 1–8% of failures. For comparison, numeral-edit failures that are neither the new nor the original answer make up 13–19% of solved numeral pairs. Printed-value records cover 96.9% of referent edits (all templates except node degree).

Table 7: **Composition of referent-edit failures (%)** on solved pairs, over failures with a printed-value record (96.9% coverage).

Backbone	Method	Failures	Original	Other printed	Other
Qwen2.5-VL-3B	Base	297	60.3	8.1	31.7
	GRPO	276	73.6	3.3	23.2
	VPPO	250	72.0	4.8	23.2
	PRPO	281	70.1	3.9	26.0
	GRPO (frozen)	282	67.4	2.8	29.8
	VPPO (frozen)	220	72.7	5.5	21.8
Qwen2.5-VL-7B	Base	209	63.6	5.3	31.1
	GRPO	267	62.9	2.3	34.8
	VPPO	232	72.0	1.3	26.7
	PRPO	295	71.5	2.0	26.4

Perception control. Referent masks are smaller than numeral masks (median 5 vs. 32 visual tokens), so a model could miss the moved label rather than fail to bind it. The readout probes (Appendix B.2) separate the two on Qwen2.5-VL. For each locked referent edit (original answer kept), we ask whether the model reads every printed value (reading probe) and whether it names the element that now carries the moved value (binding probe), both on the edited image (Table 8). In every configuration except the 3B base model, most locked edits pass both probes (58–71% for trained 3B models, 79–87% at 7B): the model reads the values and reports where the moved label now sits, yet answers from the original figure. Only for the 3B base model do most locked edits fail the reading probe (51%), which requires listing every number in the figure and is strict.

Table 8: **Readout of locked referent edits (%),** Qwen2.5-VL, edited image. n: locked edits with both probes. Reading: every printed value listed. Binding: the element carrying the moved value named correctly. 95% intervals resample source items.

Backbone	Method	n	Fail reading	Read, fail binding	Pass both
Qwen2.5-VL-3B	Base	108	50.9 [41.8, 59.8]	20.4 [13.2, 28.0]	28.7 [20.6, 37.3]
	GRPO	134	10.4 [5.4, 15.4]	18.7 [12.2, 25.6]	70.9 [63.4, 78.2]
	VPPO	105	21.9 [14.6, 29.4]	20.0 [12.6, 27.9]	58.1 [49.1, 67.0]
	PRPO	123	26.0 [18.6, 33.3]	12.2 [6.6, 18.2]	61.8 [53.3, 70.3]
Qwen2.5-VL-7B	Base	100	11.0 [5.0, 17.0]	4.0 [1.0, 8.0]	85.0 [78.0, 92.0]
	GRPO	97	15.5 [8.3, 22.9]	5.1 [1.0, 10.2]	79.4 [71.1, 86.7]
	VPPO	112	15.2 [8.9, 22.3]	5.4 [1.8, 9.8]	79.5 [71.4, 86.6]
	PRPO	110	11.8 [6.4, 18.2]	0.9 [0.0, 2.7]	87.3 [80.9, 93.6]

D CAUSAL EVIDENCE FOLLOWING: ALL LAYERS

Table 9: **Localized transfer on numeral edits (1,546 pairs),** Qwen2.5-VL-3B layers 0/12/24 and 7B layers 0/9/18. n: originally solved pairs. Acc: transfer accuracy at layer 0, unconditional (Figure 4, light bars). CEF: conditional on a solved original (Eq. 4). Out: equal number of positions outside $S(M)$, layer 0. LS: localization share at layer 0, with 95% interval.

				CEF				
Backbone	Method	n	Acc	early	mid	late	Out	LS
Qwen2.5-VL-3B	Base	819	40.7	62.9	57.9	32.7	0.5	73.0 [69.6, 76.3]
	GRPO	975	45.0	59.2	56.4	37.1	1.5	65.6 [62.3, 68.8]
	VPPO	932	43.9	58.4	54.8	38.2	2.5	65.0 [61.5, 68.4]
	PRPO	957	44.6	59.4	56.6	37.7	1.9	65.6 [62.2, 68.9]
	GRPO (frozen)	972	46.4	59.6	58.8	37.3	0.6	67.8 [64.4, 71.1]
	VPPO (frozen)	860	41.4	60.2	55.5	37.2	1.7	67.5 [64.0, 71.0]
	PRPO (frozen)	976	46.5	62.1	61.4	38.8	1.2	70.1 [66.9, 73.2]
Qwen2.5-VL-7B	Base	977	45.9	60.6	57.4	42.8	0.4	68.4 [65.2, 71.5]
	GRPO	1081	52.4	65.5	62.2	45.0	0.2	71.5 [68.6, 74.3]
	VPPO	1070	50.9	63.2	60.5	45.6	2.0	69.8 [66.9, 72.9]
	PRPO	1096	54.1	66.9	61.0	45.2	0.4	72.7 [69.9, 75.6]
	GRPO (frozen)	1088	51.9	63.6	61.9	47.3	1.3	71.1 [68.2, 74.0]

Gain share. Table 10 gives the gain share for each group. All three check values used in the main text reproduce, and the denominator is positive in every resample. At 7B, frozen and trainable GRPO carry the same share of their gain at the edit.

Localization share. Conditioning on each configuration's solved and followed pairs gives the same directions as the gain share. At 3B, the localization share falls by 7.4–8.0 points under every objective, with every interval excluding zero; with a frozen tower it still falls by 5.1 (GRPO, $[-8.9, -1.2]$) and 5.5 points (VPPO, $[-9.5, -1.4]$), and by 2.9 points for PRPO ($[-6.6, 0.9]$), so part of PRPO's shift depends on the tower. At 7B, frozen GRPO changes from base by about as much as trainable GRPO. On the 1,545 numeral pairs with a solved original, the 7B localization

Table 10: **Gain share at layer 0** (numeral edits, 1,546 pairs), with 95% intervals.

Group	ρ (%)	95% CI
Qwen2.5-VL-3B, trainable (pooled)	42.7	[17.8, 66.7]
Qwen2.5-VL-3B, frozen (pooled)	54.1	[27.3, 83.5]
Qwen2.5-VL-7B, trainable (pooled)	96.5	[66.4, 139.2]
Qwen2.5-VL-7B, GRPO trainable	99.0	[61.9, 157.5]
Qwen2.5-VL-7B, GRPO frozen	103.3	[63.0, 173.5]
7B frozen minus trainable	$+4.3$	$[-41.6, +59.2]$

share rises from 68.4% by $+3.2$ (GRPO, $[-0.1, 6.4]$), $+1.5$ (VPPO, $[-1.8, 4.8]$), $+4.3$ (PRPO, $[1.2, 7.6]$), and $+2.7$ points for frozen GRPO ($[-0.4, 5.9]$); only the PRPO interval excludes zero.

Encoder locality. For each pair we compare the layer-0 visual states of the edited and original runs and compute the share of the squared change that falls inside the edit,

$$L = \frac{\sum_{j \in S(M)} \|H_j^0(x', q) - H_j^0(x, q)\|^2}{\sum_{j \in V} \|H_j^0(x', q) - H_j^0(x, q)\|^2},$$

where V is the set of all visual positions, and average L over numeral pairs. L is 29.0% for the Qwen2.5-VL-3B base model and 28.7–29.4% for its trained configurations, and 26.6% for Qwen2.5-VL-7B base and 26.1–26.7% trained, while $S(M)$ covers about 9% of visual tokens. Because L is well below one, part of each edit's effect lies outside $S(M)$, which is why localized transfer is a lower bound.

E Pathway interventions and attention

Table 11: **Noise-image interventions, 1,000 items (%).** Attn: attention share from answer tokens to visual tokens, middle layers. V-rec: original prediction restored by clean visual states in a noise-image run. Q-corr: original prediction kept after question states are replaced by noise-run states. Q-rec: original prediction restored by clean question states. Floor: original prediction kept in the noise-image run with no patch. Mid and late are layers 12/24 (3B) and 9/18 (7B); at layer 0 every intervention restores or keeps the prediction in every configuration. PRPO with a frozen tower was not run.

			V-rec		Q-corr		Q-rec		
Backbone	Method	Attn	mid	late	mid	late	mid	late	Floor
Qwen2.5-VL-3B	Base	5.38	46.2	16.0	52.7	51.7	6.8	6.4	6.6
	GRPO	5.69	29.8	10.1	36.7	37.5	2.6	4.9	3.8
	VPPO	5.63	16.4	4.1	22.5	23.1	2.7	2.5	3.5
	PRPO	5.98	31.7	9.2	39.1	38.7	3.1	4.7	2.7
	GRPO (frozen)	5.45	32.1	13.3	38.2	37.9	5.3	5.7	4.8
	VPPO (frozen)	4.75	15.1	3.6	24.2	24.5	2.2	3.1	1.4
Qwen2.5-VL-7B	Base	4.73	43.7	20.8	52.0	52.1	8.7	6.5	6.8
	GRPO	5.44	46.3	17.6	52.9	55.2	5.7	6.6	6.3
	VPPO	4.86	16.2	5.1	22.8	23.8	1.4	3.4	1.5
	PRPO	5.26	42.0	18.6	47.0	47.3	3.7	7.1	5.1

At 3B, visual-state recovery and question-state retention fall under every objective. At 7B, both fall substantially only under VPPO (by about 28 points), while GRPO and PRPO stay within 6 points of base; attention to visual tokens nevertheless rises under all three objectives at both sizes, so attention share does not track the causal change at either size. Restoring question states alone recovers the original prediction about as often as the unpatched noise-image run keeps it (Q-rec vs. Floor).

Certified-donor transfer. The edited image's states are copied into the original-image run at the entry to the middle decoder layer (12 for 3B, 9 for 7B) at all visual positions (T_V), question-text

positions (T_Q), or every prompt position (T_{all}), and the prediction is scored against the certified new answer among solved pairs. Because layers below the patch keep the recipient's cache, T_{all} is the ceiling at that depth. The harness reproduces the committed CEF outputs exactly (50 of 50 pairs, two configurations), and tower status is verified from weight fingerprints.

Table 12: **Certified-donor transfer at middle depth (%).** Share of solved pairs whose prediction becomes the certified new answer.

		Value edits			Referent edits		
Backbone	Method	T_V	T_Q	T_{all}	T_V	T_Q	T_{all}
Qwen2.5-VL-3B	Base	66.2	1.9	65.6	34.5	0.7	34.1
	GRPO	70.1	1.7	69.3	55.9	4.9	56.7
	VPPO	69.2	1.9	67.9	55.6	7.8	55.6
	PRPO	67.8	1.5	67.5	52.2	4.6	51.0
	GRPO (frozen)	70.4	0.8	69.5	50.0	2.4	52.2
	VPPO (frozen)	65.6	1.0	65.8	54.8	6.4	54.4
	PRPO (frozen)	69.6	0.4	69.6	51.4	1.9	51.7
Qwen2.5-VL-7B	Base	71.5	0.8	71.7	63.3	0.9	63.3
	GRPO	72.8	0.5	73.2	58.9	1.0	60.3
	VPPO	71.3	1.6	72.7	60.6	7.9	60.1
	PRPO	71.6	0.7	71.2	51.2	0.8	53.3
	GRPO (frozen)	71.8	1.0	72.2	58.7	1.8	58.0

F STANDARD BENCHMARKS

Table 13: **Sighted / blind accuracy (%) on standard benchmarks and CARVE.** Held-out: ViRL39K held-out split. Qwen3-VL MMK12 uses a blank image; other blind evaluations omit it. Base Qwen3-VL often gives no final answer without the image, so its blind values are not comparable with the trained configurations. Qwen2.5-VL-7B PRPO emits one fixed answer on 11% of Geo3K items (65 of 601 sighted, 40 blind), none of them correct. Scores use the evaluation scorer throughout; stricter per-item matching of LaTeX fractions lowers individual cells by up to 3 points, similarly across configurations.

		Geo3K		Held-out		MMK12		CARVE	
Backbone	Method	S	B	S	B	S	B	S	B
Qwen2.5-VL-3B	Base	29.3	10.3	41.9	27.3	38.0	34.0	37.6	3.2
	GRPO	35.8	11.8	62.2	42.9	63.0	56.8	44.6	4.8
	VPPO	36.9	12.5	59.3	44.3	59.2	54.2	42.4	5.8
	PRPO	35.4	12.0	59.3	42.5	60.4	55.4	45.4	4.9
Qwen2.5-VL-7B	Base	40.6	10.5	53.4	33.1	43.0	38.8	42.0	5.1
	GRPO	44.4	12.8	68.6	47.2	75.8	67.0	49.0	5.5
	VPPO	40.6	11.8	68.2	45.1	78.6	68.8	47.9	5.4
	PRPO	24.1	13.8	75.0	50.1	76.6	70.6	49.4	5.7
Qwen3-VL-2B	Base	30.4	4.3	43.1	20.7	33.6	20.6	42.4	4.9
	GRPO	47.8	11.6	66.7	43.6	56.0	48.6	50.2	4.0
	VPPO	43.6	12.6	67.9	43.9	56.0	48.4	51.8	4.7
	PRPO	45.4	10.5	67.0	42.7	56.6	44.2	50.7	4.8
Qwen3-VL-4B	Base	49.1	10.8	57.2	30.0	46.0	36.2	58.9	10.5
	GRPO	56.7	14.3	76.5	50.5	67.0	53.6	57.7	5.3
	VPPO	64.6	15.5	82.6	52.7	74.4	61.6	62.1	5.4
	PRPO	57.1	13.3	77.9	49.2	67.4	56.0	62.2	5.5

G QUALITATIVE EXAMPLES

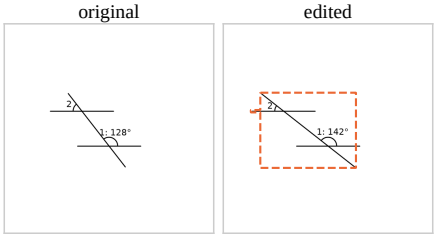
The examples below are chosen by a fixed rule rather than by inspection. Within each group, candidates are ordered by the SHA-256 hash of their identifier and the first eligible one is taken; no candidate was skipped. Images are the benchmark's native renders (532×532), shown without re-rendering.

Benchmark items. Figures 8 and 9 show one item per rendering style and evidence mode, each with one answer-changing edit. Value rewrites on MODE-L templates redraw the geometry the numeral annotates, so their masks cover more than the label (Appendix A.2).

Numeral versus referent edits. Figure 10 shows both edit types on the same scene for Qwen2.5-VL-7B Base and GRPO. Rows 1–3 are the first three scenes, from distinct templates, where a model solves the original, follows the numeral edit, and returns the original answer after the referent edit; row 4 is the first scene where both models follow both edits. In row 3, the referent edit moves the $42°$ label from A to C, so the certified answer becomes 42. Both models answer 82, the original answer. GRPO passes both readout probes, listing the printed values and naming C as the angle labeled $42°$, and its response states the new binding before answering from the old one: "The given angles are $\angle B = 56°$ and $\angle C = 42°$. Let's denote $\angle A$ as the third angle … The measure of angle C is already given as $42°$. Therefore, the measure of angle C is: 82." In rows 1 and 2, GRPO follows the numeral edit, reads the moved value correctly (probe A), and still returns the original answer after the referent edit.

What localized transfer copies. Figure 11 shows the visual-token positions $S(M)$ transferred by CEF for one numeral edit. On this item neither 3B model answers the edited image correctly, with or without transfer; the example illustrates the intervention, not its success rate.

original edited

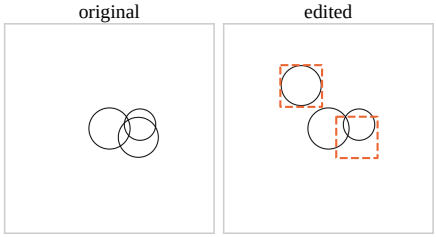


geo3k/Mode-L · T2_parallel_transversal

Two parallel lines are cut by a transversal. Find the measure of angle 2, in degrees.

52 → 38 (numeral edit: value rewrite)

original edited

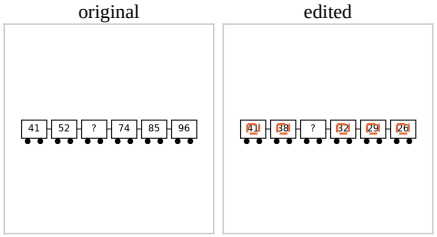


geo3k/Mode-G · T6_circle_intersection

How many points do the circles intersect at, counting all pairs?

6 → 2 (structural edit: geometry edit)

original edited

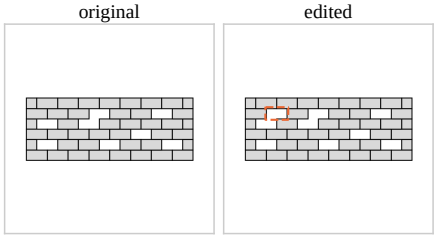


mathvision/Mode-L · T55_number_train

The numbers on the wagons increase or decrease by the same amount from one wagon to the next. What number belongs ...

63 → 35 (numeral edit: value rewrite)

original edited



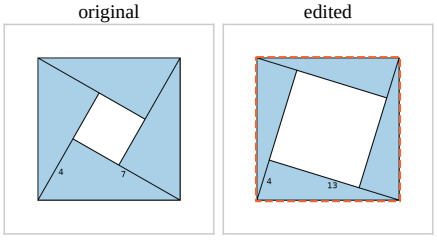
mathvision/Mode-G · T13_brick_wall

How many bricks are missing from the wall?

8 → 9 (structural edit: geometry edit)

Figure 8: CARVE-MIXED items, Geo3K and MathVision styles. For each item: original image, edited image with the edit mask outlined (dashed), question, and certified answer change $y \to y'$.

original edited

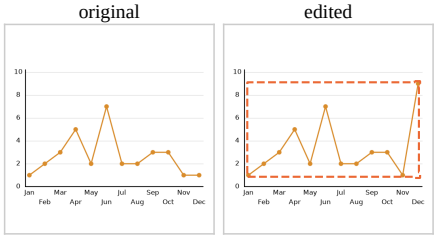


dynamath/Mode-L · T52_inner_square_area

Four identical right triangles are placed inside the square, leaving a smaller white square in the middle. T...

9 → 81 (numeral edit: value rewrite)

original edited

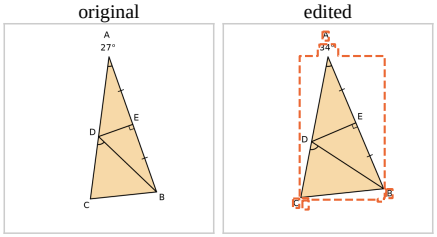


dynamath/Mode-G · T60_hottest_month

The chart shows a value for each month of the year. Which month has the highest value? Answer with the month's na...

Jun → Dec (structural edit: geometry edit)

original edited

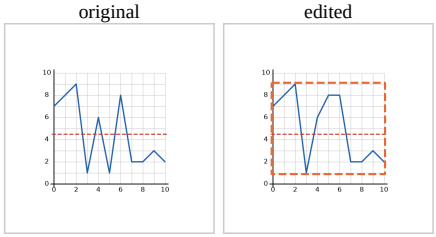


mathverse/Mode-L · T61_isosceles_bisector

In triangle ABC, AB = AC. The perpendicular bisector of AB meets AB at E and AC at D. Find the measure of angle BD...

54 → 68 (numeral edit: value rewrite)

original edited

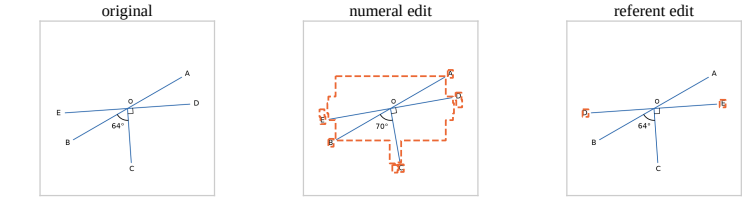


mathverse/Mode-G · T38_curve_crossings

How many times does the graph cross the dashed horizontal line?

5 → 3 (structural edit: geometry edit)

Figure 9: CARVE-MIXED items, DynaMath and MathVerse styles. Layout as in Figure 8.



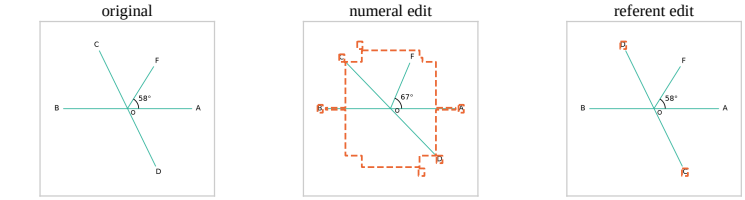
answer: 26 | answer: 20 | answer: 154

7B Base: 116 ✗ | 7B Base: 130 ✗ | 7B Base: 116 ✗

7B GRPO: 26 ✓ | 7B GRPO: 20 ✓ | 7B GRPO: 26 ✗

T34_perpendicular_mark · e2_T34_000027 · prior lock (GRPO)

Probe A reads 64 (printed: 64): yes. Probe B not defined for this template.



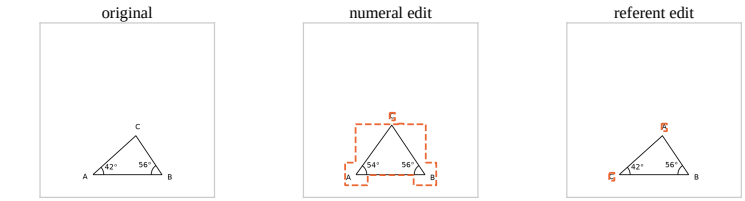
answer: 64 | answer: 46 | answer: 116

7B Base: 122 ✗ | 7B Base: 146.5 ✗ | 7B Base: 122 ✗

7B GRPO: 64 ✓ | 7B GRPO: 46 ✓ | 7B GRPO: 64 ✗

T33_intersecting_bisector · e2_T33_000056 · prior lock (GRPO)

Probe A reads 58 (printed: 58): yes. Probe B not defined for this template.



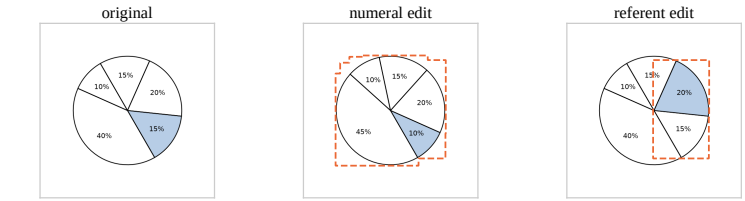
answer: 82 | answer: 70 | answer: 42

7B Base: 82 ✓ | 7B Base: 70 ✓ | 7B Base: 82 ✗

7B GRPO: 82 ✓ | 7B GRPO: 70 ✓ | 7B GRPO: 82 ✗

T1_triangle_angle_sum · e2_T1_000022 · prior lock (Base, GRPO)

Probe A reads 42, 56 (printed: 42,56): yes. Probe B — Which angle is labeled 42°? Answer with the vertex letters only. — C: yes.



answer: 54 | answer: 36 | answer: 72

7B Base: 54 ✓ | 7B Base: 36 ✓ | 7B Base: 72 ✓

7B GRPO: 54 ✓ | 7B GRPO: 36 ✓ | 7B GRPO: 72 ✓

T49_pie_slice_angle · e2_T49_000033 · both edits followed

Figure 10: Numeral and referent edits on the same scene (Qwen2.5-VL-7B). Each row shows the original image, a numeral edit, and a referent edit with its certified answer; dashed outlines mark the edit mask, and marks indicate correct and incorrect predictions. Rows 1–3: first three prior-lock scenes in hash order; row 4: first scene where both models follow both edits. Probe results are for GRPO.

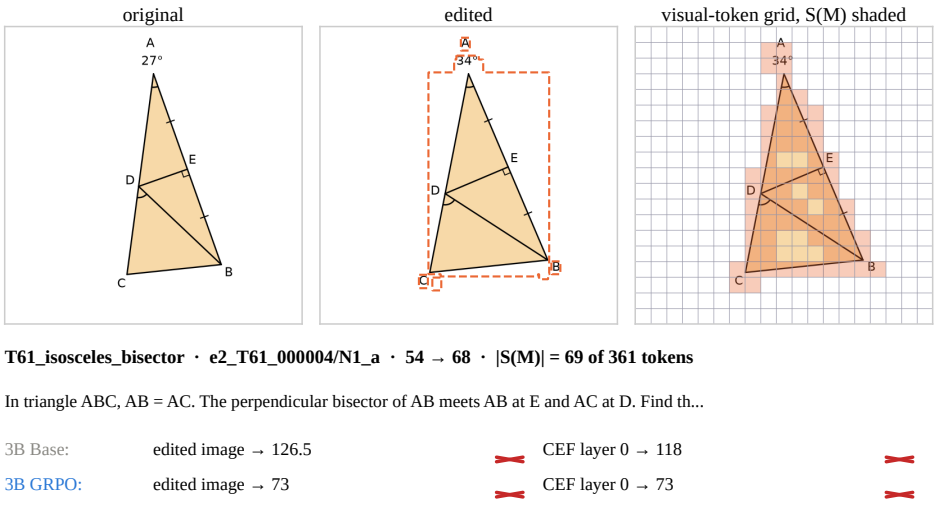


Figure 11: Positions transferred by CEF (Qwen2.5-VL-3B, layer 0). Right: the visual-token grid with the positions $S(M)$ that overlap the edit mask shaded (69 of 361 tokens). CEF replaces the original-image states at these positions with the edited-image states.