SAME REWARD, DIFFERENT SKILLS: WHEN MULTIMODAL RL LEARNS TO LOOK

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) improves vision-language benchmark scores even without visual information during training. With images at test, blind-trained models recover roughly half of the real-image gain at 3B and nearly four fifths at 7B. Prolonged real-image training can erode grounding while benchmark gains persist. Both findings expose the same gap: an image in the prompt is not an image in the learning signal. Our design rule, visual resolvability, asks that visual evidence be necessary for a correct answer and that the task remain learnable. We test it on counterfactual coordinate scenes in which the question stays fixed and the target is never named, so a correct answer requires finding the target in the image. With standard GRPO and correctness-and-format rewards, a 7B model raises its accuracy at finding the target (discovery) from 0.425 to 0.875 on held-out scenes denser than any it trained on, and it improves on question types it never trained on. Two controls locate the source of the gain. Replacing test images with gray canvases drops discovery to zero; training on gray canvases instead, at matched step 30 and in each of four seeds, yields essentially none of the gain even when the model is then tested with real images. The learned skill carries over to grounding tasks built independently of the training corpus. A caption that answers the training question, added to the same images, reward and budget, cuts the gain by nearly two thirds. Changing what reward requires changes what RL learns.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR; Lambert et al., 2025; Guo et al., 2025) improves pretrained vision-language models by rewarding correct answers to image-based questions (Huang et al., 2026; Meng et al., 2025; Wang et al., 2025a). Gains on visual reasoning benchmarks (Lu et al., 2024; Zhang et al., 2024) are read as improvements in perception as well as reasoning. This interpretation guides data curation, reward design, and checkpoint selection. It rests on an assumption: rewarding correct answers develops the visual operations needed to produce them.

On Geometry3K (Lu et al., 2021), replace every training image with a gray canvas, and the model still recovers roughly half of the real-image training gain at 3B and nearly four fifths at 7B when evaluated with real images. These are **blind gains**, learned without visual information during RL. Keeping the images is no guarantee either: prolonged real-image training with a trainable vision encoder can erode the ability to locate and read visual targets while benchmark gains persist—**grounding corrosion** (cf. Phan et al., 2026). Independent blind-trained runs lose many of the same previously solved examples, usually with identical wrong answers.

The design problem is therefore not just what the prompt contains, but what earning reward requires. **Shortcut-accessible** tasks permit correct answers without visual evidence. **Visual resolvability** is the alternative design rule: require visual evidence for correct answers while keeping the task learnable. The aim is to make the intended visual operation part of earning reward, rather than something the learner can bypass (Figure 1).

We put this rule into practice through two task controls: the visual fact and the target cue.[1] Counterfactual scenes keep the question fixed while changing one visual fact. Target cues range from a

[1] Code and data: https://anonymous.4open.science/r/learning-without-looking-2432

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.

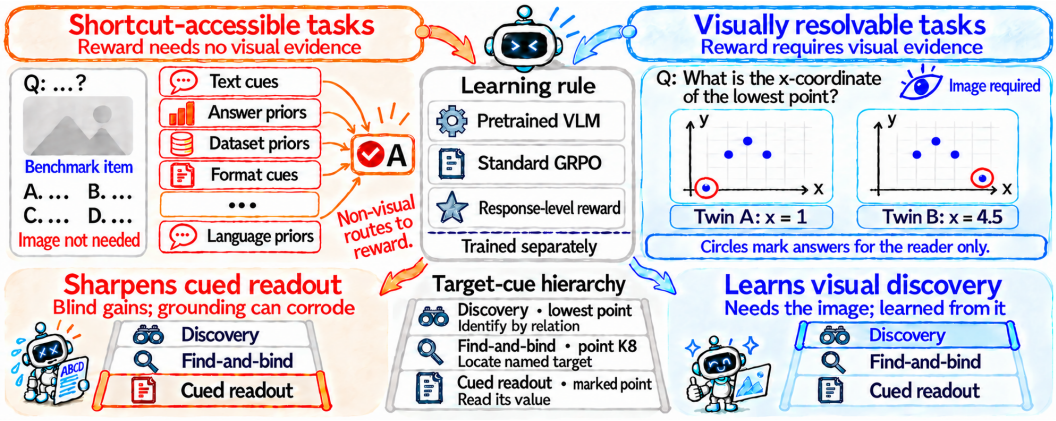


Figure 1: *Same learning rule, different lessons.* Left: shortcut-accessible tasks allow reward without visual evidence; gains in the reported 3B comparisons mainly favor cued readout. Right: constructed tasks demand visual evidence and relation-based target identification; discovery, cued readout, and named-target grounding improve. The hierarchy distinguishes supplied locations, supplied identities, and relation-defined targets. GRPO is shared across the study; rewards stay fixed within matched comparisons.

marked location or name to a relation such as "the lowest point," which leaves the target to be identified (discovery). A base-model audit measures visual benefit and learnability. We train on relation-only coordinate questions and target-identification probes using standard GRPO (Shao et al., 2024) with correctness-and-format rewards.

On a separate confirmatory set of 20-point scenes—a density absent from training—discovery accuracy rises from 0.425 to 0.875. Two visual-access controls separate needing images from learning through them: replacing test images with gray canvases reduces discovery accuracy to zero; at matched step 30, in each of four seeds, a blind-trained control evaluated with real images recovers essentially none of the discovery gain. The gains extend to question types never trained—cued readout and named-target grounding on the same scenes—and to independently constructed coordinate-grounding tasks. Mixing constructed with ordinary data turns the discovery gain up or down. The full study spans nine counterfactual task families, two public training corpora, seven audited benchmarks, and over thirty GRPO runs (Appendix K).

Whether RLVR develops capabilities or makes existing ones easier to elicit remains open (Yue et al., 2025b; Shao et al., 2026). We ask which visual operations improve and whether training images account for the gain. Related approaches modify the objective (Wang et al., 2026), response structure (Xia et al., 2026), available evidence (Zheng et al., 2026), reward (Xiao et al., 2026; Yu et al., 2025a; Liu et al., 2025), or data selection (Yu et al., 2025b; Wang et al., 2025b). We use standard GRPO and hold rewards fixed within each matched visual-access comparison; task construction specifies the evidence and operations that successful answering requires. Difficulty-based selection asks how hard a lesson is; resolvability also asks what evidence it requires. Appendix L gives the full related work.

2 Benchmark Gains Are Not Visual Acquisition

2.1 Learning Without Looking

We run 100 steps of GRPO from pretrained Qwen2.5-VL-3B-Instruct (Bai et al., 2025) on filtered Geometry3K and evaluate on its held-out test set, with the vision encoder frozen and three seeds per condition (Appendix D). Four conditions differ only in what accompanies each question during RL: the real image (REAL), image-derived text (CAPTION), no image (NONE), or a uniform gray canvas (GRAY).

Table 1 crosses training-time and test-time visual access. With real test images, REAL gains +0.238 over the base model, NONE +0.137, and GRAY +0.125. Let g_c be condition c's gain under real-image evaluation; its recovery share is g_c/g_{REAL}: 0.57 for NONE and 0.52 for GRAY. Yet the gains depend

Table 1: *Blind gains.* Training without visual information recovers more than half of REAL's benchmark gain when real images are supplied at test; with the test image removed, no condition gains more than 0.05. Base row: accuracy; other cells: gain over that base under the same evaluation; shares: fraction of REAL's gain (three-seed pooled). All cells use the same answer matcher.

	Evaluated with the real image		Test image removed	
Trained with	gain over base	share of REAL's gain	gray canvas	no image
base model (accuracy)	*0.175*	—	*0.090*	*0.068*
REAL	+0.238	1 (reference)	+0.019	+0.036
CAPTION	+0.171	0.72	+0.024	+0.037
trained without visual information				
NONE	+0.137	**0.57**	+0.015	+0.044
GRAY	+0.125	**0.52**	+0.014	+0.032

At 7B, GRAY recovers 0.78 [0.64, 0.92] of REAL's gain (Table D.4).

on images at test: remove the test images and every condition gains little (+0.036 for REAL, +0.032 for GRAY). Blind gains also persist on the 497 questions for which the base model already produces an extractable answer: GRAY and NONE recover 51% and 54% of REAL's gain (Table D.5). At 7B, GRAY recovers 0.78 (95% CI [0.64, 0.92]). An improvement can therefore depend on an image at test without having been learned from images.

2.2 How Much Does the Image Matter?

Raw blind-to-image accuracy ratios credit guessing and can blur differences in visual dependence. We audit the base 3B model using chance-corrected blind retention, $\rho_{\text{blind}} = (a_{\text{blind}} - c)/(a_{\text{image}} - c)$ for $a_{\text{image}} > c$, where c is the chance reference: the mean of $1/k$ over multiple-choice items with k options, and zero for free-form answers. BLINK (Fu et al., 2024) and the multiple-choice portion of MathVerse (Zhang et al., 2024) look alike under the raw ratio (0.83 and 0.85), but separate to 0.27 and 0.66 after correction. Audits of two other model families show a sharp contrast: without images, Gemma-3 (Gemma Team, 2025) keeps about three quarters of its free-form accuracy on ordinary questions from the RLVR corpus ViRL39K (Wang et al., 2025a), and InternVL3-9B (Zhu et al., 2025) about half, yet both score exactly zero on the grounding task below (Appendix A). The audit asks whether the image matters; to see what RL changes, we separate the operations used to answer.

2.3 Reading a Value, Finding a Target

We distinguish tasks by how much target information is supplied. **Cued readout** marks the target's location and asks for its value. **Find-and-bind** names the target—a point in a coordinate plot—and requires locating it and reading its coordinate. **Discovery** supplies only a relation, such as "the point with the smallest y-coordinate," and requires identifying the target before reading it. The cue hierarchy orders the information supplied, not model accuracy.

We measure cued readout and find-and-bind on a counterfactual suite: a *grounding task* of 600 named-point coordinate pairs, a *cued-readout task* of 300 marked-plot pairs, and a *header-cued table task* of 300 pairs. Each pair keeps the question fixed and changes one visual fact so that its two correct answers differ. For a predictor h, let $C_{j,s}(h) \in \{0, 1\}$ indicate correctness on member $s \in \{a, b\}$ of pair j. Pair accuracy (Tong et al., 2024; Li et al., 2024) over N_{pair} pairs is

$$\text{PairAcc}(h) = \frac{1}{N_{\text{pair}}} \sum_{j=1}^{N_{\text{pair}}} C_{j,a}(h)\, C_{j,b}(h). \tag{1}$$

An image-independent greedy predictor returns the same extracted answer on both members. Their accepted answers are disjoint, so every product in Eq. (1) is zero (Proposition J.1). The grounding task has an independently regenerated twin, and human review confirmed all 60 audited pairs of the suite (Appendix C).

On this suite, Geometry3K training improves cued readout under every visual-access condition: REAL gains +0.143, GRAY +0.138, and NONE +0.108 (Figure 2(a)). Cued readout accounts for

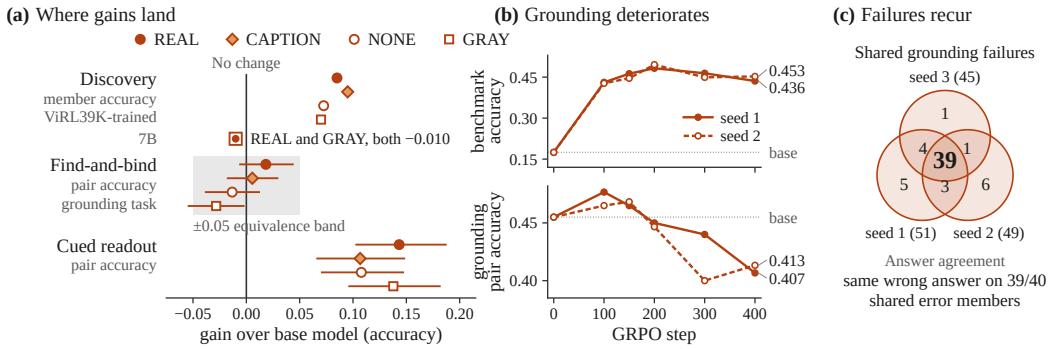


Figure 2: *Benchmark gains are not visual acquisition.* (a) Gains by operation: readout and grounding for Geometry3K-trained 3B models (three-seed means), discovery for ViRL39K-trained models (two-seed means; 7B: Geometry3K-trained pair). (b) Benchmark accuracy and grounding pair accuracy under prolonged REAL training with a trainable encoder. (c) Grounding pairs lost by three GRAY-trained runs, and their wrong answers. All evaluations use real images; error bars are 95% intervals and shading marks the grounding equivalence band.

72% of REAL's total pair-count-weighted improvement across the 1,200 pairs. On the grounding task, REAL's gain of +0.018 is equivalent to zero within the pre-specified ± 0.05 band (90% CI $[-0.003, +0.040]$; Appendix I), and GRAY declines by 0.028. Reading a marked value improves; locating a named target shows little change.

Blind gains also reach discovery. For 3B models trained under the four visual-access conditions on ViRL39K, discovery gains range from about +0.04 to +0.10, and NONE and GRAY finish within 0.04 of REAL at every density (two-seed means). At 7B, Geometry3K training barely changes 20-point discovery (0.470 $\rightarrow$ 0.460 under REAL and GRAY). In these ordinary 3B runs, gains mainly favor cued readout, and blind training reproduces most of the discovery gain.

3 When Gains Conceal Grounding Loss

Under the long-horizon real-image recipe below, grounding does more than stall: it falls below the base model's while benchmark gains persist.

Long-horizon training. We train Qwen2.5-VL-3B-Instruct for 400 GRPO steps on unfiltered Geometry3K, with real images throughout, a trainable vision encoder, and correctness-and-format rewards, in two independent runs evaluated at step 400. Benchmark accuracy rises from 0.175 to 0.436 and 0.453, while grounding-task pair accuracy, after a small rise, falls by 0.070 and 0.052 from step 100 (95% CIs exclude zero) to 0.407 and 0.413, below the base model's 0.455 from step 200 onward (Figure 2(b)). The first run's near-flat benchmark score from step 100 to 400 hides item-level churn (Appendix E). We call this loss of previously available visual grounding under reward optimization *grounding corrosion*.

Recurring errors. In a separate analysis, the three GRAY-trained 3B runs from §2.1, evaluated with real images, each lose 45–51 grounding-task pairs that the base model solved. Their degraded sets overlap with a Jaccard index of 0.661, against a permutation-null mean of 0.011 for sets of the same sizes (Figure 2(c)). Among the degraded pairs, 40 members are wrong in all three runs, with identical answers on 39: the runs agree not only on what fails, but on what replaces the correct answer. Those 39 answers span 11 distinct values. On 29 of these 39 members, at least one REAL-trained run gives the same wrong answer: these item-structured errors recur across visual-access conditions. Whatever produces them, it is not the training images.

Grounding at 7B. The larger model shows a different balance: real-image training improves grounding even though much of the benchmark gain remains available to blind training. In the ordinary 7B access comparison, REAL gains +0.035 (95% CI [0.013, 0.058]) on the grounding task and +0.038 on its twin. GRAY's changes (+0.012 and +0.013) have intervals that include zero, as do both models' changes on cued readout (Table D.3).

4 THE MISSING LINK: WHAT REWARD REQUIRES

Images help at test. The puzzle is whether earning reward makes the model learn from them. Our account is that optimization can strengthen nonvisual answer strategies when these suffice to earn reward; the resulting policy may still exploit images supplied at test. Such a policy has not failed to look; nothing required it to. A target-location cue offers another shortcut: it removes search while leaving the value to be read from the image. Needing the image is not the same as needing to discover what to read.

Two properties guide task construction. **Visual necessity** ties correct answers to visual evidence rather than an image-free shortcut. **Learnability** leaves room for informative correctness feedback: when responses are almost uniformly right or wrong, there is little variation in correctness to guide learning. Together, these properties define a **visually resolvable** training problem. The *Visual Resolvability Principle* is a design rule: build learnable tasks whose successful solutions require the visual operation we want to develop. We instantiate it in counterfactual coordinate scenes, where both properties can be measured and controlled directly.

5 MAKE LOOKING NECESSARY: MEASURE, SHAPE, VERIFY

Two controls specify what the task demands: which visual fact determines the answer, and what the prompt reveals about its target. A base-model audit measures the contribution of the image and the opportunity for correctness feedback.

5.1 MEASURE: VISUAL NECESSITY AND LEARNABILITY

We audit the base Qwen2.5-VL-7B-Instruct model before RL: 16 responses per item with its image and 16 without, at temperature 1, scored by the training reward's accuracy component. Let b_i and $q_{\text{blind},i}$ be these empirical success rates. Our empirical visual-necessity measure, $\Delta q_i = b_i - q_{\text{blind},i}$, estimates how much the image raises correct-answer probability; the corpus mean of $q_{\text{blind},i}$ is its *blind reward opportunity*.

The audited learnability band, $0 < b_i < 1$, means that both correct and incorrect responses were observed. At a fixed prompt and checkpoint, let p_i be the true correctness probability of an independently sampled response. A group of $G = 5$ responses contains both outcomes with probability $P_{\text{mix}}(p_i; G) = 1 - p_i^G - (1 - p_i)^G$; the subtracted terms are the all-correct and all-incorrect cases. The band thus targets opportunities for accuracy-based reward variation (Appendix J.2). This fixed audit describes the corpus and supplies the mixture measurements; it neither filters nor reweights training. The constructed corpus has blind reward opportunity 0.003, and 92.9% of its rows satisfy $\Delta q_i > 0$ and $0 < b_i < 1$ (Figure 3(a)).

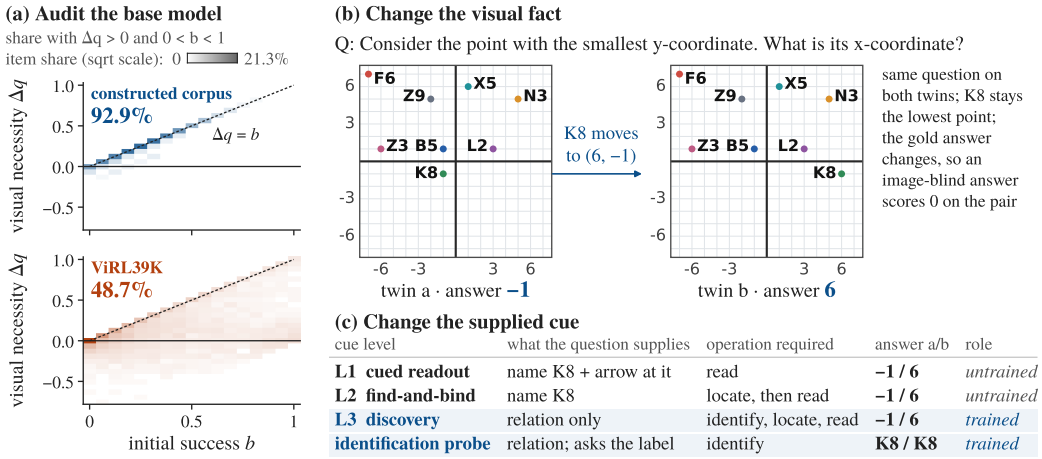
5.2 SHAPE: CHANGE THE FACT, CHANGE THE CUE

Each scene program specifies labeled points, their coordinates, and the relations from which questions and exact answers derive. Rendering a program's two twins changes one answer-relevant fact while the question and the rest of the scene stay fixed.

Changing the visual fact. In the eight-point scene of Figure 3(b), the question asks, "Consider the point with the smallest y-coordinate. What is its x-coordinate?" The target, K8, lies at $(-1, -1)$. Its counterfactual twin moves K8 to $(6, -1)$: K8 remains the lowest point, but the answer changes from -1 to 6. The question is unchanged, and solving the pair requires both answers to be correct (Eq. (1)).

Changing the supplied cue. On the same scene, cued readout (L1) names K8 and marks it with an offset arrow; find-and-bind (L2) names K8 without the arrow; discovery (L3) gives only the relation. All three ask for the same coordinate; only the supplied target information changes. Discovery requires identifying the relevant point before reading its coordinate. A separate **identification probe** asks for the label directly: "Which labeled point has the smallest y-coordinate?"

The corpus balances three pair roles: *target-stable* pairs move the target, as in the K8 scene; *target-switch* pairs change which point satisfies the relation; and *invariance* pairs change the image without



(c) Change the supplied cue

cue level	what the question supplies	operation required	answer a/b	role
L1 cued readout	name K8 + arrow at it	read	**−1 / 6**	*untrained*
L2 find-and-bind	name K8	locate, then read	**−1 / 6**	*untrained*
L3 discovery	relation only	identify, locate, read	**−1 / 6**	*trained*
identification probe	relation; asks the label	identify	**K8 / K8**	*trained*

Figure 3: *Measure and shape.* (a) Audited visual benefit against initial success, constructed corpus and ViRL39K. (b) The K8 scene and its counterfactual twin. (c) Its target-cue hierarchy.

changing the answer. Answers must follow relevant changes and ignore irrelevant ones. From each of 720 scene programs we take the discovery question and probe on both twins, giving 2,880 training prompts on 8- and 12-point scenes. The 20-point scenes are reserved for evaluation; L1 and L2 are evaluated but never trained. Discovery accuracy uses target-stable and invariance pairs, with target-switch pairs reported separately.

5.3 Verify: Does the Test Require Looking?

We track learning on a fixed development set and assess acquisition on a separate confirmatory set. Five criteria qualify each scene density: discovery neither trivial nor unreachable for the base model; discovery harder than both cued levels; zero image-removed accuracy on discovery and its probe; question-blind captions insufficient to answer; and failure of the specified linear classifiers to distinguish the twins. We set the numerical thresholds before generating 450 new scene programs, 150 per density, disjoint from training and development. The 20-point subset meets all five and supplies the confirmatory result; Appendix B reports every density's outcome on both sets and the protocol timing.

5.4 Train: Standard GRPO

We train the 7B base model with standard GRPO. The vision encoder and projector stay frozen; only the language model is trained. For each image–question prompt we sample five responses at temperature 1 and score each response o independently against the gold answer y:

$$r(o, y) = \frac{1}{2}\,\mathbf{1}\{\text{correct}(o, y)\} + \frac{1}{2}\,\mathbf{1}\{\text{format}(o)\}. \tag{2}$$

Correctness tests the extracted answer for equivalence to y; format requires exactly one nonempty `<answer>...</answer>` span. GRPO normalizes rewards within each prompt's five responses. This reward stays fixed across the constructed-corpus visual-access and mixture comparisons. Twins are separate training rows with separate response groups: pairing shapes the data; rewards remain response-level. Training spans 100 steps in four segments; restarts reset AdamW moments, re-anchor the KL reference, and restart the corpus order. All matched 30-step comparisons end before the first restart.

6 The Regime Flips

Training on relation-only questions and identification probes improves discovery and the cued tasks on the same held-out scenes.

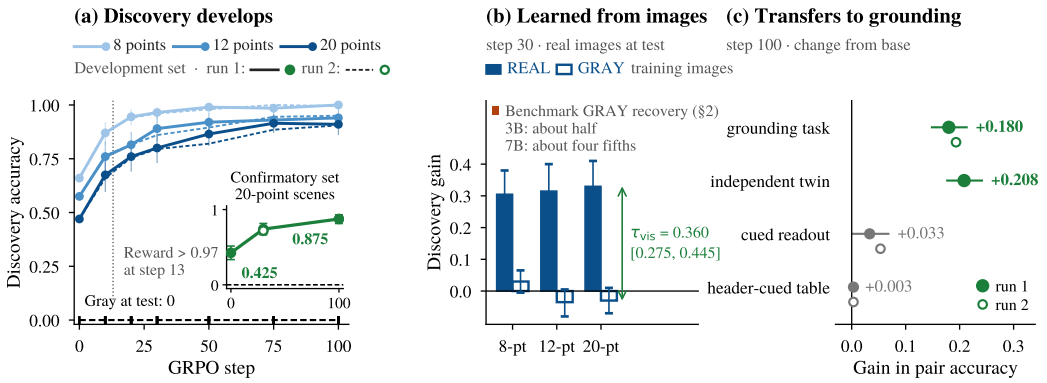


Figure 4: *The regime flips.* (a) Discovery member accuracy on the development set (inset: confirmatory 20-point result). (b) Step-30 discovery gains under REAL and GRAY training, evaluated with real images. (c) Step-100 change in pair accuracy. Error bars: 95% intervals over scene programs in (a, b) and over pairs in (c).

6.1 LEARNING TO FIND WHAT TO READ

On the 20-point confirmatory set, discovery accuracy rises from 0.425 to 0.740 after 30 steps and to 0.875 after 100 (in four segments; §5.4), a gain of 0.450 (95% CI [0.365, 0.540]; Table G.2). The same scenes also show gains on the cued tasks, which were never trained: at step 100, cued readout rises from 0.610 to 0.890 and find-and-bind from 0.620 to 0.950. At matched step 30, the two cued tasks gain +0.210 and +0.280, against +0.030 and +0.060 for GRAY. On target-switch pairs, discovery improves from 43 to 78 correct answers out of 100 at step 30, against 44 for GRAY, and to 88 at step 100 (Table G.5).

Learning continues long after the reward saturates: training reward exceeds 0.97 by step 13, while discovery on the 12- and 20-point development scenes rises through step 75 (Figure 4(a)). The result replicates: a second run of the same configuration ends within 0.010 of the first at every development density (Table G.1), and three independent seeds reach 0.730–0.760 on the confirmatory set at step 30; seed 1 reaches 0.895 at step 100 (Table G.6).

The gain does not hinge on the relation-only question. Training instead on the named-target question (find-and-bind) with the same probes, three seeds each, leaves confirmatory discovery +0.042 below the relation-only corpus (seed-level 95% interval [−0.027, +0.110]; Table G.7): identification and grounding, trained separately, compose into discovery. Identification is itself learned: on members the base model could not identify, REAL lifts discovery from 0.143 to 0.629, against 0.485 to 0.764 where it could; GRAY improves neither (Table G.8).

6.2 USING IMAGES, LEARNING FROM IMAGES

One control tests what the model uses; the other tests what taught it.

Images at test. Replacing test images with gray canvases reduces discovery and probe accuracy to exactly zero for both REAL runs at every development checkpoint and the evaluated confirmatory checkpoints. With the image gone, the trained models answer like the untrained one, with default responses outside the answer set (Appendix G).

Images during training. GRAY uses the same 2,880 prompts, starting checkpoint, reward, and budget, but replaces every training image with a gray canvas: it is shown everything REAL is shown, except the image. We evaluate both models with real images at matched step 30. The visually attributable gain $\tau_{\text{vis}}(t)$ is REAL's gain minus GRAY's over the common image-present base (Appendix F). On development scenes, REAL gains 0.305–0.330 across densities, GRAY changes by -0.035 to +0.030, and $\tau_{\text{vis}}(30) = 0.360$ (95% CI [0.275, 0.445]) on the never-trained 20-point density (Figure 4(b)); the confirmatory set shows the same contrast (Table G.3). The blind-gain pattern reverses: on Geometry3K, blind training recovered much of the real-image gain; here it recovers essentially none. The reversal holds seed by seed: across four independent seeds, REAL exceeds

GRAY by +0.354 on the confirmatory set (seed SD 0.019) and by +0.339 on an untouched set of 1,332 members (Table G.7).

Answers without looking. GRAY removes the image; a third arm keeps it and adds a shortcut: the question-blind caption of each training image goes into the prompt, with the same scenes, reward, budget and seeds, so that at the training densities the answer can be read from text. Reward is earned as before (accuracy component 0.97–0.99 by step 30); the skill is not. With the image alone, this arm gains +0.120 on confirmatory discovery against REAL's +0.323, below REAL in every seed (gap +0.203, seed-level 95% interval [+0.163, +0.243]); even with its caption at test it reaches at most 0.655 (Table G.9). Same reward, different skill.

6.3 DISCOVERY TRAINING TRANSFERS TO COORDINATE GROUNDING

Training on discovery and identification also improves find-and-bind on the grounding task and its independently regenerated twin. Both predate the training corpus and were never used in training; they were built to catch shortcuts, not to be trained on. At step 100, grounding-task pair accuracy rises from 0.768 to 0.948 (+0.180, 95% CI [0.147, 0.215]) and twin pair accuracy from 0.728 to 0.937. The second run reaches 0.962 on the grounding task, and seed 1 at step 100 reaches 0.942. The cued-readout and header-cued table tasks, built in other visual formats, change by at most 0.053 (Figure 4(c); Table G.4). At matched step 30, REAL gains +0.162 on grounding and the GRAY-trained control +0.002: the transfer to coordinate grounding comes from training images. Across the four seeds, REAL exceeds GRAY on the grounding task by +0.141 (seed SD 0.029; Table G.7).

7 TURNING THE DIAL

7.1 CHANGING THE MIXTURE, CHANGING THE GAIN

We mix constructed data with ordinary ViRL39K examples while holding the training budget fixed. In each 240-prompt step, a fraction $f \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ is constructed and the rest ordinary. Every condition starts from the same 7B model, keeps the reward and recipe, and sees 7,200 prompt presentations in 30 steps. We manipulate the constructed fraction f; the base-model audit measures the resulting stream's resolvability,

$$m(f) = f\,\overline{\Delta q}_{\text{constructed}} + (1-f)\,\overline{\Delta q}_{\text{ViRL},f}. \tag{3}$$

Here m measures a correct-answer probability difference, not a data fraction, and spans 0.126 to 0.249. Each grid point is a single run at the same budget.

On the never-trained 20-point density, ordinary data alone leaves discovery where it was (-0.010), a one-third constructed mixture gains +0.275, and the full corpus gains +0.330 (Figure 5(a)). One-third constructed data thus delivers over 80% of the full-corpus gain at the same budget. The full corpus gives the largest gain at every density (Table H.1). Paired comparisons on the same scenes favor the full corpus over either mixture by about 0.06, and the two mixtures do not differ (Table H.2).

The one-third mixture keeps ordinary-task accuracy. With images, it matches the untrained model on 4,239 held-out ViRL39K items (0.519 against 0.509), MathVista testmini (0.696 against 0.682) and BLINK (0.547 against 0.536); ordinary-only training reaches 0.548 on ViRL39K and the fully constructed run 0.419, its deficit concentrated on multiple-choice items (Appendix H).

At equal constructed exposure (2,400 presentations) the one-third mixture is ahead of the fully constructed run, 0.745 against 0.675 (Appendix H).

7.2 SAME TASKS, DIFFERENT LEARNING CURVES

The smaller model improves too, but on a different timetable. We train Qwen2.5-VL-3B on the same constructed corpus with the same recipe and 30-step budget; its identification-probe accuracy remains zero under gray-canvas evaluation (Table H.3). On held-out scenes at the training densities, 3B at step 30 roughly matches 7B at step 10: 0.865 against 0.870, and 0.730 against 0.760 (Figure 5(b)). On the never-trained density it reaches 0.545, against 0.675. Grounding-task performance improves at both scales (+0.130 at 3B, +0.162 at 7B; step 30).

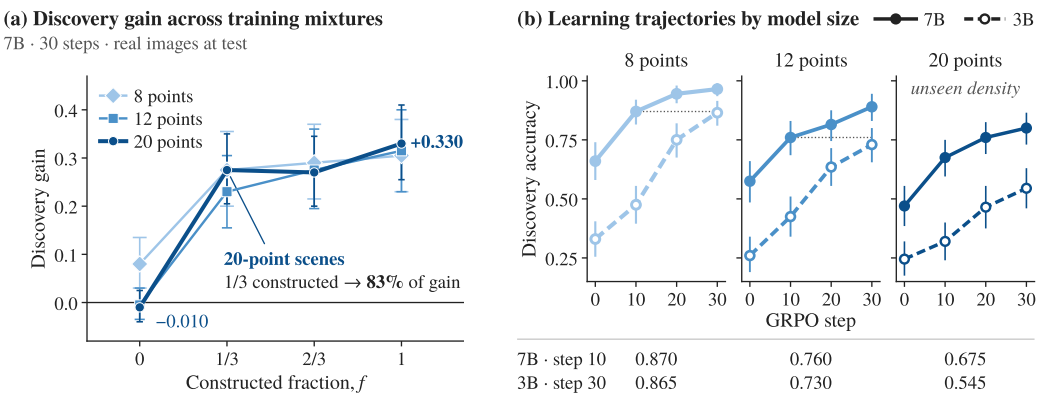


Figure 5: *Composition changes the gain; model size changes the trajectory.* (a) Step-30 discovery gain of each training mixture against the constructed fraction f. (b) Discovery accuracy of 3B and 7B models trained on the constructed corpus. Error bars: 95% intervals over scene programs.

8 Discussion

Specifying the reward does not finish specifying the learning problem. The verifier credits answers; the task determines what producing those answers requires. A correctness reward can reinforce an answer-producing shortcut or support visual discovery. In the mixture experiment, changing composition changes acquisition while model, reward, optimizer, and budget stay fixed. Task design, not only reward design, shapes the lesson.

Score is not skill. A higher score can conceal where a gain came from and which skill was lost along the way. Benchmark improvement, dependence on visual input, and learning from visual evidence are distinct claims. Image removal tests whether successful answers depend on the image; matched blind training tests whether the improvement was learned from it. A claim of visual learning must name the operation acquired and the information that taught it.

The learning problem is the lever. Reward engineering and data selection are means; the practical question is whether the learning problem makes the intended operation necessary for success. Data construction is one way to write that requirement into the task; the named-target ablation shows it need not take one question form (§6.1). The operation should also guide evaluation: benchmark gains can persist while grounding erodes, and discovery keeps improving after reward saturates. Checkpoint selection should track the skill being taught, not reward or aggregate score alone.

Scope. Coordinate scenes are a deliberate testbed, in which visual necessity can be verified directly and the target cue varied while the question stays fixed. Within it, training on relation-defined discovery and identification also improves question types that were never trained: cued readout and named-target grounding at a held-out scene density, and grounding tasks built before the corpus. The discovery gain appears in every run trained on real constructed images, at both model sizes. Applying the principle to other visual formats and to natural images, where visual necessity must first be measured, is the next test of its generality.

Beyond vision. In the constructed-corpus runs, the vision encoder and projector stay frozen; training changes only the language model. A given image supplies the same visual tokens before and after, yet the model becomes better at identifying a target, grounding it, and reading its value. Learning to see, in these experiments, was learning to look. The broader proposal concerns evidence, not pixels: an audio task can demand an acoustic distinction that a transcript omits; a video task can require temporal order; a tool-use task can make the returned result determine the answer. In each, the goal is to tie verifiable success to a chosen source of evidence while keeping the operation learnable. What matters is not how many modalities a prompt contains but which evidence the model must learn to use to succeed. That turns the question around: instead of observing which capabilities emerge, we design the learning problems that develop them.

The verifier decides what counts; the data determine what it takes. Changing what reward requires changes what RL learns.

AI USE STATEMENT

AI tools were used solely to polish the language and improve the readability of the manuscript. The authors conducted the literature review, developed the research ideas, carried out the experiments, selected the references, and wrote the initial draft. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Anonymized code: https://anonymous.4open.science/r/learning-without-looking-2432. The repository holds the scene program and the counterfactual grounding suite with its regenerated twin (Appendices B and C), corpus preparation and the training mixtures of §7.1, the base-model audit (§5.1, Appendix A), the training reward and answer matcher (§5.4, Appendix I), every training recipe with the modifications to the training framework (Appendix D), and the evaluation and statistics code (Appendices F and I). From the released per-item evaluation outputs, training streams, and audit rows, one script rebuilds the paper's tables and figure values without a GPU and compares each with the printed value; six printed values, such as the training-reward curve, come from trainer logs and are not rebuilt. Re-running training and evaluation requires GPUs; the repository's scripts fetch the released scene sets, grounding suite, corpora, and the 7B model trained on the constructed corpus.

REFERENCES

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

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

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: Reasoning transfer from LLMs to vision-language models via on-policy distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Mingkang Dong, Hongyi Cai, Jie Li, Sifan Zhou, Bin Ren, Kunyu Peng, and Yuqian Fu. VisNec: Measuring and leveraging visual necessity for multimodal instruction tuning. In *European Conference on Computer Vision (ECCV)*, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. In *Conference on Language Modeling (COLM)*, 2025.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Tiancheng Han, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Hoang Phan, Xianjun Yang, Yuanshun Yao, Jack Zhang, Shengjie Bi, Xiaocheng Tang, Madian Khabsa, Lijuan Liu, and Deren Lei. Beyond reasoning gains: Mitigating general-capability forgetting in large reasoning models. In *Findings of the Association for Computational Linguistics (ACL)*, 2026.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. In *International Conference on Machine Learning (ICML)*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. SoTA with less: MCTS-guided sample selection for data-efficient visual reasoning self-improvement. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Yuhang Zhou, Xiaoyu Liu, Yongyuan Liang, Ming Li, Ziyi Zang, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. ViCrit: A verifiable reinforcement learning proxy task for visual perception in VLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025c.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations (ICLR)*, 2026.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves MLLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-R1: Mitigating shortcuts in visual reasoning with reinforcement learning. *Transactions on Machine Learning Research*, 2026.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-R1: Advancing multimodal reasoning capabilities of MLLMs via visual perception reward. In *International Conference on Learning Representations (ICLR)*, 2026.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception-R1: Pioneering perception policy with reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025a.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

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 *European Conference on Computer Vision (ECCV)*, 2024.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Visual Necessity Audit

For each benchmark we compare accuracy with the image against accuracy with the image removed. For the retention ρ_{blind} of §2.2, the chance reference is $c = N^{-1} \sum_i 1/k_i$ for a multiple-choice task with k_i options on item i and $c = 0$ for free-form answers; mixed-format benchmarks are reported by format. Retention is a ratio of aggregate above-chance accuracies, not an average of per-item ratios; it is defined when $a_{\text{image}} > c$, need not lie in $[0, 1]$, and is not clipped. Table A.1 audits the untrained 3B model on seven benchmarks: BLINK, MathVerse, MMVP, MMMU (Yue et al., 2024), HallusionBench (Guan et al., 2024), MMStar and MathVista. Table A.2 repeats the audit for two model families we did not train, Gemma-3 and InternVL3-9B; these are additional base-model audits, not tests of the intervention on other families.

Table A.1: Visual-necessity audit of the untrained Qwen2.5-VL-3B model on seven benchmarks: accuracy with the image and with the image removed (blind), the chance level (mean of the item-specific $1/k$ for multiple choice; 0 for free-form answers), naive retention (blind ÷ with-image) and chance-corrected retention (blind−chance)/(with-image−chance), with 95% item-paired bootstrap intervals (10,000 draws). Mixed-format benchmarks are reported by format; HallusionBench is reported under both plausible chance levels.

benchmark (format)	n	with image	blind	chance	naive retention	corrected retention
BLINK (MC, pooled)	1901	0.493	0.409	0.377	0.829 [0.781, 0.879]	0.271 [0.089, 0.454]
MathVerse (MC, pooled)	2180	0.465	0.394	0.260	0.848 [0.807, 0.889]	0.655 [0.568, 0.743]
MathVerse (free-form)	1755	0.055	0.019	0.000	0.340 [0.240, 0.458]	0.340 [0.240, 0.458]
MMVP (MC, $k = 2$)	300	0.660	0.500	0.500	0.758 [0.668, 0.854]	0.000 [−0.417, 0.342]
MMMU dev+val (MC, pooled)	988	0.506	0.413	0.263	0.816 [0.764, 0.873]	0.617 [0.512, 0.729]
MMMU dev+val (free-form)	62	0.097	0.048	0.000	0.500 [0.000, 1.000]	0.500 [0.000, 1.000]
HallusionBench (free-form, chance 0)	1129	0.598	0.475	0.000	0.794 [0.748, 0.843]	0.794 [0.748, 0.843]
HallusionBench (Yes/No, chance 0.5)	1129	0.598	0.475	0.500	0.794 [0.748, 0.843]	−0.258 [−0.673, 0.039]
MMStar (MC, pooled)	1500	0.554	0.261	0.269	0.471 [0.430, 0.513]	−0.029 [−0.108, 0.049]
MathVista testmini (MC, pooled)	539	0.725	0.512	0.332	0.706 [0.646, 0.765]	0.458 [0.351, 0.564]
MathVista testmini (free-form)	460	0.504	0.115	0.000	0.228 [0.174, 0.287]	0.228 [0.174, 0.287]

Table A.2: The same audit for two other model families on a 4,096-item ViRL39K sample (the image is removed rather than replaced): with-image and blind accuracy, the chance level, and naive and chance-corrected retention with 95% item-paired bootstrap intervals. Corrected retention is defined only when with-image accuracy exceeds the chance level (§2.2). On the multiple-choice pool Gemma-3 scores below chance with the image, so its ratio is undefined; InternVL3-9B exceeds chance by only 0.026, so its ratio is defined but unstable and is reported unclipped. Without the image, both models score 0.000 pair accuracy on the grounding task and its twin (§2.2).

model	pool	n	with image	blind	chance	naive retention	corrected retention
Gemma-3	ViRL39K free-form	2789	0.430	0.312	0.000	0.727 [0.690, 0.765]	0.727 [0.690, 0.765]
Gemma-3	ViRL39K multiple choice	1215	0.135	0.086	0.268	0.634 [0.533, 0.744]	undefined
InternVL3-9B	ViRL39K free-form	2789	0.269	0.130	0.000	0.485 [0.439, 0.533]	0.485 [0.439, 0.533]
InternVL3-9B	ViRL39K multiple choice	1215	0.294	0.205	0.268	0.697 [0.619, 0.782]	−2.439 [−17.961, 6.505]

The two pools hold 4,004 of the 4,096 sampled items; 92 multiple-choice items whose number of options could not be determined are excluded.

B TASK CONSTRUCTION

Construction. Each scene program yields two twin images that differ in one fact, and three question levels on the same scene (Table B.1): L1 marks the target's position, L2 names the target, and L3 gives only the relation; a probe asks which point satisfies the relation. Pairs have one of three roles, in equal numbers: *target-stable* pairs move the target, so the answer changes while the same point satisfies the relation (the K8 scene); *target-switch* pairs change which point satisfies it; and *invariance* pairs change the image without changing the answer.

Scoring. Discovery accuracy is member accuracy over the target-stable and invariance pairs, 200 questions from 100 scene programs per density; the target-switch pairs, the cued levels and the probe on the same scenes are reported separately (Table G.5). Pair accuracy is used for the grounding suite of Appendix C.

Training selection and splits. From each of 720 scene programs with 8 or 12 points the corpus takes four prompts, the L3 question and the probe on both twins (2,880 prompts, 960 per role). L1 and L2 prompts are never used in training, 20-point scenes are reserved for evaluation, and training, development and confirmatory scenes come from disjoint scene programs. The seed replication of Table G.6 adds an untouched evaluation set, described there. On the corpus, 133 of 46,080 image-free audit responses were correct; this is the blind reward opportunity of 0.003 reported in §5.1.

Criteria and their outcome. The confirmatory test of §5.3 qualifies an evaluation cell (one scene density) by five criteria, listed with their outcome for every density in Table B.4. Caption-only accuracy (criterion 4) uses question-blind captions written by Qwen2.5-VL-72B-Instruct (served with vLLM in bf16, temperature 0, at most 384 tokens); the untrained 7B model then answers from the caption alone. The artifact screen (criterion 5) trains four ridge-regularized linear classifiers ($\lambda = 10^{-2}$) to tell the two twins of an L3 pair apart, from a DINOv2-small image embedding, the PNG file size, 21 image and frequency statistics, or 13 file features, on the target-stable and target-switch pairs (200 images per density) with five-fold cross-validation grouped by pair. A cell passes when, for every classifier, the out-of-fold $\max(\mathrm{AUC}, 1-\mathrm{AUC})$ is at most 0.55 and the upper end of the 95% pair-bootstrap interval of the unfolded AUC is at most 0.62, at each of three cross-validation seeds, which change folds and bootstrap draws on the same images; passing the screen shows that these classifiers cannot tell the twins apart, not that no artifact exists.

The development set, used to track learning, meets all five criteria at no density, so acquisition is assessed on a separate confirmatory set of 450 new scene programs, 150 per density; its 20-point cell meets all five and supplies the confirmatory result (Table G.2). The criteria and their numerical limits were set before the confirmatory set was generated; the requirement that the artifact screen pass at each of the three seeds was added after generation began and before any criterion was evaluated.

Ordinary training. Tables B.2 and B.3 give discovery and the task-layer gains of the models trained with ordinary rewards (§2.3).

Table B.1: The three levels and the identification probe on one development scene (eight labeled points; the K8 scene of Figure 3(b)): the question templates, the cue each level supplies, and the gold answers on the two twins, which differ in one fact (K8 moves from $(-1, -1)$ to $(6, -1)$). L2 and the probe use the L3 image; L1 adds only the offset arrow.

level	question	visual cue	gold (twin a)	gold (twin b)
L1 (arrow cue; position given)	Point K8 has the smallest y-coordinate. What is the x-coordinate of point K8?	offset arrow at K8	-1	6
L2 (target named; identity given)	Point K8 has the smallest y-coordinate. What is the x-coordinate of point K8?	none	-1	6
L3 (discover + ground + read; nothing given)	Consider the point with the smallest y-coordinate. What is its x-coordinate?	none	-1	6
probe (identification only)	Which labeled point has the smallest y-coordinate?	none	K8	K8

Decoding is greedy with at most 32 new tokens. Results on these tasks are reported as member accuracy; pair accuracy (Eq. (1)) is used for the grounding suite.

Table B.2: Discovery on the coordinate development set (member accuracy with real images) for the models trained with ordinary rewards: the ViRL39K-trained 3B conditions (two seeds; step 100) and the Geometry3K-trained 7B pair (step 100), with the untrained models. Discovery (L3) and the identification probe at 8, 12, and 20 points. Two-seed means are computed from the per-seed counts; intervals are not reported for these values.

model	discovery 8-pt	12-pt	20-pt	probe 8-pt	12-pt	20-pt
3B, ViRL39K-trained (step 100); two seeds						
untrained 3B	0.330	0.260	0.245	0.705	0.680	0.630
REAL, two-seed mean	0.383	0.338	0.330	0.728	0.700	0.610
seeds 1 / 2	0.380 / 0.385	0.330 / 0.345	0.330 / 0.330	0.725 / 0.730	0.710 / 0.690	0.620 / 0.600
CAPTION, two-seed mean	0.370	0.323	0.340	0.728	0.715	0.635
seeds 1 / 2	0.365 / 0.375	0.325 / 0.320	0.345 / 0.335	0.715 / 0.740	0.710 / 0.720	0.610 / 0.660
NONE, two-seed mean	0.368	0.303	0.318	0.730	0.700	0.638
seeds 1 / 2	0.365 / 0.370	0.300 / 0.305	0.305 / 0.330	0.740 / 0.720	0.705 / 0.695	0.650 / 0.625
GRAY, two-seed mean	0.373	0.298	0.315	0.713	0.683	0.625
seeds 1 / 2	0.370 / 0.375	0.295 / 0.300	0.320 / 0.310	0.725 / 0.700	0.685 / 0.680	0.630 / 0.620
7B, Geometry3K-trained (step 100)						
untrained 7B	0.660	0.575	0.470	0.940	0.910	0.840
REAL	0.765	0.575	0.460	0.940	0.905	0.855
GRAY	0.720	0.570	0.460	0.945	0.905	0.845

Table B.3: Task-layer gains of the Geometry3K-trained 3B conditions at step 100 (pair accuracy; three-seed means with 95% paired-item intervals; per-seed gains beneath). The equivalence test for the grounding task is in Appendix I.

condition	cued readout: gain [95% CI]	find-and-bind	header-cued table
untrained 3B (pair accuracy)	0.320	0.455	0.867
REAL	+0.143 [0.102, 0.188]	+0.018 [−0.007, 0.044]	+0.019 [−0.002, 0.042]
seeds 1 / 2 / 3	+0.157 / +0.127 / +0.147	+0.012 / +0.022 / +0.022	+0.030 / +0.013 / +0.013
CAPTION	+0.107 [0.066, 0.149]	+0.006 [−0.018, 0.030]	+0.021 [−0.001, 0.044]
seeds 1 / 2 / 3	+0.130 / +0.113 / +0.077	+0.003 / 0.000 / +0.013	+0.023 / +0.027 / +0.013
NONE	+0.108 [0.070, 0.148]	−0.013 [−0.039, 0.013]	+0.023 [0.000, 0.049]
seeds 1 / 2 / 3	+0.093 / +0.097 / +0.133	−0.010 / −0.010 / −0.020	+0.023 / +0.020 / +0.027
GRAY	+0.138 [0.096, 0.182]	−0.028 [−0.055, −0.002]	+0.023 [0.002, 0.047]
seeds 1 / 2 / 3	+0.140 / +0.137 / +0.137	−0.030 / −0.032 / −0.023	+0.027 / +0.017 / +0.027

Cued readout accounts for 72.1% of REAL's pair-count-weighted improvement across the 1,200-pair suite: its gain times its pair count, divided by the same product summed over the three tasks. Hierarchy discovery, measured by member accuracy, is not part of this denominator.

Table B.4: Outcome of the five evaluation criteria of §5.3 on the untrained 7B model, for every density of the development set (used to track learning) and of the separate confirmatory set: (1) discovery accuracy between 0.20 and 0.80; (2) discovery at least 0.05 below both cued levels; (3) exactly zero accuracy on discovery and its probe with a gray canvas and with no image; (4) caption-only accuracy at most half of with-image discovery accuracy; (5) the artifact screen of Appendix B: out-of-fold $\max(\mathrm{AUC}, 1-\mathrm{AUC})$ of four classifiers at most 0.55, and the upper end of the unfolded interval at most 0.62, at each of three cross-validation seeds. Criteria 1–3 are met in every cell. Only the 20-point confirmatory cell meets all five; it supplies the confirmatory result.

			(4) caption-only accuracy				
set	density	discovery (L3)	accuracy	limit ($\frac{1}{2}$ L3)	outcome	(5) artifact screen	criteria met
development	8-pt	0.660	0.580	0.330	not met	met	4 of 5
	12-pt	0.575	0.445	0.288	not met	not met: 0.557 in 1 of 3 seeds	3 of 5
	20-pt	0.470	0.185	0.235	met	not met: 0.565 in 1 of 3 seeds	4 of 5
confirmatory	8-pt	0.585	0.570	0.293	not met	met	4 of 5
	12-pt	0.560	0.410	0.280	not met	met	4 of 5
	20-pt	0.425	0.130	0.213	met	met	5 of 5

C Independently Constructed Grounding Evaluations

Each task pairs two renders of one scene that differ in a single answer-relevant fact, so an answer that ignores the image cannot solve the pair (Proposition J.1). The tasks were built before the constructed training corpus existed and share no scenes with it. Table C.1 lists them with their roles and untrained levels and summarizes their validation.

Table C.1: The independently constructed grounding tasks: the 1,200-pair counterfactual suite (three tasks with distinct roles, reported separately) and its independently regenerated twin (fresh seeds, no iteration). Untrained pair accuracy (both members correct) for the 3B and 7B models; 7B levels with 95% intervals. Validation: Qwen2.5-VL-72B-Instruct given question-blind captions reaches at most 0.054; every image-removed condition scores exactly 0.000, with all responses collapsing to one answer; 60 of 60 audited pairs were confirmed by human review; Gemma-3 and InternVL3-9B also score exactly 0.000 without the image.

task	pairs	role	untrained 3B	untrained 7B [95% CI]
grounding task (named-point coordinate pairs)	600	find-and-bind (primary)	0.455	0.768 [0.733, 0.802]
header-cued table task	300	cued control, high baseline	0.867	0.993 [0.983, 1.000]
cued-readout task (marked-plot pairs)	300	cued readout, location marked	0.320	0.673 [0.620, 0.727]
twin of the grounding task	600	regenerated twin	–	0.728 [0.692, 0.762]
twin of the cued-readout task	300	twin control	–	0.623 [0.570, 0.677]
twin of the header-cued table task	300	twin control	–	0.997 [0.990, 1.000]

The twin was evaluated for the 7B models only.

D Training and Canonical Access Results

Recipes. Table D.1 lists the training recipe of every run behind the main-text results. Training uses EasyR1 (revision dd71bbd) with local patches for the visual-access conditions; the repository lists the patches and the recorded revision of every run. The vision encoder and its projector into the language model are frozen in every run except the long-horizon runs, so only the language model is trained; the saved visual weights of the constructed-corpus checkpoints are identical to the base model's.

Segmented training. The constructed-corpus runs of §§6–7 share an initial 30-step budget; the two REAL runs of seed 0 and the REAL run of seed 1 continue to 100 steps in segments: at steps 30, 50 and 75 a fresh trainer resumes from the previous checkpoint, so the AdamW moments reset, the KL reference is re-anchored to the resumed model, and the corpus order restarts from its beginning. Every matched 30-step comparison ends before the first restart.

Reward received by the control. The GRAY-trained control receives format reward throughout and only sparse correctness reward: its mean accuracy reward rises from 0.0008 at step 1 to 0.031 at step 30 (zero on probe rows), while its format reward rises from 0.87 to 1.00. In the three further seeds

(Table D.7), the control's mean accuracy reward is 0.0017–0.0050 at step 1 and 0.021–0.048 at step 30, against 0.972–0.991 for the REAL and named-target runs.

Replication structure. An independent seed changes two things here: the shuffled order in which the corpus is presented, and the seed of rollout sampling. The Geometry3K, ViRL39K and long-horizon runs draw both afresh for each seed. The two 100-step REAL constructed-corpus runs of seed 0 instead repeat one configuration, including its seeds and data order, so the pair measures how closely a repeated run reproduces the first; the corpus is presented in the same fixed order in both. Seeds 1–3 of the constructed-corpus conditions (REAL, GRAY-trained, named-target and shortcut-accessible; Tables G.6 and G.9) draw both afresh, with data-order and rollout-sampling seeds 1/1, 2/2 and 3/3 against 0/0 for seed 0, and run 30 steps each; the REAL run of seed 1 then continues to step 100 on the seed-0 schedule. Table D.1 therefore counts seeds and runs in separate columns, and Table D.6 gives the seeds, the runs and the resampled unit behind each result reported in this paper.

Table D.2 gives Table 1 per seed; the CAPTION condition and the caption test input use question-blind captions written by Qwen2.5-VL-3B-Instruct. Table D.4 gives the counts behind the 7B share of §2.1. Table D.3 gives the 7B pair on the independently constructed grounding suite and its twin (§3). Table D.5 splits the blind gains of Table 1 by whether the untrained model's real-image answer could be extracted (497 of 601 questions); its note defines extractability.

E CORROSION

Table E.1 gives both long-horizon trajectories of §3: benchmark accuracy peaks at step 200, and grounding pair accuracy falls from step 100 to step 400 by 0.070 (95% CI [0.043, 0.098]) and 0.052 ([0.020, 0.083]), ending 0.048 ([0.017, 0.082]) and 0.042 ([0.005, 0.078]) below the base model. In the first run, 137 of 601 benchmark items change correctness between steps 100 and 400 despite a net gain of three correct answers: a flat score, a churning model. Relative to §2.1, these runs change the horizon, the encoder, the corpus filter and the reward implementation together, so they show the loss under this recipe rather than isolating one cause.

Table E.2 gives the per-run degraded sets, their overlap against the permutation null, and the shared wrong answers behind Figure 2(c); its note characterizes those answers.

F GAINS, ATTRIBUTION, AND MIXTURE MEASUREMENT

Fix a task, scoring rule, model size, and checkpoint budget. Let $a_c^v(t)$ be the accuracy of training condition c at step t under evaluation input v, and a_0^v that of the base model. The gain and the visually attributable gain of §6.2 are

$$g_c(t) = a_c^{\text{image}}(t) - a_0^{\text{image}}, \qquad \tau_{\text{vis}}(t) = g_{\text{REAL}}(t) - g_{\text{GRAY}}(t) = a_{\text{REAL}}^{\text{image}}(t) - a_{\text{GRAY}}^{\text{image}}(t). \quad \text{(F.1)}$$

Both trained models are evaluated with real images at the matched step, so the shared base cancels. The interval on τ_{vis} comes from a paired comparison of the two models on the same items, not from arithmetic on the endpoints of the two marginal intervals. Table G.7 reports the same contrast per seed, with a t-based interval over seeds (Appendix I). Test-time dependence is a different contrast: it changes the evaluation input of one checkpoint, $a_c^{\text{image}}(t) - a_c^{\text{removed}}(t)$, with the removal named (test-image deletion in §2, a gray canvas in §§6–7). The recovery share of §2.1 is g_c/g_{REAL} with gains averaged over seeds before division; it is a ratio of averaged gains, not an average of per-seed ratios, and it is not confined to $[0, 1]$.

The measured resolvability $m(f)$ of Eq. (3), the audited mass $m(f)$ of the mixture in Appendix H, combines two component means. With $\mathcal{C}$ the 2,880 audited constructed training rows and $\mathcal{V}_f$ the unique audited ViRL39K items in the stream of condition f,

$$\overline{\Delta q}_{\text{constructed}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \Delta q_i, \qquad \overline{\Delta q}_{\text{ViRL},f} = \frac{1}{|\mathcal{V}_f|} \sum_{i \in \mathcal{V}_f} \Delta q_i. \quad \text{(F.2)}$$

The constructed mean is the same in every condition, and the ordinary mean depends on f; at $f = 1$ the mass is the constructed mean, and no mean over an empty ordinary subset is evaluated. Both means weight unique audited items equally. The standardization of Appendix J.3 uses a different audit population, the realized training rows weighted by their presentations.

Table D.1: Training recipes. Every run uses GRPO with 5 rollouts per prompt, learning rate 10^{-6} and KL coefficient 0.01. In every recipe the verifier scores each response's extracted answer against its gold answer and, where noted, its format; no image-dependent auxiliary loss or pairwise training reward is added. Constructed-corpus runs (§§6–7): rollout batch 240, global batch 48. Geometry3K, ViRL39K and long-horizon runs: rollout batch 512, global batch 128 (the long-horizon runs on 4 GPUs with tensor parallelism 2). The last two columns count, per condition, the independent seeds and the training runs: an independent seed draws its own shuffled data order and rollout sampling; a run beyond the seed count repeats one configuration. The dose response of §7.1 has four points, the three mixtures listed here and the standard run at step 30 ($f = 1$).

experiment	backbone, vision encoder	training images	reward	training data	steps; evaluated at	seeds	runs
Geometry3K access conditions (§2.1, §3)	3B, frozen	REAL / GRAY / NONE / CAPTION, one condition per run	answer correctness	Geometry3K, filtered (1,288 items)	100	3	3
7B access pair (§2.1, §3)	7B, frozen	REAL; GRAY	answer correctness	Geometry3K	100	1	1
long-horizon runs (§3)	3B, trainable	real images	0.5·accuracy + 0.5·format	Geometry3K, unfiltered	400, in segments; evaluated at 100, 150, 200, 300, 400	2	2
resolvable training, standard reward (§6)	7B, frozen	real images	0.5·accuracy + 0.5·format	constructed corpus, 2,880 items (discovery and probe, both twins)	seeds 0 and 1: 100, in segments (seed 0 evaluated at 10, 20, 30, 50, 75, 100; seed 1 at 30 and 100). Seeds 2 and 3: 30	4	5 (two of seed 0)
GRAY-trained control (§6.2)	7B, frozen	gray canvases	as above	same corpus	30; evaluated at 10, 20, 30 (seed 0) or 30 (seeds 1–3)	4	4
named-target ablation (§6.1; Table G.6)	7B, frozen	real images	as above	same scene programs, with the find-and-bind (L2) question in place of the discovery question; probe unchanged	30	3	3
shortcut-accessible twin (§6.2; Table G.9)	7B, frozen	real images, with the question-blind 72B caption of each training image in the prompt	as above	same 2,880 prompts; caption placed after the image and before the question	30	3	3
3B capacity comparison (§7.2)	3B, frozen	real images	as the standard reward	constructed corpus, identical scenes	30; evaluated at 10, 20, 30	1	1
mixtures $f = 0, \frac{1}{3}, \frac{2}{3}$ (§7.1)	7B, frozen	real images	as the standard reward	constructed corpus mixed with decontaminated ViRL39K in 240-prompt steps	30; evaluated at 10, 20, 30	1	1 each
ViRL39K access conditions (§2.3)	3B, frozen	REAL / GRAY / NONE / CAPTION, one condition per run	answer correctness	decontaminated ViRL39K, single-image items (23,542)	100	2	2

Segments. In the 100-step constructed-corpus runs, each restart (steps 30, 50 and 75) warm-starts a fresh trainer from the merged checkpoint, with the effects described above; the step-30 checkpoint, re-evaluated after the extension, matched its original evaluation, and every 30-step run (GRAY-trained control, 3B comparison, mixtures) ends before the first restart. The long-horizon runs were also trained in segments (seed 1 continues a 100-step run to step 400 in 50-step segments; seed 2 runs from the base model to step 400).

Table D.2: Table 1 per seed, with the untrained and trained models scored by the same answer matcher: accuracy of each condition's step-100 model under the four test conditions (601 Geometry3K test items), the gain over the untrained model under the same test condition, and the recovery share (real-tested gain ÷ REAL's real-tested gain). Intervals are not computed for the three-seed means and shares; the per-seed values show their spread.

condition	seed	acc. real	acc. gray	acc. none	acc. caption	gain real	gain gray	gain none	gain caption	share
base model	–	0.175	0.090	0.068	0.210	–	–	–	–	–
REAL	1	0.423	0.106	0.108	0.311	+0.248	+0.017	+0.040	+0.102	–
	2	0.419	0.112	0.098	0.301	+0.245	+0.022	+0.030	+0.091	–
	3	0.398	0.110	0.106	0.326	+0.223	+0.020	+0.038	+0.117	–
CAPTION	1	0.361	0.118	0.112	0.318	+0.186	+0.028	+0.043	+0.108	0.752
	2	0.346	0.108	0.106	0.291	+0.171	+0.018	+0.038	+0.082	0.701
	3	0.331	0.117	0.097	0.326	+0.156	+0.027	+0.028	+0.117	0.702
NONE	1	0.325	0.095	0.097	0.286	+0.150	+0.005	+0.028	+0.076	0.604
	2	0.308	0.117	0.121	0.294	+0.133	+0.027	+0.053	+0.085	0.544
	3	0.303	0.103	0.118	0.261	+0.128	+0.013	+0.050	+0.052	0.575
GRAY	1	0.311	0.110	0.108	0.286	+0.136	+0.020	+0.040	+0.076	0.550
	2	0.306	0.097	0.088	0.270	+0.131	+0.007	+0.020	+0.060	0.537
	3	0.281	0.106	0.103	0.245	+0.106	+0.017	+0.035	+0.035	0.478

Table D.3: The ordinary 7B access pair of §3 on the independently constructed grounding suite and its twin (Geometry3K training, step 100): untrained pair accuracy and the change under REAL and GRAY training, with 95% paired-item bootstrap intervals (2,000 draws). The interval excludes zero for REAL on the grounding task and on its twin, and includes zero in every other cell.

task	pairs	untrained 7B	REAL Δ [95% CI]	GRAY Δ [95% CI]
grounding task	600	0.768	+0.035 [0.013, 0.058]	+0.012 [−0.010, 0.033]
twin of the grounding task	600	0.728	+0.038 [0.017, 0.062]	+0.013 [−0.010, 0.037]
cued-readout task	300	0.673	+0.030 [−0.007, 0.067]	+0.017 [−0.013, 0.047]
twin of the cued-readout task	300	0.623	+0.013 [−0.030, 0.060]	+0.013 [−0.023, 0.050]
header-cued table task	300	0.993	0.000 [−0.010, 0.010]	−0.003 [−0.010, 0.000]
twin of the header-cued table task	300	0.997	0.000 [−0.010, 0.010]	−0.007 [−0.017, 0.000]

Table D.4: The 7B access pair behind the share in Table 1 (Geometry3K training; 601 Geometry3K test items; the answer matcher of Table 1): correct answers with real test images and with a gray canvas, with 95% item-bootstrap intervals, and the gain over the untrained model with real images, with its item-paired interval (5,000 draws). The recovery share is GRAY's gain divided by REAL's: 116/149 = 0.779 [0.642, 0.921].

trained with	step	real image: correct [95% CI]	gain over untrained [95% CI]	gray canvas: correct [95% CI]
untrained 7B	–	141/601 = 0.235 [0.201, 0.270]	–	48/601 = 0.080 [0.058, 0.103]
REAL	100	290/601 = 0.483 [0.443, 0.521]	+0.248 [0.203, 0.291] (149/601)	75/601 = 0.125 [0.100, 0.151]
GRAY	100	257/601 = 0.428 [0.388, 0.466]	+0.193 [0.153, 0.235] (116/601)	79/601 = 0.131 [0.103, 0.160]

Table D.5: The blind gains of Table 1 by whether the untrained 3B model's real-image answer can be extracted (extractable: 497 of 601 questions): the untrained model's accuracy, the gain over it with real test images, averaged over the three seeds per question, and each condition's share of REAL's gain, with 95% intervals from a 10,000-draw bootstrap over questions. The three runs of a condition answer the same 601 questions, so questions, not the 1,803 responses, are resampled.

questions	n	untrained	REAL gain	CAPTION gain	NONE gain	GRAY gain
all questions	601	0.175	+0.238 [0.202, 0.275]	+0.171 [0.136, 0.207]	+0.137 [0.103, 0.171]	+0.125 [0.092, 0.158]
extractable	497	0.211	+0.200 [0.160, 0.240]	+0.140 [0.100, 0.179]	+0.107 [0.069, 0.145]	+0.101 [0.064, 0.139]
not extractable	104	0.000	+0.423 [0.343, 0.503]	+0.324 [0.250, 0.401]	+0.279 [0.212, 0.349]	+0.237 [0.176, 0.301]

questions	CAPTION share	NONE share	GRAY share
all questions	0.719 [0.614, 0.821]	0.574 [0.467, 0.673]	0.523 [0.416, 0.622]
extractable	0.698 [0.565, 0.828]	0.537 [0.395, 0.664]	0.507 [0.364, 0.636]
not extractable	0.765 [0.620, 0.930]	0.659 [0.519, 0.813]	0.561 [0.438, 0.691]

An answer is extractable when the answer extractor returns one, which is distinct from following the answer-tag format. The untrained model's answers are extractable for 497 questions with the image and 502 with the caption, with accuracy 0.211 and 0.251 among them, so its higher caption accuracy is not a difference in extraction coverage. The split is descriptive: it shows that blind gains are not confined to questions whose answer the untrained model could not state, not how much of any gain is due to format.

Table D.6: Run basis of each reported result. *Seeds* counts the independent seeds of a condition and *runs* its training runs (Table D.1); where runs exceed seeds, the further run repeats one configuration. The fourth column names the table that prints the individual seeds or runs, and the last the unit that each result's 95% intervals resample with the trained models held fixed.

result (section; table)	seeds	runs	individual seeds or runs	95% intervals resample
blind recovery shares, 3B (§2.1; Table 1)	3	3	Table D.2, per seed	point estimates over seeds
GRAY recovery share, 7B (§2.1; Table D.4)	1	1	–	benchmark items
operation-level gains, 3B (§2.3; Table B.3)	3	3	Table B.3, per seed	grounding pairs
discovery, ViRL39K-trained 3B (§2.3; Table B.2)	2	2	Table B.2, per seed	point estimates over seeds
grounding loss under prolonged training (§3; Table E.1)	2	2	Table E.1, per run	grounding pairs
discovery on the confirmatory set (§6.1; Tables G.2, G.6)	4	5	Table G.2 (seed 0, per run); Table G.6 (seeds 1–3; seed 1 also at step 100)	scene programs (seed 0); seeds, t-based (Table G.7)
visually attributable gain τ_{vis} (§6.2; Tables G.3, G.7)	4	5 (REAL), 4 (GRAY)	Table G.6, per seed	scene programs (seed 0); seeds, t-based (Table G.7)
transfer to the grounding tasks (§6.3; Tables G.4, G.7)	4	5	Table G.4 (seed 0, per run); Table G.6 (seeds 1–3; seed 1 also at step 100)	grounding pairs (seed 0); seeds (Table G.7)
named-target ablation (§6.1; Table G.7)	3	3 per condition	Table G.6, per seed	seeds, t-based (Table G.7)
shortcut-accessible twin (§6.2; Tables G.9, G.10)	3	3	Table G.9, per seed	seeds, t-based (Table G.10)
mixture contrasts at step 30 (§7.1; Tables H.1 and H.2)	1	1 per mixture	–	scene programs
3B against 7B (§7.2; Table H.3)	1	1 per model	–	scene programs

Table D.7: Training rewards of seeds 1–3 (Table G.6): the overall reward of Eq. (2) with its format and accuracy components at steps 1 and 30, and the mean accuracy component over steps 19–30. The GRAY-trained control earns format reward from the start and almost no accuracy reward at any point; the REAL, named-target and shortcut-accessible runs approach the ceiling of both components by step 30.

condition	seed	step 1: overall / format / accuracy	step 30: overall / format / accuracy	mean accuracy, steps 19–30
REAL	1	0.5975 / 0.8658 / 0.3292	0.9858 / 1.0000 / 0.9717	0.975
REAL	2	0.5900 / 0.8875 / 0.2925	0.9863 / 1.0000 / 0.9725	0.982
REAL	3	0.6095 / 0.8825 / 0.3365	0.9930 / 0.9992 / 0.9867	0.970
GRAY-trained	1	0.4350 / 0.8683 / 0.0017	0.5175 / 1.0000 / 0.0350	0.033
GRAY-trained	2	0.4413 / 0.8800 / 0.0025	0.5238 / 0.9992 / 0.0483	0.041
GRAY-trained	3	0.4425 / 0.8800 / 0.0050	0.5104 / 1.0000 / 0.0208	0.024
named-target	1	0.5829 / 0.8708 / 0.2950	0.9942 / 1.0000 / 0.9883	0.987
named-target	2	0.5810 / 0.8825 / 0.2795	0.9896 / 0.9992 / 0.9800	0.979
named-target	3	0.5998 / 0.8855 / 0.3140	0.9954 / 1.0000 / 0.9908	0.991
shortcut-accessible	1	0.7138 / 0.8892 / 0.5383	0.9904 / 1.0000 / 0.9808	0.977
shortcut-accessible	2	0.7000 / 0.8825 / 0.5175	0.9850 / 0.9992 / 0.9708	0.972
shortcut-accessible	3	0.7354 / 0.9008 / 0.5700	0.9946 / 1.0000 / 0.9892	0.981

Seed 1, continued to step 100 (Table G.6): accuracy component 0.9875, 0.9933 and 0.9908 at steps 50, 75 and 100, format 1.0000 throughout, KL 0.041–0.054.

Table E.1: The two long-horizon runs of §3 (Qwen2.5-VL-3B, trainable encoder, real images, unfiltered Geometry3K): benchmark accuracy (correct answers out of 601) with 95% item-bootstrap intervals, and grounding-task pair accuracy with its interval and its change from the untrained model (the 600 pairs, each its own scene, resampled with the same 10,000 draws at every checkpoint). Peaks are marked.

seed	step	benchmark [95% CI]	grounding [95% CI]	Δ vs base [95% CI]
1	0	0.175 [0.145, 0.206]	0.455 [0.415, 0.495]	–
	100	0.431 [0.393, 0.471]	0.477 [0.437, 0.517]	+0.022 [−0.008, 0.052]
	150	0.463 [0.424, 0.504]	0.465 [0.425, 0.505]	+0.010 [−0.020, 0.040]
	200	0.483 [0.444, 0.524] (peak)	0.450 [0.412, 0.490]	−0.005 [−0.035, 0.025]
	300	0.464 [0.424, 0.506]	0.440 [0.400, 0.480]	−0.015 [−0.047, 0.017]
	400	0.436 [0.398, 0.474]	0.407 [0.367, 0.447]	−0.048 [−0.082, −0.017]
2	0	0.175 [0.145, 0.206]	0.455 [0.415, 0.495]	–
	100	0.428 [0.389, 0.468]	0.465 [0.425, 0.505]	+0.010 [−0.020, 0.040]
	150	0.446 [0.404, 0.488]	0.468 [0.428, 0.508]	+0.013 [−0.020, 0.047]
	200	0.496 [0.454, 0.537] (peak)	0.447 [0.407, 0.487]	−0.008 [−0.042, 0.025]
	300	0.449 [0.409, 0.488]	0.400 [0.362, 0.438]	−0.055 [−0.090, −0.018]
	400	0.453 [0.411, 0.493]	0.413 [0.373, 0.453]	−0.042 [−0.078, −0.005]

From step 100 to step 400, grounding-task pair accuracy changes by −0.070 [−0.098, −0.043] in the first run and −0.052 [−0.083, −0.020] in the second.

Table E.2: Shared errors across the three GRAY-trained runs of §3 (grounding task, 600 pairs). The sets are defined with the earlier answer-matcher revision (Appendix I). Top: per run, the degraded set (pairs the untrained model solves and the run fails), pairs gained, and wrong members (individual pair members answered incorrectly within the degraded set). Bottom: overlap of the degraded sets (Jaccard; intersection/union) against a 10,000-draw permutation null of sets of the same sizes ($p = 10^{-4}$ for every comparison), and the rate of identical wrong answers on the wrong members the runs share.

run	degraded pairs	pairs gained	wrong members
seed 1	51	24	52
seed 2	49	22	53
seed 3	45	23	46

sets	Jaccard	permutation null mean	identical wrong answer, shared wrong members [95% CI]
seeds 1 vs 2	0.724 (42/58)	0.097 (95th pct. 0.149)	0.976 [0.877, 0.996] (41/42)
seeds 1 vs 3	0.811 (43/53)	0.093 (95th pct. 0.143)	1.000 [0.920, 1.000] (44/44)
seeds 2 vs 3	0.741 (40/54)	0.091 (95th pct. 0.146)	0.976 [0.874, 0.996] (40/41)
all three	0.661 (39/59)	0.011 (95th pct. 0.026)	0.975 [0.871, 0.996] (39/40)

Degraded sets are drawn from the 283 of 600 pairs the untrained model solves under that matcher revision. Pairs failed by seed 1 / 2 / 3 only: 5 / 6 / 1; by exactly two seeds (1&2 / 1&3 / 2&3): 3 / 4 / 1; by all three: 39; union 59. The 39 identical answers take 11 distinct values; the most frequent value, 1, occurs 14 times, each time against a gold answer of −1. Several explanations drawn from the scene can fit one answer, and each answer is counted under every explanation it fits: the gold plus one, skipping zero (21); the gold with its sign flipped (15); the nearest point's x-coordinate (6); the target's own y-coordinate (3). Eleven answers match only some coordinate of the scene, which on a dense scene does not show that the point was selected. On the same 39 member slots, the three REAL-trained runs give the identical wrong answer in 29 slots in at least one run and in 20 in all three (NONE: 37 and 28; CAPTION: 32 and 21). Re-scored with the current matcher, the degraded sets have 48, 47 and 43 pairs, 37 of them shared by all three runs.

G ACQUISITION, ATTRIBUTION, AND TRANSFER

Table G.1 gives development-set acquisition at steps 0, 30 and 100 for both runs, with the identification probe; Table G.2 every confirmatory-set cell; Table G.3 the attribution of §6.2 on both sets; Table G.4 transfer to the independently constructed grounding tasks at the same three checkpoints; and Table G.5 the same-scene hierarchy and the target-switch pairs on the 20-point confirmatory scenes. The development trajectories at all six checkpoints are plotted in Figure 4(a); Figure 4(c) shows the transfer changes at step 100. Tables G.6 and G.7 give the seed replication of the matched comparison and the named-target ablation; Table G.8 gives the composite-operation analysis, and Tables G.9 and G.10 the shortcut-accessible twin.

A gray canvas does not make the models refuse; it makes them give the same default answer every time. With a gray canvas, the REAL runs and the untrained model answer every question in the required format but with a fixed default: 0 on discovery, which is not among the 14 possible coordinates, and a letter that labels no point on the probe. Every such cell therefore scores zero; the responses are not refusals, which occur only when the image is removed altogether. The seed-0 GRAY-trained control and the 3B model can default to a coordinate that occurs and score a few percent on discovery with a gray canvas (Tables G.2, G.3 and H.3); on the 20-point confirmatory set the control and the GRAY-trained seeds 1–3 score 0 of 200.

Table G.1: Development-set acquisition (Figure 4(a)): discovery (L3) member accuracy with real images at steps 0, 30 and 100 for both runs, by density, and its change from the untrained model, with 95% intervals from resampling scene programs (Appendix I); last column, the identification probe with real images at 8 / 12 / 20 points. With the test image replaced by a gray canvas, discovery and the probe score 0 of 200 at every one of the six checkpoints of both runs.

run	step	8-pt: level [95% CI]; Δ [95% CI]	12-pt	20-pt	probe 8 / 12 / 20-pt
untrained 7B	0	0.660 [0.580, 0.740]	0.575 [0.485, 0.660]	0.470 [0.385, 0.555]	0.940 / 0.910 / 0.840
run 1	30	0.965 [0.935, 0.990]	0.890 [0.830, 0.945]	0.800 [0.730, 0.865]	1.000 / 0.975 / 0.975
		Δ +0.305 [0.230, 0.380]	Δ +0.315 [0.230, 0.400]	Δ +0.330 [0.255, 0.410]	
	100	200/200 (100 programs)	0.940 [0.895, 0.980]	0.910 [0.860, 0.950]	1.000 / 0.990 / 0.995
		Δ +0.340 [0.260, 0.420]	Δ +0.365 [0.280, 0.455]	Δ +0.440 [0.360, 0.525]	
run 2	30	0.960 [0.925, 0.985]	0.860 [0.795, 0.920]	0.795 [0.730, 0.860]	1.000 / 0.960 / 0.960
		Δ +0.300 [0.225, 0.380]	Δ +0.285 [0.205, 0.370]	Δ +0.325 [0.245, 0.405]	
	100	0.995 [0.985, 1.000]	0.950 [0.915, 0.980]	0.905 [0.860, 0.945]	0.995 / 0.985 / 0.990
		Δ +0.335 [0.255, 0.415]	Δ +0.375 [0.290, 0.465]	Δ +0.435 [0.350, 0.520]	

Seed replication and the named-target ablation. Three further independent seeds (1–3) repeat the matched step-30 comparison of §6.2 with fresh data order and rollout sampling; seed 0 is the original pair (REAL runs 1 and 2, which share a seed, and the GRAY-trained control). Each new seed also trains a third condition, the **named-target corpus**: the same scene programs with the find-and-bind (L2) question in place of the relation-only (L3) question and the identification probe unchanged, so that discovery itself is never trained. Table G.6 gives every run at step 30 on the 20-point confirmatory set and on an additional untouched set (333 scene programs per pair role; 1,332 discovery members and 333 target-switch pairs) that had not been used in any development or analysis before this evaluation; Table G.7 summarizes the contrasts over seeds.

The attribution replicates seed by seed. REAL exceeds GRAY on confirmatory discovery by +0.354 (seed SD 0.019; every seed at least +0.335) and on the untouched set by +0.339 (SD 0.021); GRAY changes from the untrained model by −0.033 (SD 0.032) and never by more than −0.075. Transfer to the grounding task replicates as well: REAL minus untrained +0.145 (SD 0.016), REAL minus GRAY +0.141 (SD 0.029). The step-30 REAL levels themselves are 0.730–0.760 (mean 0.746, SD 0.014). Seed 1 then continues to step 100 on the seed-0 schedule, with restarts at steps 30, 50 and 75, and reaches 0.895 on confirmatory discovery (seed 0: 0.875), 0.900 on the untouched set, 0.942 and 0.925 on the grounding task and its twin, and 0.905 and 0.960 on cued readout and find-and-bind (identification probe: 297 of 300). The two step-100 endpoints are reported directly, without a two-seed interval. The named-target corpus reaches 0.695–0.720 on confirmatory discovery without ever training it, +0.042 below the relation-only corpus (seed-level 95% interval [−0.027, +0.110]; +0.045 [−0.005, +0.095] on the untouched set), and transfers to the grounding task about as strongly (seed means 0.928 against 0.907). A model taught to find a named point and to name the point a relation picks out can, without further instruction, find the point a relation picks out.

Table G.2: All confirmatory-set cells (150 new scene programs per density, generated after the evaluation criteria of §5.3 were set): discovery member accuracy with real images and its change from the untrained model, and accuracy with the test image replaced by a gray canvas, with 95% intervals from resampling scene programs; a cell with 0 of 200 correct is given as its count and number of scene programs. The last column gives how many of the five criteria the evaluation cell met before training; it describes the cell, not a run's outcome. The 20-point cell met all five and supplies the confirmatory result of §6.1; the 8- and 12-point cells did not meet the caption criterion (Appendix B).

			real images		gray canvas	
run	step	density	level [95% CI]	Δ vs untrained [95% CI]	level [95% CI]	criteria met
untrained 7B	0	8-pt	0.585 [0.500, 0.665]	–	0/200 (100 programs)	4 of 5
untrained 7B	0	12-pt	0.560 [0.470, 0.645]	–	0/200 (100 programs)	4 of 5
untrained 7B	0	20-pt	0.425 [0.335, 0.515]	–	0/200 (100 programs)	5 of 5
REAL, run 1	30	8-pt	0.935 [0.895, 0.970]	+0.350 [0.275, 0.430]	0/200 (100 programs)	4 of 5
REAL, run 1	30	12-pt	0.860 [0.805, 0.910]	+0.300 [0.225, 0.380]	0/200 (100 programs)	4 of 5
REAL, run 1	30	20-pt	0.740 [0.665, 0.815]	+0.315 [0.235, 0.400]	0/200 (100 programs)	5 of 5
REAL, run 1	100	8-pt	0.990 [0.975, 1.000]	+0.405 [0.325, 0.490]	0/200 (100 programs)	4 of 5
REAL, run 1	100	12-pt	0.940 [0.905, 0.970]	+0.380 [0.295, 0.470]	0/200 (100 programs)	4 of 5
REAL, run 1	100	20-pt	0.875 [0.815, 0.930]	+0.450 [0.365, 0.540]	0/200 (100 programs)	5 of 5
REAL, run 2	30	8-pt	0.915 [0.865, 0.960]	+0.330 [0.250, 0.415]	0/200 (100 programs)	4 of 5
REAL, run 2	30	12-pt	0.850 [0.795, 0.905]	+0.290 [0.215, 0.370]	0/200 (100 programs)	4 of 5
REAL, run 2	30	20-pt	0.720 [0.645, 0.795]	+0.295 [0.215, 0.375]	0/200 (100 programs)	5 of 5
GRAY-trained control	30	8-pt	0.590 [0.510, 0.670]	+0.005 [−0.035, 0.050]	0.050 [0.020, 0.090]	4 of 5
GRAY-trained control	30	12-pt	0.580 [0.490, 0.665]	+0.020 [−0.025, 0.070]	0.060 [0.025, 0.100]	4 of 5
GRAY-trained control	30	20-pt	0.385 [0.300, 0.475]	−0.040 [−0.085, 0.000]	0/200 (100 programs)	5 of 5

With a gray canvas the GRAY-trained control changes by +0.050 [0.020, 0.090] and +0.060 [0.025, 0.100] from the untrained model at 8 and 12 points; every other gray-canvas cell, its 20-point cell included, is 0 of 200: each response is a fixed default outside the answer set.

Table G.3: Attribution at matched step 30: the REAL gain and the GRAY-trained control's change (both evaluated with real images) and their difference τ_{vis} (Eq. (F.1)), on the development set (dev) and replicated on the confirmatory set (conf); task: discovery (L3) or the identification probe, with each model's gray-canvas count at the same checkpoint (correct/total). 95% intervals from resampling scene programs, with the same draws for both models.

						gray-canvas count	
set	density	task	REAL Δ [95% CI]	control Δ [95% CI]	τ_{vis} [95% CI]	REAL	control
dev	8-pt	L3	+0.305 [0.230, 0.380]	+0.030 [−0.005, 0.065]	+0.275 [0.205, 0.350]	0/200	11/200
dev	8-pt	probe	+0.060 [0.025, 0.105]	+0.005 [−0.020, 0.035]	+0.055 [0.020, 0.095]	0/200	0/200
dev	12-pt	L3	+0.315 [0.230, 0.400]	−0.035 [−0.080, 0.005]	+0.350 [0.265, 0.435]	0/200	9/200
dev	12-pt	probe	+0.065 [0.025, 0.110]	−0.005 [−0.035, 0.020]	+0.070 [0.030, 0.120]	0/200	0/200
dev	20-pt	L3	+0.330 [0.255, 0.410]	−0.030 [−0.070, 0.010]	+0.360 [0.275, 0.445]	0/200	8/200
dev	20-pt	probe	+0.135 [0.085, 0.190]	+0.020 [0.000, 0.045]	+0.115 [0.070, 0.170]	0/200	0/200
conf	8-pt	L3	+0.350 [0.275, 0.430]	+0.005 [−0.035, 0.050]	+0.345 [0.270, 0.425]	0/200	10/200
conf	8-pt	probe	+0.090 [0.045, 0.145]	+0.020 [0.005, 0.040]	+0.070 [0.025, 0.120]	–	–
conf	12-pt	L3	+0.300 [0.225, 0.380]	+0.020 [−0.025, 0.070]	+0.280 [0.210, 0.355]	0/200	12/200
conf	12-pt	probe	+0.080 [0.035, 0.135]	−0.005 [−0.035, 0.020]	+0.085 [0.035, 0.140]	–	–
conf	20-pt	L3	+0.315 [0.235, 0.400]	−0.040 [−0.085, 0.000]	+0.355 [0.270, 0.440]	0/200	0/200
conf	20-pt	probe	+0.150 [0.090, 0.215]	+0.025 [0.000, 0.050]	+0.125 [0.070, 0.185]	–	–

Confirmatory discovery on gray test images is 0 of 200 for every run. Without the image, per-twin accuracy on the untouched set is 0.069–0.078 for every model, the untrained one included; the contrast between REAL and GRAY is unaffected, since both conditions share that floor.

Composite operations. Discovery requires identifying the target and then locating and reading it. Table G.8 splits the 200 confirmatory members by whether the untrained model answers the identification probe for that member correctly (165 members) or not (35) and reports step-30 discovery within each subset for seed 0. REAL raises discovery in both: from 0.485 to 0.764 where identification was already available, and from 0.143 to 0.629 where it was not. GRAY leaves the second subset unchanged and the first slightly lower. The gain is not confined to targets the base model could already identify; identification is itself acquired, which is what the compositional reading of the named-target ablation requires. The split conditions on a base-model outcome and is reported descriptively, without intervals.

Answers available without looking. The GRAY-trained control removes the image and with it almost all correctness reward (Table D.7). A third condition keeps the image and the reward and adds a text shortcut: the question-blind Qwen2.5-VL-72B caption of each training image (the captions

Table G.4: Transfer to the independently constructed grounding tasks (Figure 4(c)) at steps 0, 30 and 100 for both runs: pair accuracy with 95% intervals and, beneath, the paired change from the untrained model (5,000-draw item bootstrap). Changes are computed from unrounded accuracies, so they can differ by 0.001 from the difference of the rounded levels (as do the labels of Figure 4(c)). Step-30 twin levels are from the evaluation of Table G.6; their changes are differences of rounded levels and carry no interval.

run	step	grounding task (600 pairs)	twin (600)	cued-readout task (300)	header-cued table (300)
untrained 7B	0	0.768 [0.733, 0.802]	0.728 [0.692, 0.762]	0.673 [0.620, 0.727]	0.993 [0.983, 1.000]
run 1	30	0.930 [0.908, 0.950]	0.900	0.680 [0.627, 0.733]	0.997 [0.990, 1.000]
		Δ +0.162 [0.130, 0.193]	Δ +0.172	Δ +0.007 [−0.023, 0.037]	Δ +0.003 [0.000, 0.010]
	100	0.948 [0.930, 0.965]	0.937 [0.917, 0.955]	0.707 [0.657, 0.757]	0.997 [0.990, 1.000]
		Δ +0.180 [0.147, 0.215]	Δ +0.208 [0.175, 0.243]	Δ +0.033 [0.000, 0.070]	Δ +0.003 [0.000, 0.010]
run 2	30	0.935 [0.915, 0.955]	0.905	0.690 [0.640, 0.743]	0.997 [0.990, 1.000]
		Δ +0.167 [0.137, 0.198]	Δ +0.177	Δ +0.017 [−0.013, 0.047]	Δ +0.003 [0.000, 0.010]
	100	0.962 [0.947, 0.977]	–	0.727 [0.677, 0.777]	0.997 [0.990, 1.000]
		Δ +0.193 [0.162, 0.227]		Δ +0.053 [0.020, 0.087]	Δ +0.003 [0.000, 0.010]

Table G.5: The same-scene hierarchy on the 20-point confirmatory scenes, a density never trained (100 scene programs in the top block): member accuracy with real images with its 95% interval over scene programs, its change from the untrained model, and, for pairs whose answer changes between the twins, the number of scene programs with both members correct. Top: the four tasks on the target-stable and invariance pairs, the population of the discovery result (200 members). Bottom: pair roles separately (100 members each); the cued levels have no target-switch pairs. L1 and L2 ask the same question, and L1 adds only the arrow; L1, L2 and the 20-point density were never trained, while the probe was.

task	untrained	REAL run 1, 30	REAL run 2, 30	GRAY-trained control, 30	REAL run 1, 100
L1 cued readout	0.610 [0.525, 0.695]	0.820 [0.755, 0.880] Δ +0.210 [0.140, 0.285]	0.810 [0.745, 0.870] Δ +0.200 [0.130, 0.270]	0.640 [0.555, 0.720] Δ +0.030 [−0.010, 0.075]	0.890 [0.835, 0.935] Δ +0.280 [0.200, 0.360]
L2 find-and-bind	0.620 [0.530, 0.705]	0.900 [0.850, 0.945] Δ +0.280 [0.200, 0.360]	0.875 [0.815, 0.930] Δ +0.255 [0.180, 0.330]	0.680 [0.600, 0.760] Δ +0.060 [0.020, 0.105]	0.950 [0.915, 0.980] Δ +0.330 [0.250, 0.415]
L3 discovery	0.425 [0.335, 0.515]	0.740 [0.665, 0.815] Δ +0.315 [0.235, 0.400]	0.720 [0.645, 0.795] Δ +0.295 [0.215, 0.375]	0.385 [0.300, 0.475] Δ −0.040 [−0.085, 0.000]	0.875 [0.815, 0.930] Δ +0.450 [0.365, 0.540]
identification probe	0.825 [0.755, 0.890]	0.975 [0.940, 1.000] Δ +0.150 [0.090, 0.215]	0.975 [0.940, 1.000] Δ +0.150 [0.090, 0.215]	0.850 [0.785, 0.905] Δ +0.025 [0.000, 0.050]	0.985 [0.960, 1.000] Δ +0.160 [0.100, 0.225]

task	untrained	REAL run 1, 30	REAL run 2, 30	GRAY-trained control, 30	REAL run 1, 100
L3, target-switch	0.430 [0.320, 0.540] both: 13/50	0.780 [0.700, 0.860] Δ +0.350 [0.250, 0.450] both: 30/50	0.750 [0.670, 0.830] Δ +0.320 [0.230, 0.420] both: 27/50	0.440 [0.340, 0.550] Δ +0.010 [−0.060, 0.080] both: 12/50	0.880 [0.810, 0.940] Δ +0.450 [0.340, 0.560] both: 39/50
probe, target-switch	0.890 [0.830, 0.940] both: 39/50	0.990 [0.970, 1.000] Δ +0.100 [0.050, 0.160] both: 49/50	0.990 [0.970, 1.000] Δ +0.100 [0.050, 0.160] both: 49/50	0.910 [0.860, 0.960] Δ +0.020 [−0.020, 0.060] both: 41/50	100/100 (50 programs) Δ +0.110 [0.060, 0.170] both: 50/50
L3, target-stable	0.410 [0.290, 0.530] both: 14/50	0.730 [0.630, 0.820] Δ +0.320 [0.230, 0.420] both: 29/50	0.720 [0.620, 0.820] Δ +0.310 [0.200, 0.420] both: 29/50	0.380 [0.270, 0.490] Δ −0.030 [−0.090, 0.030] both: 12/50	0.890 [0.820, 0.950] Δ +0.480 [0.370, 0.590] both: 40/50
L3, invariance	0.440 [0.300, 0.580]	0.750 [0.630, 0.860] Δ +0.310 [0.180, 0.450]	0.720 [0.600, 0.830] Δ +0.280 [0.160, 0.400]	0.390 [0.260, 0.520] Δ −0.050 [−0.120, 0.010]	0.860 [0.760, 0.940] Δ +0.420 [0.290, 0.550]

On target-switch discovery, REAL minus the GRAY-trained control at step 30 is +0.340 [0.250, 0.430]. A cell with every member correct is given as its count.

of criterion 4, Appendix B) is placed in the prompt after the image and before the question, with the 2,880 prompts, reward, budget, starting checkpoint and data-order and rollout seeds of REAL seeds 1–3 otherwise unchanged. At the training densities the caption alone answers 0.580 (8 points) and 0.445 (12 points) of discovery questions, so the answer is often available without looking; at 20 points it is not (Table B.4). Table G.9 gives the three seeds at step 30, tested with the image alone and with the image and caption; Table G.10 gives the contrasts over seeds.

Table G.6: Seed replication (§§6.1–6.3): seeds 1–3 of the REAL, GRAY-trained and named-target conditions at step 30, and the REAL run of seed 1 continued to step 100, beside seed 0 (Tables G.2–G.5). Discovery member accuracy on the 20-point confirmatory set (200 members) and on the untouched set (1,332 members); pair accuracy on the grounding task and its twin; target-switch discovery with both members correct on the confirmatory (of 50 pairs) and untouched (of 333) sets; find-and-bind (200 members) on the confirmatory scenes. Counts are given where recorded.

seed	condition	discovery, confirmatory	discovery, untouched	grounding task	twin	target-switch, conf.	target-switch, untouched	find-and-bind
–	untrained 7B	0.425 (85/200)	0.471 (628/1,332)	0.768	0.728	13/50	0.318	0.620
0	REAL, run 1	0.740	0.791	0.930	0.900	30/50	0.586	0.900
0	REAL, run 2 (same seed)	0.720	0.770	0.935	0.905	27/50	0.562	0.875
0	GRAY-trained	0.385	0.440	0.770	0.730	12/50	0.252	0.680
1	REAL	0.760 (152/200)	0.792 (1,055/1,332)	0.907	0.885	30/50	0.583	0.915
1	REAL, step 100	0.895 (179/200)	0.900 (1,199/1,332)	0.942	0.925	–	–	0.960
1	GRAY-trained	0.415 (83/200)	0.451 (601/1,332)	0.778	0.742	13/50	0.267 (89/333)	0.685
1	named-target	0.705 (141/200)	0.725 (966/1,332)	0.943	0.912	24/50	0.502 (167/333)	0.960
2	REAL	0.730 (146/200)	0.769 (1,024/1,332)	0.922	0.898	26/50	0.544 (181/333)	0.925
2	GRAY-trained	0.350 (70/200)	0.414 (551/1,332)	0.752	0.715	10/50	0.213 (71/333)	0.645
2	named-target	0.720 (144/200)	0.742 (988/1,332)	0.903	0.873	29/50	0.529 (176/333)	0.980
3	REAL	0.755 (151/200)	0.762 (1,015/1,332)	0.893	0.882	33/50	0.595 (198/333)	0.890
3	GRAY-trained	0.420 (84/200)	0.453 (603/1,332)	0.787	0.737	14/50	0.282 (94/333)	0.705
3	named-target	0.695 (139/200)	0.720 (959/1,332)	0.938	0.923	25/50	0.508 (169/333)	0.955

Table G.7: Levels and contrasts over seeds at step 30 (seed 0 uses REAL run 1): per-seed values, their mean and sample standard deviation, and, for the contrasts used as seed-level claims in §§6.1–6.3, the range and the t-based 95% interval over seeds (Appendix I).

quantity	endpoint	seeds	per-seed values	mean	seed SD	range	seed-level 95% interval
REAL level	discovery, confirmatory	0–3	0.740, 0.760, 0.730, 0.755	0.746	0.014	0.030	–
REAL − GRAY	discovery, confirmatory	0–3	+0.355, +0.345, +0.380, +0.335	+0.354	0.019	0.045	[+0.323, +0.384]
REAL − GRAY	discovery, untouched	0–3	+0.351, +0.341, +0.355, +0.309	+0.339	0.021	0.046	[+0.306, +0.372]
REAL − untrained	discovery, confirmatory	0–3	+0.315, +0.335, +0.305, +0.330	+0.321	0.014	–	–
REAL − untrained	discovery, untouched	0–3	+0.320, +0.321, +0.297, +0.291	+0.307	0.015	–	–
GRAY − untrained	discovery, confirmatory	0–3	−0.040, −0.010, −0.075, −0.005	−0.033	0.032	–	–
REAL − untrained	grounding task	0–3	+0.162, +0.139, +0.154, +0.125	+0.145	0.016	–	–
REAL − GRAY	grounding task	0–3	+0.160, +0.129, +0.170, +0.106	+0.141	0.029	–	–
relation-only − named-target	discovery, untouched	1–3	+0.067, +0.027, +0.042	+0.045	0.020	0.040	[−0.005, +0.095]
relation-only − named-target	discovery, confirmatory	1–3	+0.055, +0.010, +0.060	+0.042	0.028	0.050	[−0.027, +0.110]

Table G.8: Step-30 discovery on the 20-point confirmatory set (200 members; seed 0, REAL run 1 and the GRAY-trained control), by whether the untrained model's identification probe for the same member is correct.

subset	members	untrained	REAL	GRAY-trained
probe correct	165	0.485 (80/165)	0.764 (126/165)	0.436 (72/165)
probe wrong	35	0.143 (5/35)	0.629 (22/35)	0.143 (5/35)
all	200	0.425 (85/200)	0.740 (148/200)	0.385 (77/200)

Reward is earned as under REAL: the accuracy component starts at 0.52–0.57 (REAL: 0.29–0.34) and reaches 0.97–0.99 by step 30 (Table D.7). The skill is not. With the image alone, confirmatory discovery is 0.550, 0.515 and 0.570, a gain of +0.120 over the untrained model against +0.323 for REAL; every shortcut seed is below every REAL seed, with a seed-level gap of +0.203 [+0.163, +0.243]. The untouched set shows the same gap (+0.195 [+0.132, +0.258]), and grounding transfer is +0.043 against +0.125 to +0.154 for REAL. With its caption at test the arm reaches 0.610–0.655, still below every REAL seed's image-only level, so the gap is not a change of input format; what the arm learned was to read the caption. Gray test images give 0 of 200 in every seed.

Table G.9: The shortcut-accessible twin at step 30, seeds 1–3, tested with the image alone (the paper's evaluation) and with the image and its caption (the training input): discovery member accuracy on the confirmatory (200) and untouched (1,332) sets, pair accuracy on the grounding task (600) and discovery with gray test images. REAL and GRAY-trained seeds 1–3 and the untrained model are in Table G.6.

seed	test input	discovery, confirmatory	discovery, untouched	grounding task	gray test images
1	image only	0.550 (110/200)	0.589 (784/1,332)	0.820 (492/600)	0/200
2	image only	0.515 (103/200)	0.553 (737/1,332)	0.802 (481/600)	0/200
3	image only	0.570 (114/200)	0.595 (793/1,332)	0.812 (487/600)	0/200
1	image and caption	0.645 (129/200)	0.669 (891/1,332)	–	–
2	image and caption	0.610 (122/200)	0.646 (861/1,332)	–	–
3	image and caption	0.655 (131/200)	0.682 (908/1,332)	–	–

Table G.10: Contrasts over seeds for the shortcut-accessible twin (image only), in the format of Table G.7; REAL values from Table G.7 for comparison.

quantity	endpoint	per-seed values (1, 2, 3)	mean	seed SD	seed-level 95% interval	REAL reference
shortcut level	discovery, confirmatory	0.550, 0.515, 0.570	0.545	0.028	[0.476, 0.614]	0.760, 0.730, 0.755
shortcut − untrained	discovery, confirmatory	+0.125, +0.090, +0.145	+0.120	0.028	[+0.051, +0.189]	+0.335, +0.305, +0.330
REAL − shortcut	discovery, confirmatory	+0.210, +0.215, +0.185	+0.203	0.016	[+0.163, +0.243]	REAL − GRAY: +0.345, +0.380, +0.335
shortcut level	discovery, untouched	0.589, 0.553, 0.595	0.579	0.023	[0.523, 0.635]	0.792, 0.769, 0.762
REAL − shortcut	discovery, untouched	+0.203, +0.215, +0.167	+0.195	0.025	[+0.132, +0.258]	REAL − GRAY: +0.341, +0.355, +0.309
shortcut − untrained	grounding task	+0.052, +0.033, +0.043	+0.043	0.009	[+0.020, +0.066]	+0.139, +0.154, +0.125

H DOSE, AUDIT STANDARDIZATION, AND CAPACITY

Dose response. Table H.1 gives all four mixture conditions of §7.1. Each grid point is a single run from the common start checkpoint with the same reward, budget (30 steps, 240 prompts per step, five responses per prompt) and stream definition; the constructed prompts per step are 0, 80, 160 and 240. On the never-trained density the two intermediate mixtures are tied (+0.275 and +0.270), and the fully constructed condition gives the largest gain on every density. On this fixed-budget grid the audited mass $m(f)$ is nearly linear in f (correlation 0.997; largest residual from a straight line 0.006) and serves as the audit's descriptor of each mixture.

Audit association and equal exposure. Table H.2 reports the association between step-30 accuracy and the audited mass, raw and standardized to a common initial-success profile (Appendix J.3); the raw slope depends strongly on the condition without constructed data, and the standardized mass is itself nonlinear in f. It also gives both equal-exposure comparisons: at equal constructed-prompt exposure the one-third mixture is ahead of the fully constructed run on all three densities, and the two-thirds mixture is within 0.035 of it (the 12- and 20-point intervals include zero). Its last block compares the mixtures directly at step 30.

Ordinary-task accuracy. On the decontaminated held-out ViRL39K split (4,239 items, none shared with any training stream), greedy step-30 accuracy with real images is 0.548, 0.519, 0.522 and 0.419 for $f = 0, \frac{1}{3}, \frac{2}{3}, 1$, each at least 0.20 above the same model's image-removed accuracy. The fully constructed run's deficit lies on the 1,215 multiple-choice items (0.325, against 0.649–0.660 for the other three conditions); on numeric and free-text items it is within 0.05 of the other conditions. The untrained model scores 0.509 with images (2,157 of 4,239; 0.250 without) and 0.640 on the multiple-choice items, so the two mixtures match it and ordinary-only training exceeds it. Table H.4 extends the comparison to two audited benchmarks: with images, the one-third mixture matches the untrained model on MathVista testmini (0.696 against 0.682) and BLINK (0.547 against 0.536), with blind retention unchanged; the fully constructed run falls on both.

Capacity. Table H.3 gives the capacity comparison of §7.2 per checkpoint.

Table H.1: Dose response at step 30 by constructed fraction f (same 7B starting model, reward, recipe and 7,200 prompt presentations in all four conditions): audited mass $m(f)$ (Eq. (3)), standardized mass $\widetilde{m}(f)$ (Appendix J.3) and the share of audit-joined training rows with $0 < b_i < 1$ (in-band); per density, discovery accuracy with real images and its gain over the untrained model with a 95% interval from resampling scene programs; and grounding-task pair accuracy. With a gray canvas, discovery accuracy is 0 of 200 in every cell of every mixture.

f	$m(f)$	$\widetilde{m}(f)$	in-band	8-pt: level; gain [95% CI]	12-pt	20-pt (never trained)	grounding [95% CI]
0	0.126	0.094	0.752	0.740 +0.080 [0.030, 0.135]	0.570 −0.005 [−0.035, 0.030]	0.460 −0.010 [−0.040, 0.025]	0.797 [0.763, 0.828]
$\frac{1}{3}$	0.165	0.229	0.900	0.935 +0.275 [0.200, 0.355]	0.805 +0.230 [0.155, 0.305]	0.745 +0.275 [0.205, 0.350]	0.908 [0.885, 0.930]
$\frac{2}{3}$	0.199	0.240	0.926	0.950 +0.290 [0.215, 0.370]	0.850 +0.275 [0.195, 0.360]	0.740 +0.270 [0.200, 0.345]	0.927 [0.905, 0.947]
1	0.249	0.251	–	0.965 +0.305 [0.230, 0.380]	0.890 +0.315 [0.230, 0.400]	0.800 +0.330 [0.255, 0.410]	0.930 [0.908, 0.950]

The in-band share was not recorded for $f = 1$: the stream profiles cover only the three conditions that contain ordinary data.

Table H.2: Audit association, equal-exposure and direct comparisons for the mixtures. Top: ordinary least-squares slope (with intercept, over the four conditions) of step-30 held-out accuracy on the audited mass $m(f)$ and on the standardized mass $\widetilde{m}(f)$ (Appendix J.3); 95% intervals resample scene programs (grounding pairs for the grounding task) with the audited masses held fixed (Appendix I). Middle: equal constructed-prompt exposure. Discovery accuracy of a mixture at step 30 and of the fully constructed run at the step where it had seen the same number of constructed prompt presentations (five responses each), with their difference and its 95% interval. The two runs of each comparison differ in total presentations, so these comparisons match constructed exposure, not total training. Bottom: direct comparisons of the mixtures at step 30.

endpoint	slope on $m(f)$ [95% CI]	slope on $\widetilde{m}(f)$ [95% CI]
8-pt discovery	1.670 [1.187, 2.190]	1.441 [1.005, 1.907]
12-pt discovery	2.446 [1.813, 3.088]	1.942 [1.432, 2.471]
20-pt discovery (never trained)	2.482 [1.886, 3.083]	2.079 [1.565, 2.596]
grounding task	1.009 [0.787, 1.233]	0.861 [0.669, 1.054]

	$f = \frac{1}{3}$ at step 30 vs. $f = 1$ at step 10			$f = \frac{2}{3}$ at step 30 vs. $f = 1$ at step 20		
constructed prompt presentations	2,400	2,400		4,800	4,800	
total prompt presentations	7,200	2,400		7,200	4,800	
density	mixture	$f = 1$	difference [95% CI]	mixture	$f = 1$	difference [95% CI]
8-pt	0.935	0.870	+0.065 [0.030, 0.105]	0.950	0.945	+0.005 [0.000, 0.015]
12-pt	0.805	0.760	+0.045 [0.015, 0.080]	0.850	0.815	+0.035 [−0.010, 0.085]
20-pt (never trained)	0.745	0.675	+0.070 [0.030, 0.115]	0.740	0.760	−0.020 [−0.060, 0.015]

density	$f = \frac{2}{3}$ minus $f = \frac{1}{3}$ [95% CI]	$f = 1$ minus $f = \frac{1}{3}$ [95% CI]	$f = 1$ minus $f = \frac{2}{3}$ [95% CI]
8-pt	+0.015 [0.000, 0.035]	+0.030 [0.010, 0.055]	+0.015 [0.000, 0.035]
12-pt	+0.045 [0.000, 0.095]	+0.085 [0.045, 0.130]	+0.040 [0.000, 0.080]
20-pt (never trained)	−0.005 [−0.050, 0.045]	+0.055 [0.025, 0.085]	+0.060 [0.020, 0.100]

I Scoring and Uncertainty

Answer matching. A response is correct when its extracted answer is equivalent to the gold under the answer matcher, and trained and untrained models are always scored with the same matcher. One exception to the current matcher revision affects interpretation: the shared-error sets of Table E.2 were defined with the earlier revision, under which the untrained 3B model solves 283 of the 600 grounding pairs rather than 273. The current revision changes member accuracy by at most 0.043 per run relative to it, and in a conservative direction: image-removed accuracies fall, pair accuracies without the image stay at exactly 0.000, and gains over the base model rise rather than fall.

Units and aggregation. The grounding suite is scored by pair accuracy (Eq. (1): both members correct); the constructed discovery tasks and their probe by member accuracy. Results over seeds are means of per-seed values, and recovery shares divide seed-averaged gains (Appendix F).

Table H.3: Capacity (Figure 5(b)): the 3B and 7B models trained on the same constructed corpus, per checkpoint and density (development set): discovery with real images and its change from the same model untrained, with 95% intervals from resampling scene programs; the identification probe with real images and with a gray canvas; discovery with a gray canvas; and grounding-task pair accuracy where evaluated. With a gray canvas the probe is exactly zero for both models at every checkpoint, while the 3B model's discovery accuracy is small and nonzero at every checkpoint, including before training: it answers with a fixed coordinate that is sometimes correct.

			discovery, real images		identification probe		discovery	
model	step	density	level [95% CI]	Δ vs untrained [95% CI]	real	gray canvas	gray canvas	grounding task [95% CI]
3B	0	8-pt	0.330 [0.255, 0.405]	–	0.705	0.000	0.080	0.455 [0.415, 0.495]
		12-pt	0.260 [0.190, 0.340]	–	0.680	0.000	0.025	
		20-pt	0.245 [0.175, 0.320]	–	0.630	0.000	0.055	
	10	8-pt	0.475 [0.395, 0.555]	+0.145 [0.065, 0.225]	0.875	0.000	0.080	–
		12-pt	0.425 [0.340, 0.510]	+0.165 [0.085, 0.245]	0.755	0.000	0.025	
		20-pt	0.320 [0.240, 0.400]	+0.075 [−0.010, 0.160]	0.740	0.000	0.055	
	20	8-pt	0.750 [0.675, 0.820]	+0.420 [0.325, 0.510]	0.960	0.000	0.105	–
		12-pt	0.635 [0.555, 0.715]	+0.375 [0.285, 0.465]	0.870	0.000	0.030	
		20-pt	0.465 [0.375, 0.555]	+0.220 [0.120, 0.315]	0.880	0.000	0.050	
	30	8-pt	0.865 [0.810, 0.915]	+0.535 [0.450, 0.620]	0.985	0.000	0.060	0.585 [0.547, 0.625]
		12-pt	0.730 [0.655, 0.800]	+0.470 [0.375, 0.560]	0.945	0.000	0.040	
		20-pt	0.545 [0.460, 0.630]	+0.300 [0.200, 0.395]	0.960	0.000	0.070	
7B	0	8-pt	0.660 [0.580, 0.740]	–	0.940	0.000	0.000	0.768 [0.733, 0.802]
		12-pt	0.575 [0.485, 0.660]	–	0.910	0.000	0.000	
		20-pt	0.470 [0.385, 0.555]	–	0.840	0.000	0.000	
	10	8-pt	0.870 [0.815, 0.920]	+0.210 [0.145, 0.275]	0.985	0.000	0.000	–
		12-pt	0.760 [0.685, 0.830]	+0.185 [0.115, 0.260]	0.950	0.000	0.000	
		20-pt	0.675 [0.595, 0.750]	+0.205 [0.135, 0.280]	0.940	0.000	0.000	
	20	8-pt	0.945 [0.905, 0.980]	+0.285 [0.215, 0.360]	0.995	0.000	0.000	–
		12-pt	0.815 [0.745, 0.875]	+0.240 [0.165, 0.320]	0.960	0.000	0.000	
		20-pt	0.760 [0.690, 0.825]	+0.290 [0.215, 0.370]	0.975	0.000	0.000	
	30	8-pt	0.965 [0.935, 0.990]	+0.305 [0.230, 0.380]	1.000	0.000	0.000	0.930 [0.908, 0.950]
		12-pt	0.890 [0.830, 0.945]	+0.315 [0.230, 0.400]	0.975	0.000	0.000	
		20-pt	0.800 [0.730, 0.865]	+0.330 [0.255, 0.410]	0.975	0.000	0.000	

Table H.4: The untrained 7B model and the step-30 $f = 1$ and $f = \frac{1}{3}$ checkpoints on MathVista testmini (1,000 items, 539 of them multiple-choice) and BLINK (1,901 multiple-choice items), with and without the image, with the chance reference c and the chance-corrected blind retention ρ_{blind} of §2.2. Retention is reported unclipped; without the image the $f = 1$ model answers BLINK below chance, so its retention is negative.

benchmark	model	with image	no image	c	ρ_{blind}
MathVista testmini	untrained 7B	0.682 (682/1,000)	0.366 (366/1,000)	0.179	0.372
MathVista testmini	$f = 1$, step 30	0.532 (532/1,000)	0.258 (258/1,000)	0.179	0.224
MathVista testmini	$f = \frac{1}{3}$, step 30	0.696 (696/1,000)	0.367 (367/1,000)	0.179	0.364
BLINK	untrained 7B	0.536 (1,018/1,901)	0.416 (791/1,901)	0.377	0.247
BLINK	$f = 1$, step 30	0.457 (869/1,901)	0.363 (690/1,901)	0.377	−0.175
BLINK	$f = \frac{1}{3}$, step 30	0.547 (1,039/1,901)	0.420 (798/1,901)	0.377	0.252

Intervals. Plotted and tabulated intervals are 95% intervals. Each resamples the evaluation unit of its task (the scene program, the grounding pair or the benchmark item) with the trained models held fixed, and Table D.6 names that unit and the run basis for each result.

Constructed coordinate tasks (discovery, the cued levels and the probe; Figures 4(a, b) and 5; Tables G.1, G.2, G.3, H.1, H.2, H.3 and G.5): the scene program, which contributes both twins, is the independent unit. Levels, gains and contrasts resample scene programs with replacement within their pair role, apply the same draws to every compared model and keep the point-estimate weighting (10,000 draws). A cell whose draws are all equal, such as 0 of 200 correct, is reported as its count and number of scene programs, without an interval.

Grounding suite and benchmarks: each grounding pair is its own scene and each benchmark item its own question, so pairs or items are resampled, paired across the compared models: 5,000 draws for the constructed-corpus transfer and the Geometry3K access results (Tables G.4 and D.4), 2,000 for the 7B access pair on the grounding suite and the long-horizon benchmark accuracy, and 10,000 for the long-horizon grounding columns of Table E.1, the grounding endpoint of the slopes in Table H.2, the extraction split of Table D.5 and the audit of Appendix A. Figure 2(a) plots three-seed mean gains with such intervals for cued readout and grounding, and two-seed point estimates without intervals

for discovery; Figure 4 plots the two runs separately. In Table 1 the 7B share carries a bootstrap interval and the three-seed 3B shares are point estimates, with per-seed values in Table D.2.

Other quantities: the identical-wrong-answer rates of Table E.2 carry Wilson intervals, and the overlap of the degraded sets is compared with a 10,000-draw permutation null. The slopes of Table H.2 hold the audited masses fixed.

Seed-level intervals (Table G.7): for a contrast measured on n independent seeds, the interval is the mean $\pm\, t_{n-1,0.975}\,\mathrm{SD}/\sqrt{n}$, with SD the sample standard deviation over seeds ($n = 4$ for REAL against GRAY, $n = 3$ for the named-target ablation). Seeds, not evaluation items, are the unit; the item-resampled intervals of seed 0 remain those of Tables G.2–G.5.

Equivalence test. The grounding-task gain of REAL in §2.3 is compared with the ± 0.05 band by two one-sided tests (Schuirmann, 1987); the band was fixed before pilot training, and the test procedure was chosen afterwards. The test uses the 90% interval of the seed-averaged paired difference, from a 20,000-draw bootstrap over the 600 pairs: [−0.003, +0.040], inside the band. It is distinct from the 95% interval plotted in Figure 2(a), [−0.007, +0.044].

J MATHEMATICAL FOUNDATIONS AND AUDIT COMPARISON

J.1 ANSWER-CHANGING PAIRED EVALUATION

For pair j, write the inputs as $(I_{j,a}, q_j)$ and $(I_{j,b}, q_j)$, and let $\mathcal{A}_{j,a}$ and $\mathcal{A}_{j,b}$ be the sets of extracted answers accepted for its two members. For an answer-changing pair these sets are disjoint: numerical or symbolic equivalence is allowed within each set, but no response matches both gold answers.

Proposition J.1 (image-independent greedy predictors have zero pair accuracy)**.** *Suppose every evaluated pair has the same nonvisual input on its two members, disjoint accepted-answer sets, and a fixed deterministic decoding procedure. If the predictor's extracted answer is independent of the image, its pair accuracy in Eq. (1) is zero.*

Proof. On each pair, such a predictor returns one answer u_j on both members. Hence $C_{j,a}C_{j,b} = \mathbf{1}\{u_j \in \mathcal{A}_{j,a} \cap \mathcal{A}_{j,b}\} = 0$. Averaging these products gives the claim. □

The proposition concerns answer-changing pair accuracy. It does not describe the member accuracy of the constructed discovery tasks or the target-stable identification probe, whose accuracy without the image is measured directly (§6.2).

J.2 AUDIT AND CORRECTNESS FEEDBACK

For item i, let C_{ik}^{image} and C_{ik}^{none} be the binary outcomes of the training reward's accuracy check on audit response k, with $K = 16$ responses in each condition:

$$b_i = \frac{1}{K}\sum_{k=1}^{K} C_{ik}^{\mathrm{image}}, \qquad q_{\mathrm{blind},i} = \frac{1}{K}\sum_{k=1}^{K} C_{ik}^{\mathrm{none}}, \qquad \Delta q_i = b_i - q_{\mathrm{blind},i}. \tag{J.1}$$

These are empirical rates, and the signed difference Δq_i is kept without clipping. Let p denote the underlying probability that a sampled response is correct at a fixed prompt and checkpoint; b_i estimates it for the base model. A group of G independent responses is unmixed only when every response is correct or every response is wrong, so it contains both outcomes with probability

$$P_{\mathrm{mix}}(p; G) = 1 - p^G - (1-p)^G, \tag{J.2}$$

which is positive exactly when $0 < p < 1$ and largest at $p = 1/2$, where it equals $1 - 2^{1-G}$ (15/16 for the training groups, $G = 5$). With $K = 16$ the same expression is the probability that the audit observes both outcomes, $0 < b_i < 1$; that is a sample event, not exact knowledge of p.

Write the reward of response j as $r_j = \frac{1}{2}C_j + \frac{1}{2}F_j$, with F_j the format indicator. With bars denoting means within a prompt's group,

$$r_j - \bar{r} = \tfrac{1}{2}(C_j - \bar{C}) + \tfrac{1}{2}(F_j - \bar{F}). \tag{J.3}$$

When format is constant within a group, correctness variation supplies the reward contrast; format variation can also contribute. P_{mix} therefore measures an opportunity for accuracy-based feedback, not a guarantee of a nonzero GRPO update.

Table J.1 records this at the start of training for the seeds of Table G.6. With real images, 58–67% of discovery groups and 88–89% of probe groups contain both outcomes at step 1 (means over steps 1–5: 0.63–0.66 and 0.60–0.62); with gray canvases, 2–5% of discovery groups and no probe group do, and 95–100% of groups are all wrong. Nonzero advantages nevertheless arise in 61–64% of the gray-canvas discovery groups and 45–48% of its probe groups at step 1: format varies within a group even when correctness does not. That source is short-lived. The nonzero-advantage fraction of the GRAY-trained runs falls to 0.26–0.32 over steps 1–5, their format reward reaches 0.999–1.000 by step 30 (Table D.7), and their mean accuracy reward over steps 19–30 is 0.024–0.041, so by then an update occurs only in the rare group that contains a correct response.

Table J.1: Reward variation within response groups at the start of training, for the runs of Table G.6, by training-row type (discovery, the identification probe, or find-and-bind for the named-target corpus). Columns: mean accuracy and format components at step 1; the fraction of groups containing both a correct and an incorrect response ("mixed", the event of Eq. (J.2)); the fraction of groups with any within-group reward variation ("nonzero advantage"); the same two fractions averaged over steps 1–5; and the step-1 fractions of all-correct and all-wrong groups.

condition	seed	rows	step-1 accuracy	step-1 format	mixed, step 1	nonzero advantage, step 1	mixed, steps 1–5	nonzero advantage, steps 1–5	all correct, step 1	all wrong, step 1
REAL	1	discovery	0.223	0.857	0.633	0.808	0.645	0.698	0.008	0.358
REAL	1	probe	0.435	0.875	0.892	0.933	0.612	0.622	0.025	0.083
REAL	2	discovery	0.178	0.883	0.575	0.800	0.632	0.690	0.008	0.417
REAL	2	probe	0.407	0.892	0.875	0.942	0.598	0.610	0.017	0.108
REAL	3	discovery	0.245	0.875	0.667	0.842	0.655	0.718	0.008	0.325
REAL	3	probe	0.428	0.890	0.892	0.933	0.622	0.637	0.025	0.083
GRAY-trained	1	discovery	0.003	0.862	0.017	0.608	0.025	0.265	0.000	0.983
GRAY-trained	1	probe	0.000	0.875	0.000	0.458	0.003	0.300	0.000	1.000
GRAY-trained	2	discovery	0.005	0.875	0.025	0.617	0.034	0.283	0.000	0.975
GRAY-trained	2	probe	0.000	0.885	0.000	0.475	0.002	0.298	0.000	1.000
GRAY-trained	3	discovery	0.010	0.867	0.050	0.642	0.021	0.255	0.000	0.950
GRAY-trained	3	probe	0.000	0.893	0.000	0.450	0.007	0.317	0.000	1.000
named-target	1	find-and-bind	0.155	0.867	0.517	0.875	0.547	0.635	0.000	0.483
named-target	1	probe	0.435	0.875	0.892	0.933	0.607	0.620	0.025	0.083
named-target	2	find-and-bind	0.137	0.873	0.475	0.842	0.512	0.628	0.000	0.525
named-target	2	probe	0.422	0.892	0.883	0.942	0.635	0.647	0.017	0.100
named-target	3	find-and-bind	0.183	0.883	0.608	0.900	0.585	0.675	0.000	0.392
named-target	3	probe	0.445	0.888	0.900	0.942	0.595	0.608	0.025	0.075

Within a seed, the REAL and named-target runs share the starting checkpoint, the probe rows and the sampling seed, so their step-1 probe rows coincide and diverge only afterwards.

J.3 COMPARING AUDIT MEASUREMENTS UNDER A COMMON SUCCESS PROFILE

Take the audit-joined rows of each condition's realized training stream, keeping their presentation multiplicities, and assign each row the stratum

$$s = \big(\min\{9, \lfloor 10\, b_i \rfloor\},\ \mathbf{1}\{0 < b_i < 1\}\big). \tag{J.4}$$

Let $\mathcal{S}$ be the strata represented in all four conditions, $n_{f,s}$ the row counts, and $\mu_{f,s}$ the mean Δq_i within condition f and stratum s. The common weights and the standardized mass are

$$w_s = \frac{\sum_f n_{f,s}}{\sum_{u \in \mathcal{S}} \sum_f n_{f,u}}, \qquad \widetilde{m}(f) = \sum_{s \in \mathcal{S}} w_s\, \mu_{f,s}. \tag{J.5}$$

Because $\Delta q_i = b_i - q_{\text{blind},i}$, the same weights applied to the stratum means $\bar{b}_{f,s}$ and $\bar{q}_{f,s}$ of those rows give

$$\widetilde{m}(f) = \underbrace{\sum_{s \in \mathcal{S}} w_s\, \bar{b}_{f,s}}_{\widetilde{B}(f)} - \underbrace{\sum_{s \in \mathcal{S}} w_s\, \bar{q}_{f,s}}_{\widetilde{Q}(f)}, \tag{J.6}$$

which separates the with-image and blind components of the standardized audit; within the coarse bins, $\widetilde{B}(f)$ need not be identical across conditions.

Common stratum weights place the audit measurements on a shared initial-success profile. Each outcome remains the measured held-out accuracy of its original training run; the standardization changes the audit summary, not the training distribution. Table H.2 reports the resulting associations and the separate equal-exposure comparisons. The slopes are ordinary least-squares slopes with an intercept over the four conditions, of step-30 held-out accuracy on $m(f)$ or on $\widetilde{m}(f)$; Appendix I describes their bootstrap intervals, which resample held-out scene programs paired across conditions with the audit measurements held fixed.

K THE EXPERIMENTAL PROGRAM AT A GLANCE

The coordinate scenes of §5 are this work's controlled training testbed; the wider program spans nine counterfactual task families, two public training corpora, seven audited benchmarks, four models from three families, and over thirty GRPO runs of up to 400 steps (Figure K.1, Table K.1).

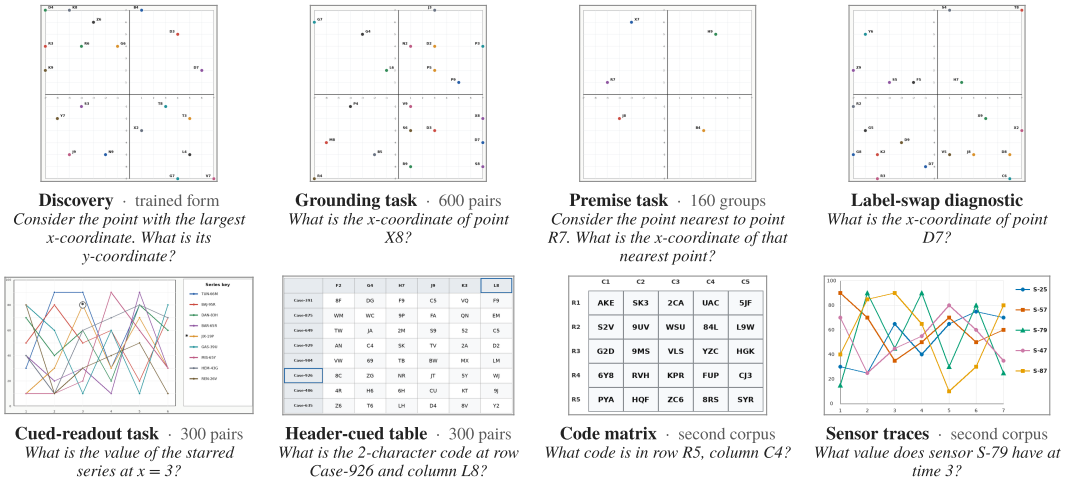


Figure K.1: Eight of the nine task families: coordinate scenes (top) and other formats (bottom).

Table K.1: The experimental program in numbers.

component	scale
constructed evaluations	a 1,200-pair suite and its regenerated 1,200-pair twin; discovery scenes at 8, 12 and 20 points, with 450 confirmatory scene programs; an untouched replication set of 333 scene programs per pair role (Table G.6); 160 premise groups (an untrained nearest-point relation task); 100 label-swap diagnostic pairs
constructed training data	2,880 discovery and probe prompts; a second corpus of 3,000 pairs and 300 answer-preserving twins
public image corpora	Geometry3K (1,288 training, 601 test items); ViRL39K (23,542 training, 4,239 held-out items)
public benchmarks	seven, more than 10,000 items, audited with and without the image at 3B and 7B (3B: Table A.1)
models	Qwen2.5-VL-3B and 7B trained; Gemma-3 and InternVL3-9B audited (Table A.2)
training runs	over thirty GRPO runs of 30 to 400 steps, one to five runs per condition (main-text runs: Table D.1)
base-model audit	6,976 items of the constructed corpus and ViRL39K, each with 16 responses with and 16 without the image

More than one kind of question. The grounding task and its twin predate the training corpus; at step 30, $+0.160$ (95% CI [0.130, 0.190]) of its gain is attributable to training images (GRAY-trained control: $+0.002$). The premise task asks for a nearest-point relation that was never trained: at step 30, final-answer pair accuracy rises from 19 to 33 of 80 pairs at 5 points and from 22 to 28 at 20 points, and is zero under a gray canvas (no GRAY-trained control). On the second corpus (3B, 120 steps, three templates), four configurations improved cued readout ($+0.15$ to $+0.23$) but not the grounding task.

Public images, in training and at test. We trained on both public corpora under four access conditions (ViRL39K, 4,239 held-out items, each condition evaluated as trained: REAL $+0.204$, GRAY $+0.149$). Constructed data was also mixed into ViRL39K training: at $f = \frac{1}{3}$, 160 of every 240 prompts are ViRL39K; never-trained discovery gains $+0.275$ against -0.010 for the all-ViRL39K run, and every mixture's held-out ViRL39K accuracy is at least 0.20 above its image-removed accuracy (§7.1).

Families, scales and horizons. Discovery performance improves at both trained scales (3B: $+0.300$ to $+0.535$ in 30 steps; Table H.3). Two families we did not train, Gemma-3 and InternVL3-9B, score 0.333 and 0.681 on the grounding task, match it on the twin within 0.005, and fall to exactly 0.000 without images. Ordinary training ran 400 steps in two runs (§3); constructed-corpus training ran 100 segmented steps in three runs (two of seed 0, one of seed 1), long after reward saturates at step 13: the grounding task reads 0.948, 0.958, 0.948 at steps 50, 75, 100 (second run 0.960, 0.968, 0.962; seed 1 at step 100, 0.942).

L Extended Related Work

RLVR and what it changes. Tulu 3 named reinforcement learning with verifiable rewards (RLVR), which replaces a learned reward model with a verification function (Lambert et al., 2025), and DeepSeek-R1-Zero optimizes rule-based accuracy and format rewards with GRPO (Guo et al., 2025). Vision-R1, MM-Eureka, and VL-Rethinker extend RLVR to visual reasoning (Huang et al., 2026; Meng et al., 2025; Wang et al., 2025a). Whether RLVR develops new skills or elicits existing ones is open: pass@k analyses find that base models can match or exceed RL-trained ones as sampling budgets grow (Yue et al., 2025b), and some Qwen models improve at mathematics under random or incorrect rewards (Shao et al., 2026). Process supervision instead scores each reasoning step: a reward model trained on step-level human feedback reranks sampled solutions to mathematics problems, without RL (Lightman et al., 2024). Our visual-access comparisons hold the verifier and training setup fixed within each corpus: blind training recovers much of the real-image gain on Geometry3K (§2.1) but, at matched step 30, essentially none of the discovery gain from training on visually resolvable tasks (§6.2).

Visual dependence in evaluation. Controlled contrasts expose shortcuts that sustain accuracy without the intended computation (Geirhos et al., 2020). VQA v2 counters language priors with complementary images (Goyal et al., 2017); VQA-CP changes the answer distribution of each question type between training and test while keeping the images' distribution, and existing models degrade sharply (Agrawal et al., 2018); MMVP probes visual differences between images with similar CLIP representations (Tong et al., 2024); and BLINK targets perception-intensive tasks (Fu et al., 2024). MMStar filters out text-answerable questions and evaluates the same vision-language model with and without images, while its leakage metric measures image-free improvement over the language-model backbone (Chen et al., 2024); MMMU-Pro also filters out text-answerable questions (Yue et al., 2025a). Blind gains probe the training stage instead: visual information is withheld during RL and restored at evaluation, so an improvement can depend on the image at test without having been learned from images. Our benchmark audit measures dependence by chance-corrected blind retention (§2.2), and programmatic pairs with controlled target cues identify which operation training improves (§2.3).

Paired scoring. MMVP scores a CLIP-blind pair only when both of its questions are answered correctly (Tong et al., 2024); NaturalBench pairs each question with two images that yield different answers, and its question accuracy awards a point only when the question is answered correctly for both images (Li et al., 2024). The pairs of our counterfactual suite are generated programmatically: the question stays fixed and one visual fact changes so that the two correct answers differ. Eq. (1) applies the same rule to these pairs, and because their accepted answers are disjoint, an image-independent greedy predictor scores exactly zero (Proposition J.1). The guarantee concerns pair accuracy; for discovery and its identification probe, scored per question, accuracy without the image is measured directly.

Visual skill under reward optimization. Multimodal reasoning models hallucinate more than their non-reasoning counterparts on perception tasks, and longer reasoning chains reduce attention to visual inputs (Xu et al., 2025). Reasoning-only RLVR on Qwen2.5-VL-7B lowers an out-of-domain segmentation score below the base model's while mathematics scores rise (Phan et al., 2026); VOLD, which transfers reasoning to a vision-language model with text-only training, reports a small perception decrease that an image-trained reasoning model shares and attributes it to reasoning-focused training rather than to text-only data (Bousselham et al., 2026); and after RLVR on Geometry3K with only accuracy and format rewards, perception on 50 MathVista problems does not differ significantly from the base model's (Xiao et al., 2026). Grounding corrosion (§3) is a training-time deterioration measured on counterfactual pairs: under prolonged real-image training with a trainable vision encoder, grounding pair accuracy falls below the base model's while benchmark gains persist. In a separate analysis, independent blind-trained runs lose many of the same previously solved pairs, usually with identical wrong answers that real-image-trained runs often repeat. Aggregate scores show the decline; recurring errors show which examples independent runs lose and which wrong answers they share.

Perception-aware reinforcement learning. PAPO adds to the objective an implicit perception loss that maximizes the KL divergence between response distributions under original and masked images, leaving the reward unchanged (Wang et al., 2026); Visionary-R1 imposes a caption–reason–

answer structure and rewards captions from which the question can be answered (Xia et al., 2026); DeepEyes rewards image-cropping tool use when the answer is correct and, for its fine-grained perception data, keeps examples that crops make solvable (Zheng et al., 2026). Perception-R1 adds a visual perception reward, judged by a language model, to the accuracy and format rewards (Xiao et al., 2026), and Visual-RFT trains on perception tasks with task-specific verifiable rewards, such as IoU for object detection (Liu et al., 2025). A second method also named Perception-R1 trains perception policies with GRPO and rule-based rewards (IoU for grounding, Euclidean distance for counting, edit distance for OCR) and observes that perceptual perplexity is a major factor in RL's effectiveness (Yu et al., 2025a). These methods act through the objective, the response structure, added evidence, or the reward, and the last two also through the task. We use standard GRPO with the correctness-and-format reward fixed within the constructed-corpus comparisons. Task design controls the visual fact relevant to the answer and the target information supplied (§5.2). A matched blind-trained control, with the same prompts, starting checkpoint, reward, and budget, then identifies how much of the gain comes from training-time visual information (§6.2).

Learnability, visual necessity, and data selection. DAPO's dynamic sampling keeps prompts with both successful and unsuccessful responses (Yu et al., 2025b); ThinkLite-VL selects data by the tree-search iterations a problem needs (Wang et al., 2025b). Both measure how hard a problem is for the model. Our learnability band, $0 < b_i < 1$, shares this concern; it is measured once on the base model and describes the corpus rather than filtering it (§5.1). VisNec uses the name "visual necessity" for the loss reduction the image provides under teacher forcing and selects instruction-tuning data with it, without RL (Dong et al., 2026); our Δq_i is a difference of sampled correct-answer rates with and without the image. This measure adds what DAPO and ThinkLite-VL lack: how much the image raises the base model's probability of a correct answer, which characterizes the training mixtures whose composition shapes acquisition (§7.1). Difficulty-based selection asks how hard a task is; resolvability also asks what evidence and operation successful answering require.

Two verifiable vision-centric training tasks are the closest precedents for task construction. ViCrit (Wang et al., 2025c) injects a small error into a human-written caption and rewards locating the corrupted span given the image, on the grounds that vision-centric tasks that are both challenging and verifiable are scarce; Visual Jigsaw (Wu et al., 2026) shuffles image patches, video clips or 3D views and rewards recovering the order, a self-supervised task that needs the image by construction. Both build tasks whose reward requires perception and report benchmark gains. Neither measures whether the gains are learned from the images (the matched blind-training control of §6.2), which visual operation improved (the cue hierarchy of §2.3), or what the same reward teaches when the task admits a shortcut (the shortcut-accessible twin of §6.2); those three measurements are what this paper adds. VisNec uses "visual necessity" for the reduction in teacher-forced loss supplied by the image and selects instruction-tuning data with it (Dong et al., 2026); our Δq_i measures the change in sampled correctness under the RL verifier.