IDENTIFYING, INSTALLING, AND REMOVING CAPABILITIES IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the computations installed by low-rank fine-tuning (LoRA) for translation, scale, and rotation tasks on a procedurally generated spatial benchmark in which one of two otherwise identical point clouds is transformed. Closed-form solutions to these simple stimuli provide both exact algorithms and cheaper statistics that are correct over part of the parameter range but fail over the rest. When asked in what direction one cloud of points must rotate to align with another, two 7B models fine-tuned on stimuli with rotation offsets between $0°$ and $45°$ do not calculate the answer based on correspondence between points, but instead compare the apparent orientation of each cloud. The fingerprint of this is evident on a full 0–180° set of evaluation stimuli, where performance alternates between high and inverted accuracy, with period $90°$ on statistically round clouds and $180°$ on elongated ones. A parameter-free reader matches these response curves by measuring the angle between the two clouds' principal axes and wrapping it into the symmetry period, agreeing with both model families. A full-range computation can be installed by training over the full 0–$180°$ range at 1,200 examples per task when the points in the two clouds are labeled with matching numbers (cues), or at 18,000 examples per task on cue-free stimuli, where the model obtains one correspondence from the point farthest from the centroid. We also demonstrate that measured computations can be severed from model readout. A model competent at x (left/right), y (up/down), and their XOR can be tuned to emit a single constant token on x and y ("refuse" to answer), while XOR holds at perfect accuracy. The refused model's output still flips when either sign flips, so it remains sensitive to the quantities it will not verbalize. This severing costs a fraction of the supervision that installed the initial competencies. When framed as capability removal, our experiments demonstrate that, across multiple losses designed to provide readout suppression and/or algorithm destruction, such removals are successful but not durable. Reinstallation of removed verbalization or computation requires 13 to 800 times less supervision than initial installation.

1 INTRODUCTION

Open-weight vision–language models (VLMs) are poor at spatial reasoning (Kamath et al., 2023; Fu et al., 2025; Rahmanzadehgervi et al., 2024; Tong et al., 2024). Fine-tuning increases accuracy on spatial benchmarks but leaves the enabling calculations obscured. A model that answers correctly on held-out data may perform a general, task-appropriate computation, or may have learned a cheaper shortcut statistic. The former is required to instill VLMs with general spatial competence. These scenarios can be separated using stimuli on which shortcuts and calculations disagree; however, a task's true computation must be (and is generally not) known. In this work we use a benchmark with known, closed-form computations specifically designed to probe spatial reasoning in VLMs. The Sophon benchmark (Anonymous, 2026a) procedurally generates identically distributed point clouds differing by color and one transform (translation, scale, or rotation). Each transform can be solved in closed form, and cheap statistics match the solution only over certain distribution ranges. Minimal-cue stimuli and simple questions allow direct examination of installed computations, provide a testbed for installing different computations, and provide a method to measure the success and durability of capability removal.

Two point clouds related by a similarity transform have a closed-form alignment (Kabsch, 1976; Umeyama, 1991). Translation is the difference of centroids, scale is a ratio of second moments

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.

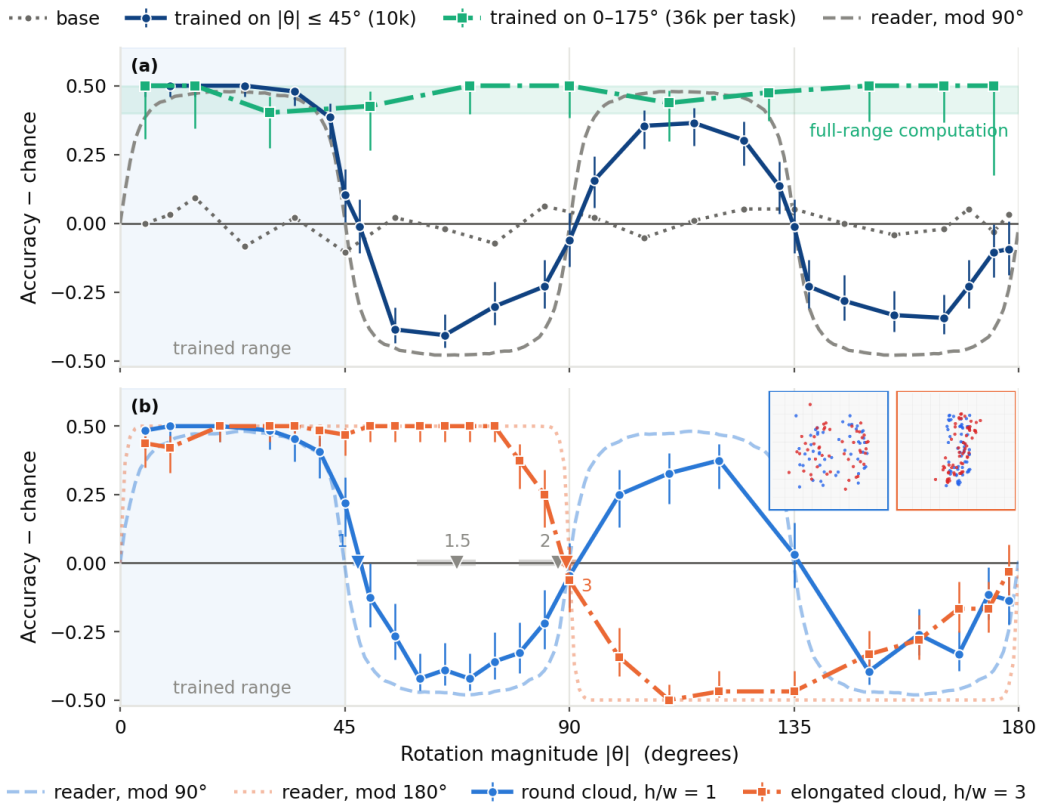

Figure 1: (a) Rotation accuracy over the full circle for a OneVision model trained on $|\theta| \leq 45°$ (shaded), plotted as accuracy minus chance. Every band is answer-balanced, so a value below zero means the model answers backwards. Points are the model after 10,000 LoRA examples per task (Wilson 95% intervals, 96 stimuli per band), the untrained base model on the same stimuli, and the six-task Qwen2.5-VL model of Section 8, trained over the full 0–$175°$ range for 36,000 examples per task, which sits at the ceiling in every band. The dashed curve is the zero-parameter orientation reader: the response a model that matches orientations would give. (b) The same sweep on round (aspect 1.0, circles) and elongated (aspect 3.0, squares) clouds for the model at 10,000 examples per task, with the reader wrapped to $90°$ and to $180°$ respectively (dashed), and the angle at which accuracy crosses chance marked on the zero line for aspect ratios 1.0, 1.5, 2.0, and 3.0 ($47.5°$, $66.2°$, $88.2°$, $90.8°$).

about the centroids, and rotation is the polar factor of the cross-covariance of the centered sets. The three quantities require one, two, and three stages of composition respectively, and rotation leaves position-marginal statistics of a cloud unchanged. Cheap statistics that agree with the closed-form solutions on the training distribution are listed in Table 1, along with stimuli designed to force the cheap statistic and the true answer to disagree.

Figure 1 demonstrates the difference between a true rotation computation and a statistic that works over the training distribution. The failure of the statistic is structured, and its structure can be changed by the design of the evaluation stimulus alone. Tuning a VLM to answer "in what direction should the points rotate" on stimuli offset by at most $45°$ yields a model that compares the apparent orientations of the two clouds. This shortcut is defined only modulo the cloud's shape symmetry, and accuracy over the full circle alternates with period $90°$ on round clouds and $180°$ on elongated ones. A reader built from the pixels with no fitted parameters predicts this curve (Section 4).

Because we know exact computations at play, we can also ask what makes a model learn, and whether learned computations can be moved or removed. Correspondence cues cut the supervision cost of full-range rotation about $15\times$, and the cue-free computation can be transplanted into a student by matching activations alone, with no answer labels (Sections 5 and 6). A model can be tuned to refuse

to answer questions it still computes (Section 7). Five removal losses applied to the rotation circuit are successful, but none durable. Reinstalling the capability takes 13 to 800 times less supervision than initial acquisition (Section 8).

Contributions. (1) A method for identifying the statistic a fine-tune installed: derive the cheap statistics of a known algorithm, predict each one's full response curve with a zero-parameter reader and a symmetry sweep, and construct stimuli on which it and the algorithm disagree (Sections 4 and 5). (2) A representation transplant: activation matching with no answer labels installs the teacher's computation in a student on a different rendering, controls show the computation and not its activation statistics is what moves, and an unsatisfiable matching target blocks the supervised route from forming (Section 6). (3) Readout control: a model that computes two quantities but verbalizes neither, while a composite is maintained (Section 7). (4) A test of removal durability framed as reacquisition cost against acquisition cost, demonstrated on five removal methods (Section 8).

2 Related work

Shortcut learning. Networks prefer statistics that are sufficient on the training distribution (Geirhos et al., 2020; McCoy et al., 2019; Hermann et al., 2020). Our training data is designed such that the target algorithm is known and decomposable, and shortcuts are known in advance. Further, these shortcuts are separable from true computation by designed evaluation stimuli. This allows direct identification of the shortcut a model learns.

Spatial failures of VLMs. Several benchmarks document failures on simple spatial questions (Rahmanzadehgervi et al., 2024; Tong et al., 2024; Fu et al., 2025; Kamath et al., 2023); the Sophon benchmark used in this work (Anonymous, 2026a) removes semantic content so that failures cannot be attributed to scene semantics. A companion paper (Anonymous, 2026b) studies what a fine-tune must change for the frozen last five decoder layers (the reader or motor regime of Gurnee et al. (2026)) to report a quantity that the layers below them (the workspace band) already encode. Here we study which computation the fine-tune installs and how that computation can be replaced or silenced.

Distillation and representation matching. Matching intermediate activations to a teacher model is standard practice (Romero et al., 2015; Hinton et al., 2015). Our geometry-matched stimulus pairs allow the application of this toolset using identical underlying scenes across different renderings. In addition, three controls (scrambled scene, annealed weight, zero labels) are used to distinguish an installed computation from mere alignment of activation statistics.

Refusal, unlearning, and tamper resistance. Refusal in language models is mediated by low-dimensional structure (Arditi et al., 2024), safety alignment is often shallow in token depth (Qi et al., 2025a), and unlearning evaluations find that removed knowledge is usually recoverable from model weights (Li et al., 2024; Eldan & Russinovich, 2023; Lynch et al., 2024; Deeb & Roger, 2024). For open-weight models the key concept is durability of removal to additional fine-tuning (Tamirisa et al., 2025; Qi et al., 2025b). We demonstrate tuned refusal-to-answer for any isolated transform while leaving the quantity linearly decodable (Section 7). Section 8 measures durability as the ratio of reacquisition supervision to installation supervision.

3 Setup

Stimuli and the exact algorithm. Each Sophon example shows a reference cloud and a moved cloud of n points related by $Q = sRP + t$. The categorical answers are left/right, up/down, inward/outward, and clockwise/counterclockwise, and the quantitative answers are the unsigned angle of rotation or the scale factor. With correspondence known, t, s, and R have closed forms (Kabsch, 1976; Umeyama, 1991): translation is a difference of centroids, scale a ratio of second moments about the centroids, and rotation the polar factor of the cross-covariance of the centered sets.

Proxy inventory. For each question type, Table 1 lists the cheap statistics that agree with it on the training distribution and the stimulus on which each disagrees with the truth. The leading rotation proxy is the difference of the clouds' principal-axis orientations, which is $180°$-periodic for an

answer asked for	cheap statistic that agrees on the training distribution	constructed stimulus on which it and the exact answer disagree
rotation direction (clockwise or counterclockwise)	difference of the two clouds' apparent orientations	rotations past the cloud's symmetry half-period (45° for a round cloud, 90° for an elongated one), where the orientation difference wraps and its sign inverts
rotation direction	displacement of one salient point	a second point at the same radius as the farthest one, so the single correspondence is ambiguous
scale direction (outward or inward)	ratio of bounding extents	a core that expands while a few outlying points collapse, so the extent shrinks as the cloud grows
scale direction	ratio of RMS radii	nearly exact; no stimulus in this battery separates it from the true scale
translation direction (left/right, up/down)	difference of centroids	exact for the categorical answer; no disagreeing stimulus exists

Table 1: The proxy inventory, abridged. Every task asks for the direction of one transform, so the answer is one of two words. Each row is a statistic that predicts that word on the training distribution and the stimulus constructed so that it and the exact algorithm give opposite answers.

elongated cloud and defined only modulo 90° for a round cloud with approximate 4-fold symmetry. A model running it must invert past the half-period and must fold a reported angle at the same period.

Models and recipes. We use Qwen2.5-VL-7B (Bai et al., 2025) and LLaVA-OneVision-7B (Li et al., 2025) with LoRA (Hu et al., 2022) on the workspace, decoder layers 0–22 ($r = 32$, $\alpha = 64$), with the reader and the vision tower frozen. Supervision checkpoints are used to bracket the onset of each capability. The models see no duplicate stimuli during training. One scene from every battery is shown in Appendix I (Fig. 19); the prompts and every hyperparameter are in Appendix J (Tables 6 and 7).

Zero-parameter readers. For each candidate statistic we build a reader from the rendered pixels (color segmentation, the statistic, its sign) with no learned or fitted parameters, and correlate the response curve with model outputs. A reader that predicts the curve of a model's accuracy as a function of stimulus angle has identified the computation. Predicting in-training range accuracy alone is insufficient.

Evaluation. Rotation sweeps bin stimuli into 10°-wide bands of $|\theta|$, each with balanced clockwise and counterclockwise representation. Every evaluation cell is balanced by construction, so a constant emitter scores exactly 0.5. Accuracies carry Wilson 95% intervals.

4 The installed rotation computation is orientation matching

A model fine-tuned on rotations of at most 45° was evaluated in bands from 5° to 178°, 96 stimuli per band (Fig. 1). The signed departure from chance (where 0.5 = perfect, 0.0 = chance, -0.5 = inverted) averaged over each quarter of the circle is $+0.49$, -0.33, $+0.29$, and -0.32 after tuning on 10,000 examples per task. The model is correct in the first and third quarters and inverted in the second and fourth. The untrained model on the identical stimuli answers at chance, which excludes the frame, the rendering, and per-band stimulus differences as causes. The zero ("chance") crossings are at 48°, 91°, and 135°, consistent with a $\sin 4\theta$ response. Supervision increases the amplitude rather than removing the pattern. Mean absolute departure from chance is 0.13, 0.25, 0.27, and 0.32 at 400, 900, 2,700, and 10,000 examples, and at 10,000 the model scores 1.00 at 15° and 0.17 at 155°. Within the training range, accuracy holds at 1.00 through 25° and falls to 0.60 at 45°; above 170°, where the two directions become indistinguishable, accuracy returns toward chance.

The $\sin 4\theta$ response is indicative of a 4-fold symmetry produced by a round point cloud, and can be tested by changing the elongation of point clouds in evaluation stimuli. Indeed, when clouds are elongated to an aspect ratio of 3, the period of inversion becomes 180° (Fig. 1), and the chance crossing moves from 47.5° at aspect ratio 1 to 90.8° at aspect ratio 3. The same fine-tuned model oscillates between high accuracy and inverted behavior twice on round clouds and once on elongated ones. Only a model computing a signed quantity and reporting it backwards can invert to performance

below chance. Beyond inversion, a statistic defined modulo $180°$ also predicts that accuracies in bands mirrored around $90°$ must sum to unity. We measure sums of 1.00 for both mirrored pairs in the battery, $45°/135°$ and $70°/110°$. Finally, in-distribution accuracy ($\leq 45°$) rises with elongation (from 0.91 at aspect 1 to 0.99 at 3), so the inverted band is not an effect of distribution shift. In the far tail (150–$178°$) both aspect ratios rise to chance and stop, implying that no other computation takes over where this one fails.

We use a reader algorithm to probe the orientation mapping hypothesis. This reader segments the two clouds by color, takes each cloud's principal axis, wraps the difference to the period, and takes its sign. The reader's measured result (also plotted in Fig. 1) predicts the response curve with $r = +0.971$ and $+0.987$ on OneVision's round and elongated clouds and $+0.959$ and $+0.961$ on Qwen2.5-VL's. Appendix E tests the reader in depth on in-distribution stimuli and a battery where the principal axis and the rotation disagree.

We test Qwen2.5-VL across eight aspect ratios from 1.0 to 3.0 and find that the crossing angle correlates -0.93 with the width-to-height ratio (1 for a round cloud, $1/3$ at the largest elongation). A similar test of OneVision over four ratios gives -0.97. The same statistic appears when the models are asked for the quantitative (unsigned) angle of rotation (Fig. 13). The model's reported angles correlate with $\theta \bmod 90°$ on round clouds and with $\theta \bmod 180°$ on elongated clouds, and with the true angle at $+0.32$–0.35 in both cases.

The two architectures have different vision encoders and different language models, and yet their accuracy response curves correlate $+0.989$ (Fig. 14). Qwen2.5-VL's rotation accuracy leaves chance by 100 examples per task and OneVision's by 400, and both reach 0.99 on held-out in-range stimuli ($n = 80$; Qwen2.5-VL at 2,700 examples, OneVision at 10,000), so an in-distribution evaluation would report rotation as solved.

A similar analysis of the statistics learned when fine-tuning scale is in Appendix A.

Why the cheap statistic is the one installed. Anonymous (2026b) find that rotation, unlike translation and scale, has no position-marginal statistic that a single pooling step can gather, and is computed by the fine-tuned decoder in several serial read/write stages. Orientation matching is a single-stage statistic, and over the $\pm 45°$ training range a $90°$-periodic readout agrees with the truth on 0.82 of stimuli and a $180°$-periodic one on 0.99. The exact computation costs about 18,000 cue-free examples per task to install (Section 5), while the statistic leaves chance within a few hundred and reaches 0.99 in range by 2,700 (Qwen2.5-VL) to 10,000 (OneVision) examples (Fig. 14). A training range that never crosses the half-period never penalizes the cheaper install.

5 Supervision cost of the exact computation: identity cues

If the cheap statistic wins because correspondence is expensive to compute, supplying correspondence should reduce the cost of the exact computation. With each of ten points labeled with a digit that matches its partner, Qwen2.5-VL (the only family from here on) learns full-range rotation to 0.944 at 1,200 examples per task and 0.992 at 1,800 (Fig. 2). The same scenes drawn as plain dots (same generator seed, identical geometry) are at constant emission through 2,700 examples and slowly improve to 0.940 at 18,000. The supervision ratio is 12–15$\times$ across draws of the LoRA initialization (triangles in Fig. 2). On the same 266 scenes the marked arm scores 0.9925 [0.973, 0.998] at 6,000 and the unmarked arm scores 0.951 [0.918, 0.971] at 21,000, its best checkpoint. The marked arm alone is correct on 13 scenes and the unmarked arm alone on 2 (exact McNemar $p = 0.007$).

Appendix H gives the pooling analysis of the cue-trained model and the cross-rendering table.

What the cue-free model calculates. We build a disagreement battery which rotates m of the ten points by $+\theta$ and the other $10 - m$ by $-\theta$, for every m from 0 to 10 in two angle bands (880 scenes; the tie at $m = 5$ is unscored). Model answers are scored against the majority (Fig. 18). A reader that follows a single landmark ramps in m, while a fit over all ten points steps at $m = 5$. On plain-dot rendering, the cue-free model at 18,000 and 21,000 examples shows three things. First, this is not the orientation statistic of Section 4: with all ten points in agreement the model scores 0.96–0.99 in both bands, including 110–$160°$, where that statistic scores 0.00. Second, this is not a full correspondence fit over all ten points: accuracy ramps across m, agreeing with the single-landmark reader within 0.07–0.09. Third, the landmark is the point farthest from the centroid. Splitting the 640 scenes with

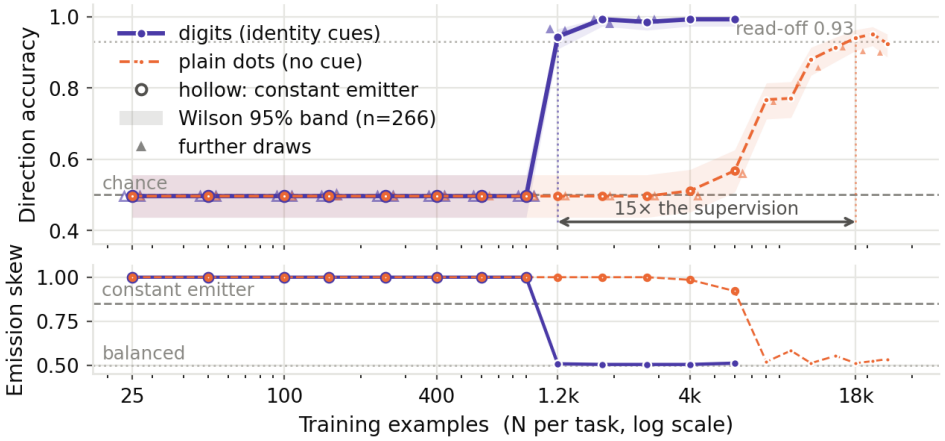

Figure 2: Full-range rotation accuracy against examples per task, with each point wearing an identity digit (marked) and with the identical geometry drawn as plain dots (unmarked). The evaluation set is 52.6% above 90°. Marked training reaches 0.94 at 1,200 and 0.99 at 1,800; unmarked training is at constant emission through 2,700, leaves it at 6,000, and reaches 0.94 at 18,000. Triangles are further draws of the LoRA initialization and dropout on the same data in the same order (two marked, one unmarked); open triangles are constant emission.

a minority by where the farthest point falls, the model is correct at 0.78 when that point is in the majority and 0.58 when it is in the minority ($z = 5.1$). No other point tested moves accuracy by more than 0.09. The cue-free model therefore takes one correspondence from the extremal point, the salient-point statistic of Table 1, and weights it heavily but not exclusively. On ordinary rotations the model scores 0.95 overall, 0.99 when one point is clearly outermost, and 0.89 when no point stands out (Appendix F).

The two models compute different things. Scored on each other's rendering, the models are asymmetric (Table 5). The cue-trained model falls to 0.54 without cues; the cue-free model retains 0.80 with them. A scrambled-cue control assigns numbers at random so that partners carry different digits. The cue-trained model falls to chance or below at every count k of cued points, so it used true correspondence rather than the presence of numbers. The cue-free model is unchanged at every k (0.801 with matched and with scrambled digits at $k = 10$), so it ignores the cues, and its decline with k is a distribution-shift effect.

6 Installing the cue-free computation by activation matching

A student model supervised only on digit renders is trained to match, at every vision-token position of the workspace, the activations of the frozen cue-free teacher on the paired dot rendering. This is accomplished via relative mean squared error alongside the ordinary answer loss. If successful, the student should answer dot scenes the way its teacher does, while a digit-trained student without the matching loss should fail like the cue-trained model of Section 5. With the matching loss the student scores 0.81 on dots, against the teacher's own 0.80; without it, the same training budget gives 0.55 (Fig. 5). Above 90° the transplanted student scores 0.77 on dots, where the orientation statistic would score near zero, so what moved is the cue-free computation. The cost to the original supervised task is minimal: the student model reaches 0.94 on digits one checkpoint later than the baseline digit-tuned model. Further, the answer loss is not even required. By matching the vision tokens and the answer-position stream with no labels during tuning, a model reaches 0.90 on digits and 0.85 on dots at 6,000 examples. This model rises above chance between 400 and 600 examples, where supervised training needs 900 to 1,200. Matching the teacher's activations on a different scene installs nothing and blocks the supervised route, and a matching weight annealed to zero mid-run lets the installed computation decay back to the baseline.

Appendix B provides the probe fingerprint of the transplanted student, the three controls (the zero-label run, the scrambled targets, and the annealed weight), and the blocking dose.

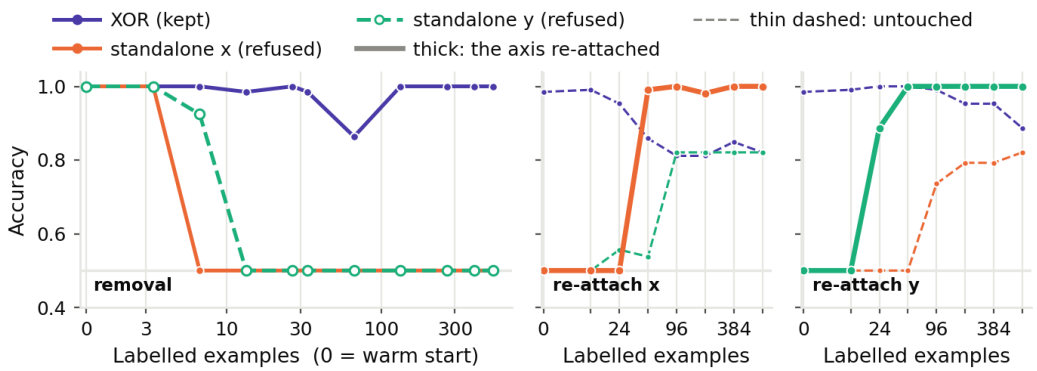

Figure 3: Refusing two readouts while their composite is maintained. Left: a cue-free Qwen2.5-VL checkpoint competent at standalone x, standalone y, and their XOR (the warm start at zero examples) is trained with a constant-token loss on the two standalone questions and the ordinary loss on XOR. x and y collapse to a single token within about ten examples while XOR stays at 0.99. Middle and right: ordinary supervision on the refused checkpoint on one axis alone reattaches rapidly, while the untouched axis (dashed) drifts back partway. Holding it refused is shown in Appendix G.

7 Readout control: refusing two quantities while their composite is maintained

Section 6 showed that a matching loss installs a computation when its target is coherent and blocks one when it is not. Here we apply the simplest such loss, cross-entropy toward a constant token, to a readout. This tests whether a model can compute a quantity yet be trained not to report its answer.

Stimuli carry a diagonal translation with independent signs and three questions: standalone x (left or right), standalone y (up or down), and their XOR (aligned or crossed), which needs both signs but reveals neither. A cue-free Qwen2.5-VL checkpoint competent across all three is trained toward a fixed token on the two standalone questions and with ordinary loss on XOR. Both standalone questions must be refused: with y and XOR still answered, x follows. A scale variant that refuses the direction while keeping the numeric factor fails the same test, since the factor's sign relative to 1 gives the direction away (Appendix C).

The refusal forms in about ten examples per axis (Fig. 3, left): x is a single token by 7 examples and y by 13, while XOR stays at 0.98–1.00 through the onset and at 1.00 from 130 examples on. Both refused readouts sit at 0.50 with zero emission entropy through 530 examples ($n = 66$ per task per checkpoint). Both signs remain decodable at the answer position of the refused checkpoint. Fifteen prompt-side attacks and in-context demonstrations leave the refusal in place; a soft prompt trained with the weights frozen does not, reopening either readout in 96–192 labeled examples. With trainable weights, 24–48 labeled examples re-attach a readout (Fig. 3, middle and right). Appendix G expands each of these.

The result demonstrates selective verbalization control against discrete prompts on a model whose weights are fixed. Both signs remain decodable at the answer position, the XOR that depends on them stays at 1.00, the refusal is specific to two near-identical questions, and no retained answer inadvertently answers a refused one. This represents a removal without durability. A soft prompt or a fine-tune with labels reopens either readout in tens of examples, and the computation it acts on cost only hundreds of examples to install. Section 8 repeats the measurement on the rotation direction, which cost about 20,000 examples per task to install, and adds losses aimed at the computation itself.

8 Removal durability: the reacquisition cost of the rotation direction

We measure the durability of a removal as a ratio of the labeled task examples needed to reacquire performance over the examples required to initially install the capability. This framing allows for benchmarking of future methods. Specifically, we require a removed capability to rise to 0.90

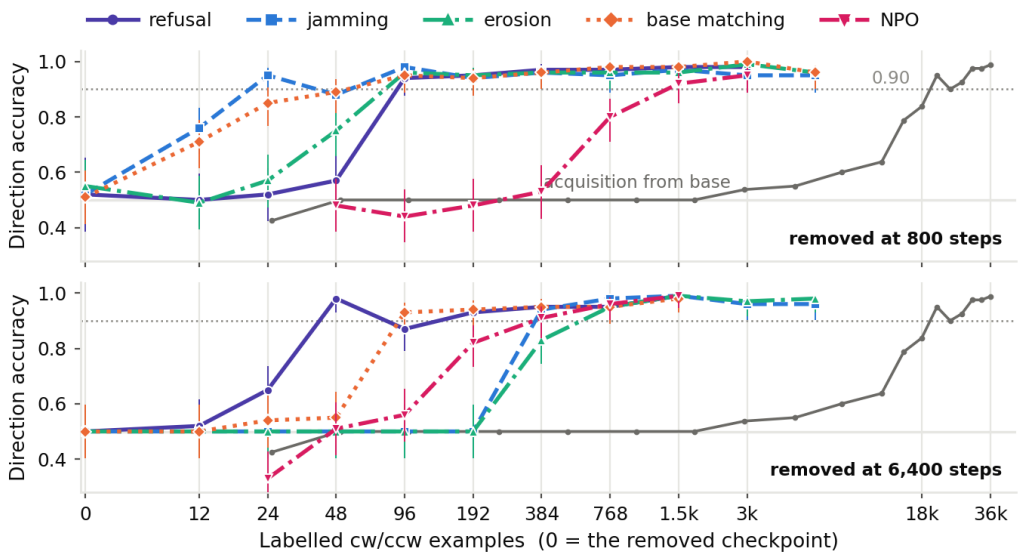

Figure 4: Re-attaching the rotation direction after each removal: direction accuracy on 100 held-out scenes per checkpoint against labeled clockwise/counterclockwise examples, continuing the adapter that made the removal (Wilson 95% intervals). Top: from the 800-step checkpoint of each loss, with acquisition from the base model in grey; zero examples is the removed checkpoint. Refusal, jamming, erosion, and base matching are back at 0.90 by 24 to 96 examples; negative preference optimization (NPO) takes 1,536. NPO's checkpoint and its first re-attach steps give no parseable answer on most scenes and are not plotted, so its curve begins at 48 examples (24 in the bottom panel). Bottom: the same five losses removed for 6,400 steps; the longer dose holds jamming and erosion at chance through 192 examples and moves their cost to 384 and 768, leaves base matching at 96, and lowers the refusal's to 48 and NPO's to 384.

accuracy on 100 held-out scenes per checkpoint. For this section we focus on categorical rotation questions with answers of clockwise / counterclockwise. The expensive nature of this install provides significant dynamic range for measuring reacquisition costs. We use two regimes: *continued adapter*, in which the adapter that made the removal keeps training with ordinary labels on the direction, which asks how shallow the removal is inside its own parameters; and *fresh adapter*, in which the removal adapter is merged into the weights and a new LoRA is trained on top, which is what anyone holding the released weights would do. The fresh adapter cannot reverse the removal update, only rebuild the capability, so the two regimes could differ widely; that they agree to within a factor of four everywhere below is itself a finding.

The starting checkpoint is a cue-free Qwen2.5-VL model trained from the base model on six tasks together (the four directions, quantitative angle and scale factor) for 36,000 examples per task, a separate run from the single-task model of Section 5. This model answers rotation direction at 0.97 over the full circle on 300 held-out scenes and the angle at a mean absolute error of 5.6°, with scale and both translations at 1.00. During training, rotation direction remained at chance through 1,800 examples per task and reached 0.84 at 18,000 and 0.95 at 21,000, so about 20,000 labeled direction examples were required, and we adopt that as our durability denominator.

Refusal. Training the answered direction toward a fixed token for 800 steps, with scale, both translations, and the angle maintained, leaves the direction accuracy at 0.52 and the angle MAE at 8.2°, and the retained tasks at 1.00. Unlike the two-axis case in Section 7, where models learned to emit single tokens, here the refused rotation readout is at chance. Refusing scale or either translation instead behaves the same way and leaves the other three tasks at 0.96–1.00 (Fig. 11 in Appendix D). Rotation direction jumps back to 0.94 by 96 labeled examples with the continued adapter, at 0.95 with the angle maintained, and at 0.92 with a fresh adapter on all layers (Fig. 4; the regimes and the angle are in Fig. 10). When reacquisition focused on direction-only stimuli, the angle MAE drifts to 17.9° at the onset and recovers by 3,072 examples. With the quantitative examples also

used in reacquisition, MAE stays below 10° throughout. The ratio of reacquisition to acquisition is a negligible 5×10^{-3}.

Losses aimed at the computation. Three losses target the computation itself, each applied from the same 36k checkpoint with cross-entropy maintaining scale and both translations. We save checkpoints for study at 800, 1,600, 3,200, and 6,400 steps. *Jamming* matches the answer-position and vision-token activations of the workspace on a rotation scene to the base model's activations on a different scene. *Erosion* targets a rotation subspace of the 36k model. At each workspace layer, we calculate the three directions along which the activation covaries with the sign, the sine, and the cosine of the angle. A probe on that subspace reads the direction at 0.98, and zeroing it at inference drops the direction from 0.98 to 0.85 and the angle error from 5.4° to 40.5° while scale and translation are untouched. The loss penalizes the squared projection of the activations onto it. *Base matching* matches the base model's activations on the same scene. At 800 steps all three leave the direction at 0.51–0.55 and the angle prompt non-numeric, with scale and both translations at 0.98–1.00. A fourth loss is the standard unlearning objective, negative preference optimization (NPO) (Zhang et al., 2024) on the rotation prompts with the pre-removal adapter as the frozen reference ($\beta = 0.1$). At 800 steps it leaves the rotation prompt without a parseable answer on 99 of 100 scenes, the retained tasks at 0.98–1.00, and drives the slowest re-attach of the set: near chance through 384 examples, 0.80 at 768, and 0.92 at 1,536 (Fig. 4).

On the in-distribution battery the five removals are indistinguishable. Against the base model's 0.94 on What'sUp CLEVR and 0.89 on BLINK spatial relations, the refusal costs the most on CLEVR (0.59), and erosion costs the most on BLINK (0.63, a format collapse that deepens with re-attachment). Jamming, base matching, and NPO stay within 0.05–0.16 of the base on both (see Appendix D for full details).

From the 800-step checkpoints the direction is back at 0.90 by 24 examples after jamming and by 96 after erosion and base matching; the fresh adapter needs 96 and 192 (Fig. 9; Table 2 in Appendix D has the numbers). Jamming re-attaches sooner than the refusal. Dose separates the losses (Fig. 4, bottom): at 6,400 steps jamming and erosion hold the direction at 0.48–0.54 through 192 examples and move the cost to 384 and 768 in both regimes, eight to sixteen times the 800-step cost and about $1/30$ of acquisition; base matching still re-attaches at 96; and the refusal and NPO re-attach sooner, at 48 and 384 rather than 96 and 1,536. Appendix D gives the layer-by-layer measurements of what each loss removed, injection at the last trained layer, the full-circle fingerprints of the re-attached models, and the seed replicates.

9 Discussion

On Sophon spatial tasks, a low-rank fine-tune installs the shallowest statistic that is sufficient on its training distribution, at every level of supervision we tried. Rotation demonstrates this most clearly as the sufficient statistic (orientation matching) and the exact algorithm (correspondence) are separated by a symmetry controlled via the stimulus. Scale is a blend of two statistics that are both nearly exact (Appendix A). Translation is exact at one stage of composition and serves as the control. None of the five losses tested removed the rotation computation durably. Each removed either a readout or the write feeding it, and each was reversible with tens to hundreds of labeled examples (1,536 in the case of NPO) on a computation that took 20,000 to build. Reacquisition cost was the only measure that separated the five; probes, activation distance from the base model, and the in-distribution retained battery did not.

Limits. The method needs a task whose exact computation is known and whose confounds can be constructed, and the statistics found here belong to the three similarity-transform tasks of one benchmark rather than to spatial reasoning at large. The models are two 7B families under one adapter recipe. The transplant and blocking results are on rotation only, and most runs are single draws: the three removal onsets, the two-axis refusal and its re-attaches, and the unmarked onset were replicated at one further seed and the marked onset at two. The retained battery is in-distribution; removals indistinguishable on it differ by up to 0.35 on external spatial benchmarks (Section 8), so selectivity is claimed on that battery only.

AI USE STATEMENT

AI assistants (Claude, Anthropic) were used under the authors' direction to review text, to write analysis, training, and plotting code, and to organize results. Experimental design, the choice of controls, and the interpretation of results are the authors'. Every number in the paper was checked by the authors against the underlying result files.

REFERENCES

Anonymous. Sophon: A procedural diagnostic for spatial reasoning in vision-language models, 2026a. Accepted at NeurIPS 2026. Benchmark and generator provided as anonymized supplementary material.

Anonymous. Gain and alignment at the readout layer of vision-language models, 2026b. under review.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

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

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023.

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 *Computer Vision – ECCV 2024*, volume 15081, pp. 148–166. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73337-6_9.

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, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, and Joshua Batson. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, https://transformer-circuits.pub/2026/workspace/index.html, July 2026.

Katherine L. Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020. ISBN 9781713829546.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873. URL https://doi.org/10.1107/S0567739476001873.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=RN5KLywTll.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan,

Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xlr6AUDuJz.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *CoRR*, abs/2402.16835, 2024. URL https://doi.org/10.48550/arXiv.2402.16835.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Xiangyu Qi, Boyi Wei, Nicholas Carlini, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, and Peter Henderson. On evaluating the durability of safeguards for open-weight LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=fXJCqdUSVG.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 18–34, December 2024.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-resistant safeguards for open-weight LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FIjRodbW6.

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.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

A SCALE

Scale as a blend of two statistics. Section 4 identified the rotation statistic with a symmetry sweep and with stimuli on which the statistic and the algorithm disagree. The same construction applies to scale. Scale has several cheap statistics that equal the true factor on the training distribution, the ratio of the two clouds' bounding extents, of their maximum radii, of their RMS radii, or of their mean radii about the centroid (Table 1). The test cell separates them from the truth. A core of points expands while three stragglers collapse toward the centroid, so the bounding extent and the maximum radius shrink while the true factor grows. A zero-parameter reader that answers from one statistic alone scores true scale 1.000, bounding extent 0.089, maximum radius 0.044, RMS radius 0.739, and mean radius 0.695 on this cell. The extent statistics follow the stragglers and answer wrong, and the averaged radii mostly follow the core. OneVision scores 0.441 at 10,000 examples per task (0.397 at 400) and Qwen2.5-VL scores 0.531 at 6,000, between the two families, which is why we describe the installed scale computation as a blend of the two. The battery decouples one statistic family from the truth, so it does not separate a model that computes scale from one that uses the RMS radius.

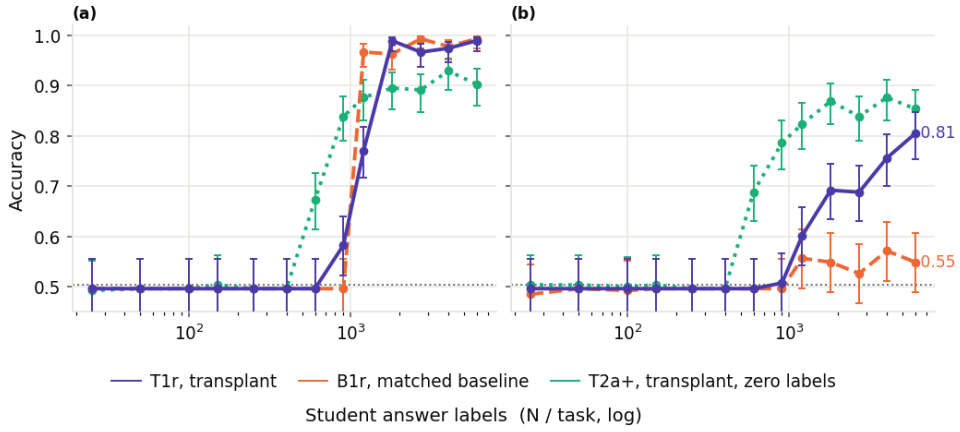

Figure 5: Accuracy on (a) the student's own rendering and (b) the other rendering, $n = 266$, Wilson 95%. T1r: a digit-rendered student trained with answer loss plus activation matching to the cue-free teacher on the paired dot scene. B1r: the same budget with answer loss only. T2a+: activation matching only, no answer labels. B2: matching to the teacher's activations on a different scene.

A quantity that is not computed. A cell at chance can mean a statistic at its zero crossing or no computation at all. Conditioning on transform magnitude separates the two, since a computed but wrong quantity changes with the magnitude while an absent one gives a constant answer. On novel color palettes, OneVision's scale answer is constant emission at every magnitude, and its per-bin accuracy equals the modal token's share of that bin's ground truth to within 0.03. Both translation axes remain computed on the same images. On those images the scale computation is absent rather than wrong.

B TRANSPLANT FINGERPRINT AND BLOCKING DOSE

This appendix holds the measurements behind the transplant of Section 6, which are the probe fingerprint of the transplanted student, the three controls, and the blocking sweep.

Probe fingerprint. Linear probes at the vision tokens and at the answer position show that the student holds both computations, selected by the input (Fig. 6). On dot renderings above 90°, the student's vision-token probe is 0.60 and its answer-position probe is 0.79, close to its behavioral accuracy of 0.77. The rise of 0.19 between the two sites matches the teacher's 0.17. On digit renderings the answer-position probe is 1.00, as in the cue-trained model. The baseline student's probes on dots are at chance at both sites. The answer-position structure the probes read in the reader is produced by frozen layers that teacher and student share, so it is present whenever the workspace below it is matched.

Controls. Three controls determine what was installed. With *scrambled targets* (B2), the teacher activations come from the next scene in the pool rather than the paired one. The marginal statistics are identical and the scene content is wrong. Every checkpoint on both renderings stays at 0.500 with constant emission. With the matching weight *annealed* to zero by the midpoint of the run, the cross-rendering score reaches 0.64 at 1,800 examples and then falls to 0.594 [0.53, 0.65]. That overlaps the no-matching baseline's 0.549 [0.49, 0.61], so the installed computation is retained only while the matching loss is applied. With *no answer labels*, matching the vision tokens alone (T2a) yields 0.65/0.62 with skewed emission. Matching the answer-position stream of the workspace as well (T2a+) yields 0.902/0.853 at 6,000 examples. Its accuracy leaves chance between 400 and 600 examples, against 900 to 1,200 for supervised training, and its own/cross gap is 0.03–0.05. The T2a+ probes show the cue-free profile on both renderings. That is consistent with its ceiling of 0.91 on digits, since it does not use the cues that the supervised model uses to reach 0.99. The budget is zero student labels and 6,000 teacher forward passes, on a teacher that cost 18,000 labeled dot examples per task.

Blocking. B2 included the same answer loss with which the baseline leaves chance between 900 and 1,200 examples, and it never left constant emission through 6,000 (Fig. 7). A matching target whose loss cannot decrease therefore prevents the supervised route from forming. A sweep over the matching weight shows this is tunable (Fig. 7). Weights of 0 and 0.05 allow the supervised route to form, and 0.25 and 1.0 do not.

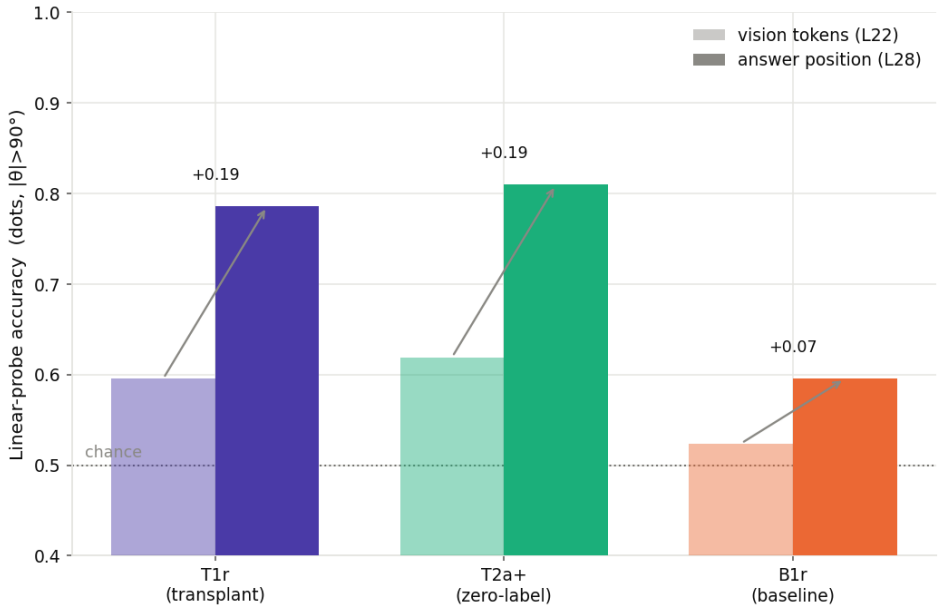

Figure 6: Linear-probe accuracy at the vision tokens and at the answer position on dot renderings above 90°, for the teacher, the transplanted student, and the supervised baseline.

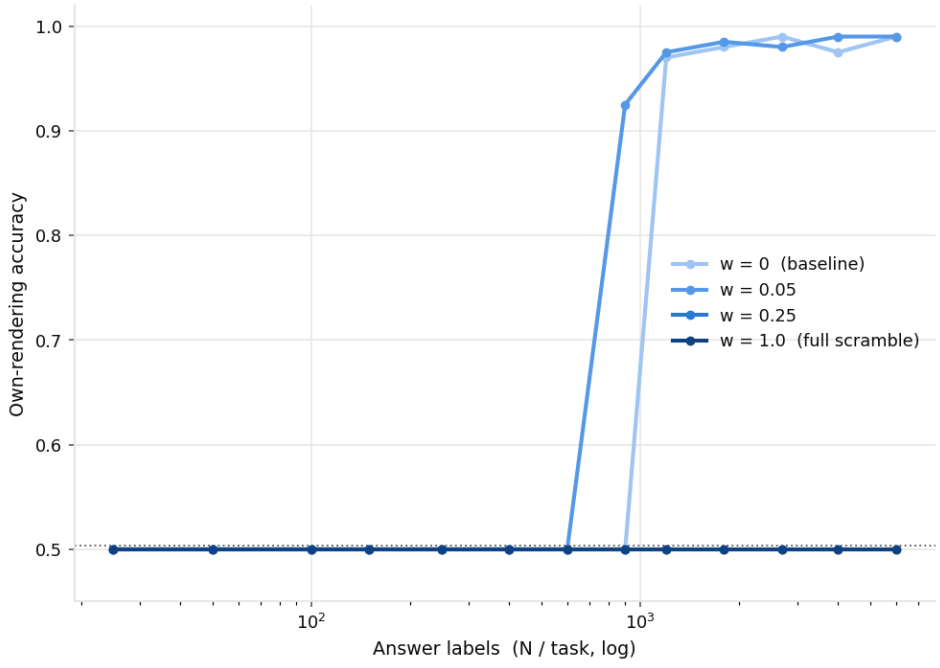

Figure 7: Accuracy against examples for four scrambled-target matching weights. Weights 0 and 0.05 leave chance between 900 and 1,200 examples; 0.25 and 1.0 remain at constant emission through 6,000.

C REFUSING THE SCALE DIRECTION WHILE KEEPING THE FACTOR

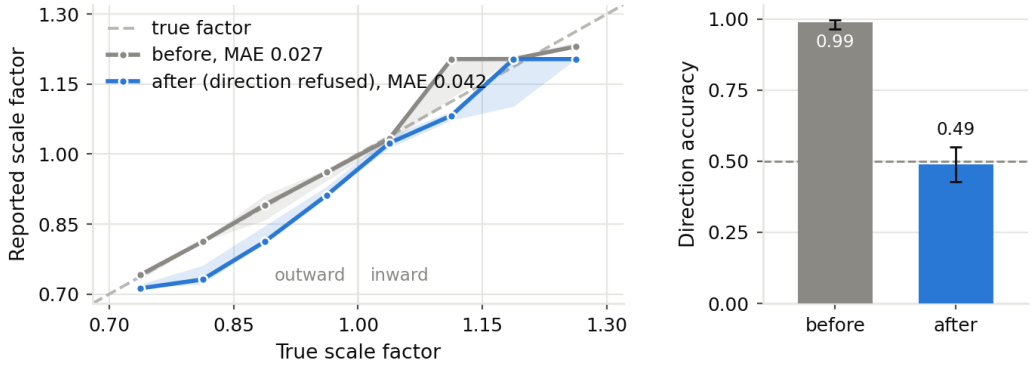

Figure 8: Refusing the scale direction while keeping the factor. Left: the reported scale factor against the true factor on 250 held-out scenes, binned by true factor (median and interquartile band), before and after the constant-token loss on the direction question; the dashed line is the true factor. Right: accuracy on the direction question before and after (Wilson 95% intervals; dashed line, chance).

Before the two-axis construction of Section 7, we tried the simpler pairing that scale offers, one categorical readout (the direction, outward or inward) and one numeric readout (the factor) of the same quantity. Starting from a Qwen2.5-VL checkpoint trained on rotation and scale and competent at both scale readouts, the direction question is trained toward a constant token for 1,600 steps while the factor question keeps its ordinary loss. On 250 held-out scenes the direction falls from 0.99 to 0.49 with mixed emission, while the reported factor stays on the true factor across 0.7–1.3, with mean absolute error 0.027 before and 0.042 after (Fig. 8). The two readouts are therefore separable. The variant does not show containment, because the retained factor's sign relative to 1 is the refused direction. A user who can ask for the factor has the answer. Section 7 uses the two-axis construction for that reason.

D REMOVAL BATTERY

The effects of each loss. Fig. 12 visualizes what each loss removes from the models through a series of probes. In the competent model, answers become decodable at both the answer token and vision token locations at layers 17–18 (answer position 0.93 at 18; vision tokens 0.88 at 18 and 0.96 from 20). The refused model probes similarly (0.89–0.91 at the answer position from layer 18 through the last layer, 0.83–0.86 at the vision tokens), and its activations projected onto the competent model's readout direction give the right sign at 0.88 at layer 20, but decay rapidly past layer 22. This is the signal of "refusal to vocalize" which this loss targets. After jamming or erosion no probe is above chance at any layer (0.45–0.65 against shuffled controls up to 0.60). At 800 steps the jammed model's vision-token activations on rotation scenes are within 0.17–0.21 relative distance of the base model's, as close as a checkpoint that never learned rotation, while scale and translation stay at 1.00 on their own scenes. The jamming loss teaches the adapter to run like the base model on rotation scenes, but that checkpoint re-attaches in 24 examples. The re-attachment weight change has Frobenius norm 3.0, against 24.0 for the jamming edit and 137.8 for the update that installed rotation. This demonstrates that the re-attach leaves the removal edit in place and routes around it, and after 24 examples the answer-position probe is back at 0.94 at layer 18 while the bulk distance of the vision stream from base is unchanged. The update that installed rotation remains untouched beneath both edits, and reacquisition requires only reconnection of a computation that is still in the weights.

Injection at the last trained layer. On the refused model we fit a logistic probe at hidden state 22, the last trained layer, and add $\alpha\, s\, \hat{\jmath}$ to that hidden state, where s is the probe score centered at its median and scaled to unit mean magnitude, and $\hat{\jmath}$ is the competent model's context-averaged Jacobian of the answer-logit difference (clockwise − counterclockwise). The residual norm there is about 231 (Table 4). On 100 held-out scenes, accuracy rises from 0.50 to 0.60, 0.81, and 0.88 at $\alpha = 64$,

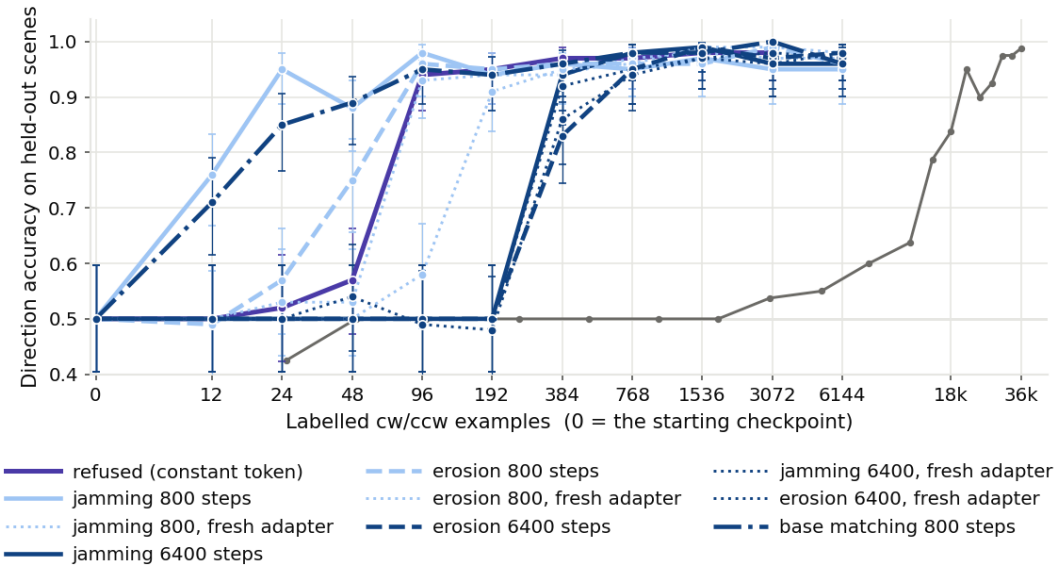

Figure 9: Re-attaching the direction after refusal and after each destruction loss, 100 held-out scenes per checkpoint, Wilson 95%. Line style is the loss (jamming solid, erosion dashed, base matching dash-dot), shade is the dose (light 800 steps, dark 6,400), thin dotted lines are the fresh-adapter regime in the same color as their continued arm; acquisition from the base model in gray. Every start has both rotation readouts dead and scale and translation at ≥ 0.98.

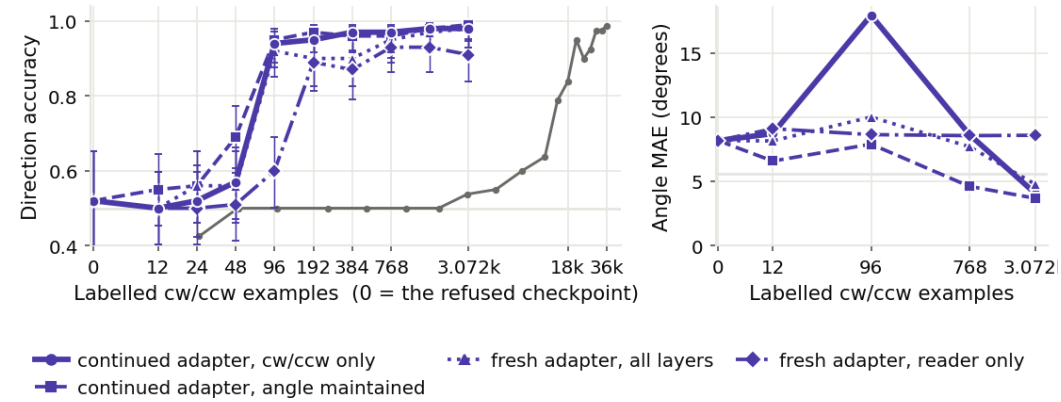

Figure 10: Re-attaching the direction from the refused checkpoint under each regime. Left: direction accuracy on 100 held-out scenes per checkpoint for the continued adapter with and without the angle maintained and for a fresh adapter on all layers or on the reader only, with acquisition from the base model in grey. Right: the angle readout (mean absolute error, 100 scenes) on a checkpoint subset; the starting model's angle is $8.2°$ and the 36k model's $5.6°$. Wilson 95% intervals.

128, and 256. A probe on the same activations reads 0.90. Shuffling the scores across scenes gives 0.51, 0.46, and 0.37; a random direction orthogonal to $\hat{\jmath}$ with the true scores stays at 0.50 at every α ($z = 7.4$ and 5.8 against the two controls at $\alpha = 256$). The model's reader transports a write made at the last workspace layer (layer 22) in a model whose own output is at chance. The refusal loss works by withholding that write, not by making the reader unresponsive to it.

Reconnected, not relearned. The re-attached models are scored over the full circle on 300 held-out scenes in six $30°$ bands. None holds the fold at $90°$ or the below-chance band that the orientation statistic of Section 4 produces. The angle prompt, untrained in these runs, answers numerically on 300 of 300 scenes in every one, with mean absolute error $17.7°$, $20.4°$, and $34.1°$ from the refused, jamming, and erosion starts against the 36k model's $8.0°$ on the same scenes. Beyond $90°$ the

reported angle tracks the true angle rather than the orientation statistic fold. Training the angle prompt alone from the 6,400-step erosion checkpoint, with no direction labels, brings the angle from non-numeric to $7.9°$ at 192 examples and $4.0°$ at 6,144, and the untrained direction from 0.51 to 0.74 at 96 and 0.95 at 192; from jamming at 6,400 the direction is 0.89 at 96 and 0.95 at 192. The competent model under the same angle-only fine-tune reaches $3.0°$ at 384 and $3.2°$ at 6,144, so the removed models' final angles of $2.0°$–$4.0°$ are the ceiling of this regime, not a residual cost of the removal. Either readout restores the other.

Replication. The three claimed onsets were redrawn at a second seed with the 36k substrate held fixed. Refusal re-attaches at 96 in both draws (0.94 and 0.94), jamming at 24 (0.95 and 0.95), and erosion at 96 (0.96 and 0.95), with retained tasks at 0.99–1.00 throughout (Table 3). Across every start, dose, and regime, the labeled supervision that reopens the rotation direction is between about $1/800$ and $1/13$ of the direction supervision in the six-task installation, and between about $1/750$ and $1/12$ of the 18,000 direction-only examples that do not install it from a checkpoint that was never competent.

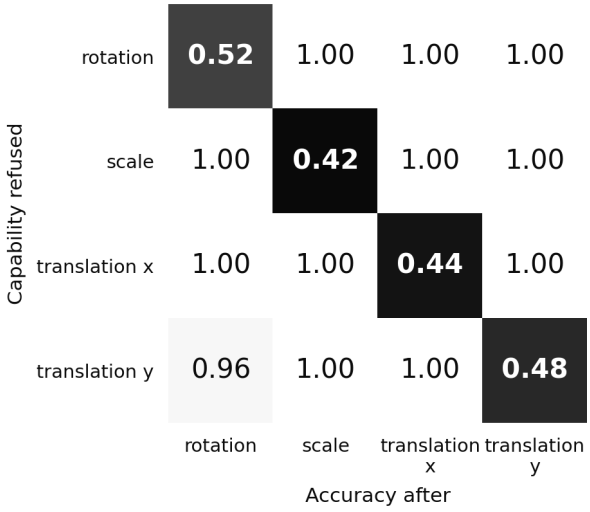

Figure 11: Refusing one direction from the 36k checkpoint with the constant-token (refusal) loss (rows) and the held-out accuracy of each task afterwards (columns), $n = 50$ per cell. The target falls to 0.42–0.52 with mixed emission while other tasks stay at 0.96–1.00. Selectivity is measured on this in-distribution battery only (Section 9).

External benchmarks. Against the base model's 0.936 on What'sUp CLEVR and 0.888 on BLINK spatial relations, the 800-step refusal scores 0.586 and 0.783, jamming 0.890 and 0.818, base matching 0.890 and 0.804, and erosion 0.890 and 0.629; at 6,400 steps jamming is 0.875 and 0.804 and erosion 0.728 and 0.510, and the competent model itself is 0.855 and 0.678. The refusal, which leaves every in-distribution retained task at 1.00, costs the most on CLEVR, and re-attaching the direction with 96 examples returns it only to 0.743; jamming's re-attach at 24 examples is at 0.875 and 0.818; NPO scores 0.777 and 0.804, and its re-attach at 1,536 examples 0.860 and, parsing 68 of 143 BLINK items, 0.343 over all items. Erosion's cost on BLINK is largely a format collapse that deepens with re-attachment: the checkpoint parses 127 of 143 items and scores 0.709 on those, and its re-attaches at 96 and 768 examples score 0.873 and 0.877 on CLEVR but parse 95 and 28 BLINK items (0.716 and 0.679 on those; 0.476 and 0.133 over all items).

E Per-stimulus and decoupling tests of the reader

This appendix tests the orientation reader of Section 4 more closely, first stimulus by stimulus on the symmetry sweep and then on a battery built to separate the principal axis from the true rotation. The moved cloud is an exact rotation of the reference, so the measured axis difference equals the

start	12	24	48	96	192	384	768	6144	first ≥ 0.90	angle at 12/96/768/6144
refused	.50	.52	.57	.94	.95	.97	.97	.98	96	9 / 18 / 9 / –
jamming 800	.76	.95	.88	.98	.94	.96	.95	.95	24	33 / 8 / 7 / 25
jamming 800, fresh adapter	.50	.53	.53	.93	.94	.94	.97	.97	96	– / 19 / 8 / 10
erosion 800	.49	.57	.75	.96	.95	.96	.96	.96	96	62 / 41 / 12 / 10
erosion 800, fresh adapter	.50	.50	.50	.58	.91	.95	.96	.98	192	66 / 64 / 20 / 11
base matching 800	.71	.85	.89	.95	.94	.96	.98	.96	96	37 / 13 / 9 / 9
jamming 6400	.50	.50	.50	.50	.50	.94	.98	.96	384	– / 63 / 9 / 15
jamming 6400, fresh adapter	.50	.50	.50	.50	.50	.92	.95	.97	384	– / – / 9 / 9
erosion 6400	.50	.50	.50	.50	.50	.83	.95	.98	768	79 / 79 / 34 / 8
base matching 6400	.50	.54	.55	.93	.94	.95	.95	(.98 at 1536)	96	not measured
refused 6400	.52	.65	.98	.87	.93	.95	.95	(.99 at 1536)	48	not measured
erosion 6400, fresh adapter	.50	.50	.54	.49	.48	.86	.94	.98	768	79 / 79 / 56 / 13
NPO 800	.01	.14	.48	.44	.48	.53	.80	–	1536	– / – / – / –
NPO 6400	.01	.33	.51	.56	.82	.91	.96	(.99 at 1536)	384	not measured

Table 2: Direction accuracy against labeled direction examples from each starting checkpoint ($n = 100$ per checkpoint, Wilson half-width about 0.05 at 0.95), the first checkpoint at or above 0.90, and the angle readout's mean absolute error in degrees at four checkpoints ("–" = non-numeric on 100 of 100 scenes). "Fresh adapter" merges the removal adapter into the weights and trains a new LoRA on all 28 layers; other rows continue the adapter that made the removal. Acquisition from the base model on the 80-scene inline evaluation: at chance through 1,800 examples per task, 0.84 at 18,000, 0.99 at 36,000 (0.97 on the 300-scene full-circle set). Scale and both translations are at or above 0.98 at every start and every checkpoint.

chain	seed	12	24	48	96	192	384
refusal, then re-attach	99	.50	.52	.57	.94	.95	.97
refusal, then re-attach	7	.52	.58	.59	.94	.95	.96
jamming 800, then re-attach	99	.76	.95	.88	.98	.94	.96
jamming 800, then re-attach	7	.70	.95	.95	.98	.99	.97
erosion 800, then re-attach	99	.49	.57	.75	.96	.95	.96
erosion 800, then re-attach	7	.50	.61	.74	.95	.95	.97

Table 3: Seed replicates of the three onsets. The 36k substrate is shared; the removal and the re-attach were redrawn. Inline evaluation used 50 rotation scenes per checkpoint at seed 99 and 100 at seed 7, both balanced.

applied rotation. The median absolute residual is $0.4°$ on elongated clouds and $2.2°$ on round ones. The reader's curve is therefore set by θ and the period, and the correlation tests the period and the phase of the response. The evidence that the statistic is an orientation is that the period follows the cloud's symmetry. This battery does not test the statistic stimulus by stimulus. Within a band θ varies by $\pm 2.5°$, which is narrower than the models' transitions. At the crossing bands the reader and the model agree on 0.58 (OneVision) and 0.55 (Qwen2.5-VL) of 192 round-cloud stimuli, against 0.50 if the two were independent. At the $90°$ band on elongated clouds the reader and OneVision both score 0.438 on 64 stimuli and agree on 0.53 of them.

A battery on which the two come apart allows us to pinpoint the installed statistic using elongated clouds (aspect 3, 30–60 points, the same renderer). The red cloud is rescaled about its centroid by 3 along the short axis and $1/3$ along the long one, before an in-range rotation of 15–$40°$. Its principal axis therefore turns by $90°$ while the exact rotation is unchanged, since the polar factor of a rotation composed with an axis-aligned scaling is the rotation, and Kabsch recovers θ on 126 of 128 scenes. A principal-axis reader answers the opposite direction on every scene. On 128 such scenes OneVision answers with the axis on 0.977 and Qwen2.5-VL on 0.781, against 0.992 for both on the unrescaled control, with both tokens emitted. A rescale of 1.2 that keeps the axis where it is costs OneVision more (0.555) than Qwen2.5-VL (0.844), so OneVision's statistic also depends on the two clouds' extents matching.

F WHERE THE CUE-FREE MODEL LOOKS

The cue-free model's correspondence landmark is the point farthest from the centroid (Section 5). Displacing only that point in the opposite direction flips the answer on 0.21–0.54 of scenes across the two bands and the two checkpoints. A random point displaced flips 0.05–0.16. The outermost point

injection at hidden state 22	$\alpha = 0$	32	64	128	256
competent $\hat{\jmath}$, refused model's own probe score	.50	.52	.60	.81	.88
same direction, scores shuffled across scenes	.50	.50	.51	.46	.37
random orthogonal direction, true scores	.50	.50	.50	.50	.50

Table 4: Direction accuracy of the refused model on 100 held-out scenes after adding $\alpha\, s\, \hat{\jmath}$ at hidden state 22. Scores are median-centered and scaled to unit mean magnitude, so α is the injected norm; the residual norm at that position is about 231. A probe on the same activations reads 0.90.

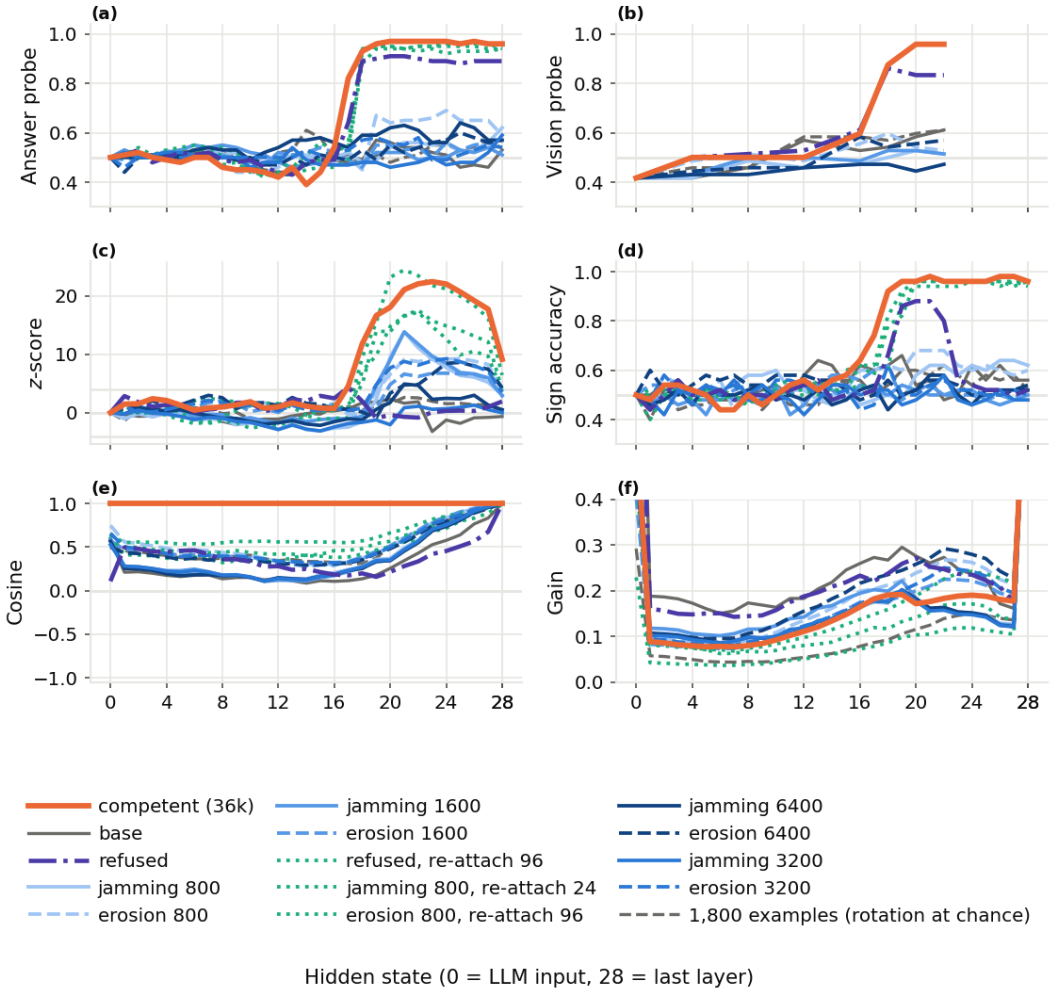

Figure 12: Layer-by-layer measurements on the direction prompt for the competent, base, refused, removed, and re-attached checkpoints, and for the 1,800-example checkpoint of acquisition, which has scale and translation but not rotation; 200 fit and 100 held-out scenes per prompt. (a) Answer-position probe. (b) Probe on the 3×3-pooled vision tokens. (c) z-score of the cosine between the model's own context-averaged Jacobian lens $\hat{\jmath}_L$ and the class-mean direction at layer L. (d) Sign accuracy of each model's activations projected onto the competent model's $\hat{\jmath}_L$. (e) Cosine between each model's $\hat{\jmath}_L$ and the competent model's. (f) $\|\hat{\jmath}_L\|$.

can overrule the other nine but does not decide in isolation. On ordinary rotations the model scores 0.99 when the outermost radius exceeds the next by 30% or more and 0.89 when the three largest radii lie within 3%, and 0.93 on clouds with two points ("twins") tied for outermost, where the model errs on scenes in which pairing the wrong twin implies the wrong sign (0.13 against 0.01, $z = 2.9$), the near-duplicate extremal-point stimulus of Table 1. We find that the model strongly attends the vision patches of the outermost point. At decoder block 8 the answer position's eight most attended

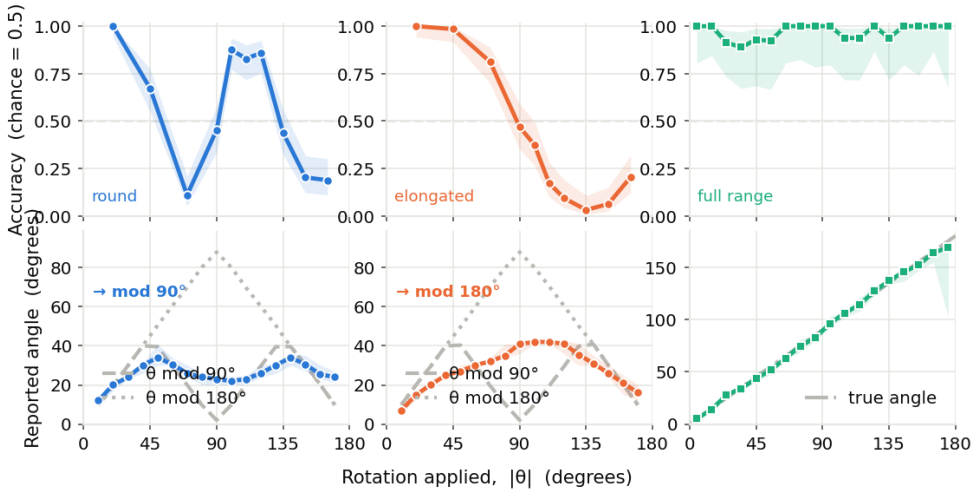

Figure 13: The same statistic feeds the numeric readout. Asked for the angle as a number, Qwen2.5-VL's report correlates with $\theta \bmod 90°$ on round clouds ($r = +0.77$) and with $\theta \bmod 180°$ on elongated clouds ($+0.90$), and with the true angle at $+0.32$–0.35 in both. Third column, green: the six-task model trained over the full range, whose direction accuracy has no period (0.97 over 300 scenes) and whose reported angle follows the true angle ($r = +0.999$ over 18 bands).

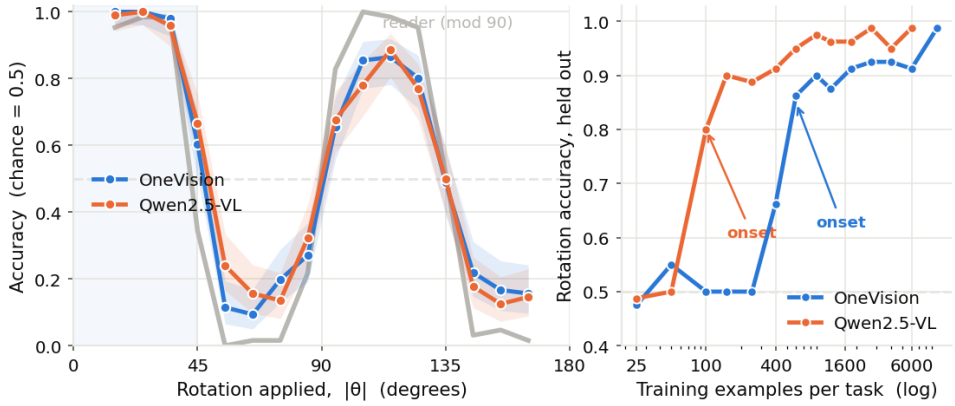

Figure 14: Two architectures trained on the same pool in the same order. The full-circle curves of OneVision and Qwen2.5-VL correlate $+0.989$ and the same reader predicts both. Qwen2.5-VL reaches the pattern at 4–6$\times$ less supervision and scores 0.988 in range.

vision patches (Fig. 15) contain the outermost point's patch on 0.44 of the clear-outlier scenes and no other point's patch on any, and its red partner's patch on 0.35 of scenes by block 14. In cases with twin outer points, the most attended patches contain one twin's patch on 0.36 and both on 0.01. That attention marks not merely a correlate but a conduit, which we test by patching. For the 160 scenes with clear outliers, we replace the residual stream at the 3×3 neighborhood of the outermost point's patch (and of its red partner's) with the activations of the same cloud rotated the other way, one hidden state at a time. Patched at the language model's input, the answer flips on 0.35 of scenes, while patched at hidden states 4, 8, 12, and 16, it flips on 0.30, 0.23, 0.14, and 0.05. The rate falls with depth as the direction is read out of those positions into the rest of the stream, and hidden state 16, where it reaches 0.05, is the depth at which the answer position stops attending to those tokens. The same patch at a random other point's neighborhood flips 0.01–0.03 at every depth, and the single patch token alone 0.02, so the conduit is that point's neighborhood. Replacing the neighborhood with an unrelated scene's activations drops accuracy from 0.99 to 0.87 at the input and only to 0.97 by hidden state 16, against 0.98–0.99 for a random point. Where it goes is the question: the same counterfactual activations written at the three question tokens *red*, *align*, and *blue* flip the answer on

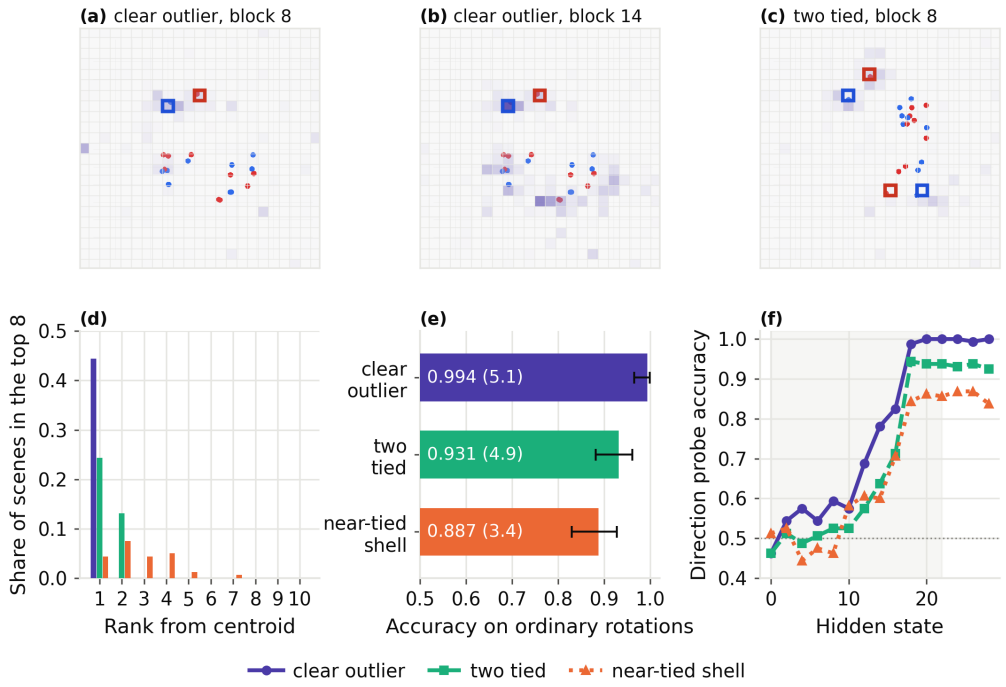

Figure 15: Where the cue-free model looks. (a, b) The answer position's head-mean attention over the vision patches at decoder blocks 8 and 14 on a scene with one clearly outermost point, and (c) at block 8 on a scene with two points tied for outermost; the outermost blue point's patch and its red partner's patch are outlined. (d) The share of scenes on which each point's patch, ranked by distance from the centroid (1 = outermost), is among the top eight attended at block 8, for the three radius-profile classes. (e) Accuracy on ordinary rotations by class with Wilson 95% intervals; the mean absolute logit margin is in parentheses. (f) The answer-position direction probe by hidden state; the shaded span is the trained band, layers 0–22. $n = 160$ scenes per class.

0.39 of scenes at hidden state 16, and written at the whole question span on 0.49, while at hidden states 4, 8, and 12 no text position moves it and the answer-option tokens, the closing template, the vision-end token, and three random question tokens never do. The direction is gathered from the patch neighborhood into the question's content words between hidden states 12 and 16, and read from there.

G Readout control details

This appendix holds the measurements behind the readout control of Section 7, which are the probe, the prompt-side attacks, and the weight-side re-attach regimes.

Probe. We fit a logistic probe on the answer-position residual at each layer, using 95 of the 136 held-out scenes, and score it on the other 41. On the refused checkpoint the probe reads the sign of x and the sign of y at 1.00 from layer 9 through the last layer. The pre-suppression checkpoint has the same profile. Shuffled-label controls score between 0.39 and 0.61. The untrained base model reads the two signs at 0.85 at layer 9 and at 1.00 from layer 15. XOR is decodable from the refused checkpoint at 1.00 from layer 12, as from the pre-suppression checkpoint, while the base model reaches at most 0.83. Both signs remain at the answer position with the depth profile the fine-tune gave them. The refusal acts on what is emitted.

Prompt-side attacks. The threat model for a closed-weight deployment is an attacker with the prompt only. We evaluated fifteen phrasings of the refused questions on the refused checkpoint and on the pre-suppression reference model (Fig. 16). The phrasings include a paraphrase, an instruction

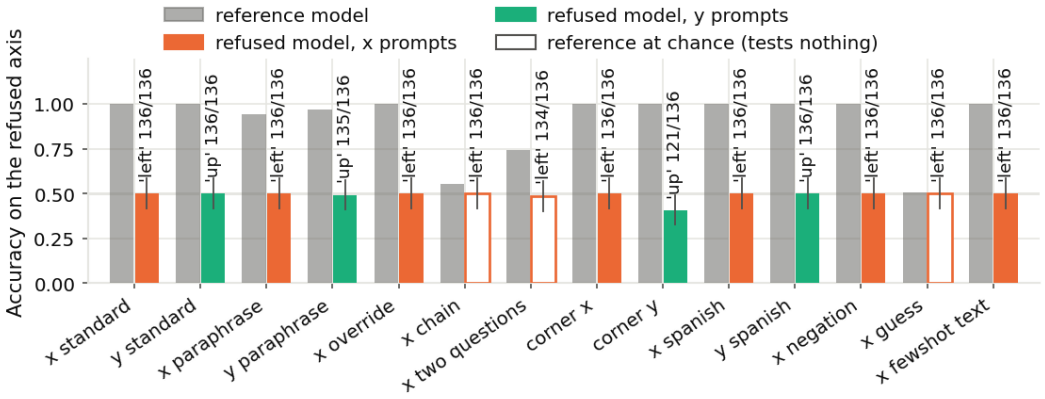

Figure 16: Fifteen prompt-side attacks on the refused checkpoint, 136 held-out scenes. Gray: what the pre-suppression reference model's answer reveals on that prompt; color: the refused model. Outlined bars are prompts on which the reference answers at chance. Above each colored bar, the refused model's most frequent answer and its count. XOR on the refused checkpoint is 0.99.

override, a chain that asks for XOR, then y, then x, a corner question, Spanish, a negation that asks for the wrong direction, a request for a best guess, and a text few-shot. The reference answers twelve of them at 0.94–1.00. On the negation it states the true direction on all 136 scenes, which reveals the answer as well as a correct one would. On all twelve the refused model emits its constant token on 136 of 136 scenes. The one exception is the corner question for y, where 15 of 136 answers leave the constant token and are anti-correlated with the truth. The three prompts on which the reference itself is at chance test nothing, and are reported but excluded.

A further attack places $k = 2$, 4, or 8 solved training scenes before the test scene. Each demonstration has its image, the refused question, and the correct answer as an assistant turn. This gives twelve cells, from two axes, three values of k, and true or shuffled demonstration labels. The refused model emits its constant token on 136 of 136 scenes in eleven of the twelve cells and on 135 in the other. The reference model answers the same contexts at 0.55–0.71 with both tokens, and shuffling the labels does not change it. Multi-image contexts are therefore a weak probe for these single-image fine-tunes. The demonstrations do not move the refused model off its token, and they recover less from the reference than the single-turn prompts do.

A soft prompt is the strongest input-side attacker. Sixteen vectors are inserted after the image and trained with every weight frozen on 12 to 768 labeled training scenes, then scored on the 136 held-out scenes. On the refused checkpoint the x readout is at 0.50 through 12 examples, 0.74 at 24, and 1.00 from 96 on, with balanced emission. The y readout is 0.99 at 96 and 1.00 from 192. The pre-refusal reference stays at 1.00 throughout. The same prefix trained on shuffled labels reaches at most 0.71, with skewed emission. The refusal therefore holds against every discrete prompt we tried. It falls, with no weight change, to an attacker who can optimize the input with about 100 labels, the same order of supervision as the weight-side re-attach.

Weight-side reacquisition and separability. With trainable weights, labeled examples of one axis, and nothing else trained, the refused readout returns in 24–48 examples (Fig. 17, left column). That is the same order as the ten it took to install the refusal. A second draw of the refusal and the re-attach (seed 7) behaves the same way. The refusal forms by 100 steps, and each axis returns in 24–48 examples. In that draw x reaches 0.97 after 24 examples and 1.00 after 48, y reaches 0.82 and 1.00 at the same points, and the untouched axis stays at 0.40–0.50 through 96. In the first draw the untouched axis returns to 0.82 by 96 examples and stays there through 768. The unmaintained XOR falls to 0.81 in the x run and to 0.89 by 768 in the y run.

With XOR maintained (middle column), XOR stays at 1.00. As x is re-attached, the untouched y prompt begins to emit x's answer words, `right` on 12–31 of 106 scenes, which is the mechanism of the drift. With the constant-token loss maintained on the other axis (right column), x returns to 0.89 at 12 examples and 1.00 at 24. The held y remains a single token on 66 of 66 scenes at every

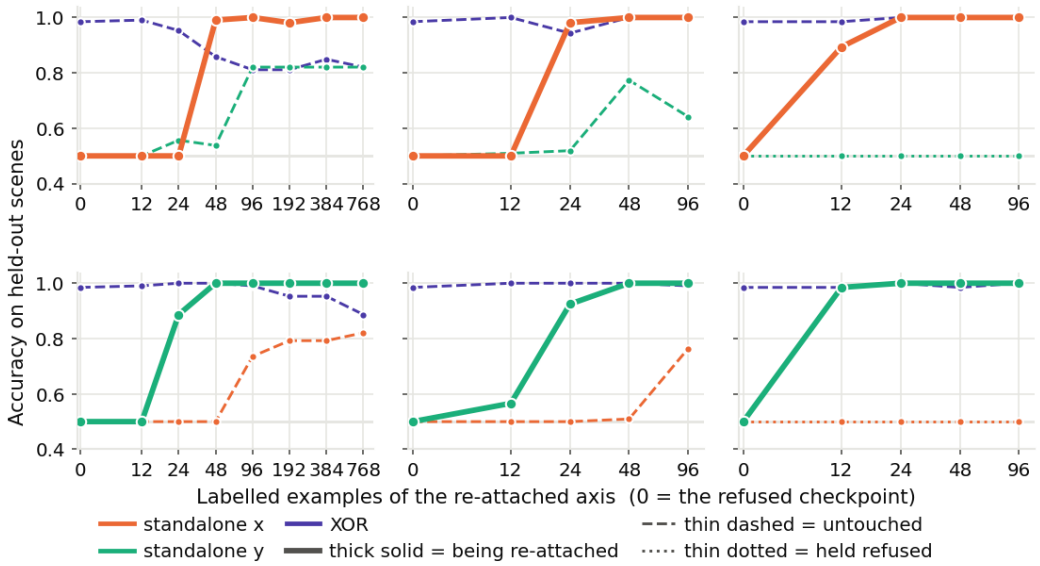

Figure 17: Re-attaching one refused readout from the refused checkpoint under three regimes. Rows: which axis receives ordinary supervision. Columns: nothing else trained; XOR maintained; the other axis held refused with the constant-token loss and XOR maintained. Colors: orange x, green y, violet XOR; thick solid = the re-attached axis, dashed = untouched, dotted = held.

model (Qwen2.5-VL)	own rendering	other rendering	change
cue-trained, 6,000 examples	0.992 (digits)	0.541 (dots)	−0.451
cue-free, 18,000 examples	0.940 (dots)	0.801 (digits)	−0.139

Table 5: Cross-rendering transfer on the same 266 scenes.

checkpoint through 96, and XOR remains at 0.98–1.00. The mirror run behaves the same way. The two refusals therefore decay together when neither is maintained and are separable when one is held.

H Additional cue results

What the cue-trained model pools. In a battery where m of the ten points rotate one way and $10 - m$ the other, scored against the majority, a single-landmark reader produces a ramp in m and a full ten-point fit a step. The cue-trained model's curve matches neither. Fitting the family "fit over a random subset of j correspondences" gives $j \approx 3$–4 (mean absolute error 0.029). A dose curve in the number of marked points saturates smoothly. Two cues close 67% of the gap between unmarked and marked accuracy, and six close 98%. A curriculum that fades the markers from ten to zero over 5,000 examples stays at 0.53–0.63, so the unmarked computation is not obtained by weaning a cue-trained model off the cues.

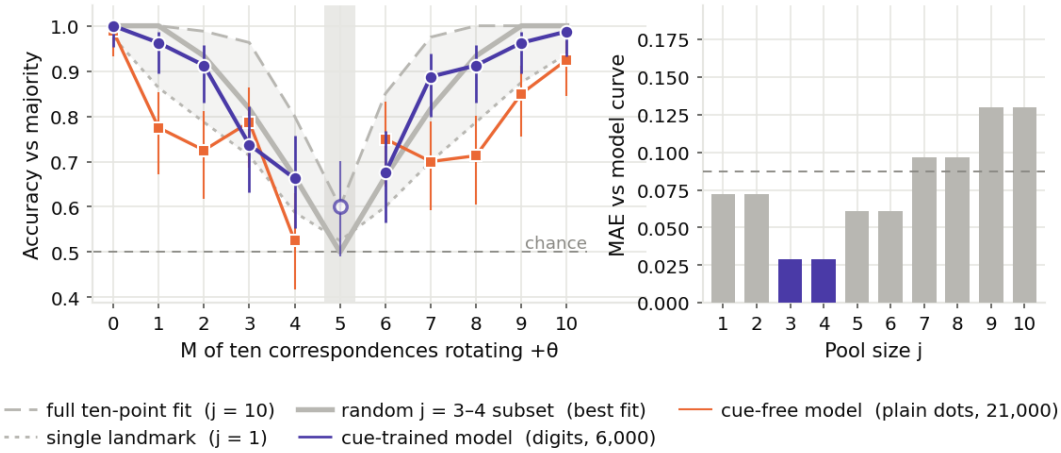

Figure 18: The disagreement battery. Left: accuracy against the majority as m of the ten points rotate $+\theta$ and the rest $-\theta$, for the cue-trained model (circles, digits, 6,000 examples) and the cue-free model (squares, plain dots, 21,000 examples), against the single-landmark and full ten-point readers and the best random-subset fit. Right: error of the cue-trained model's curve against a reader that fits j correspondences; the best pool size is 3–4. The cue-free model follows the single-landmark ramp.

I STIMULI

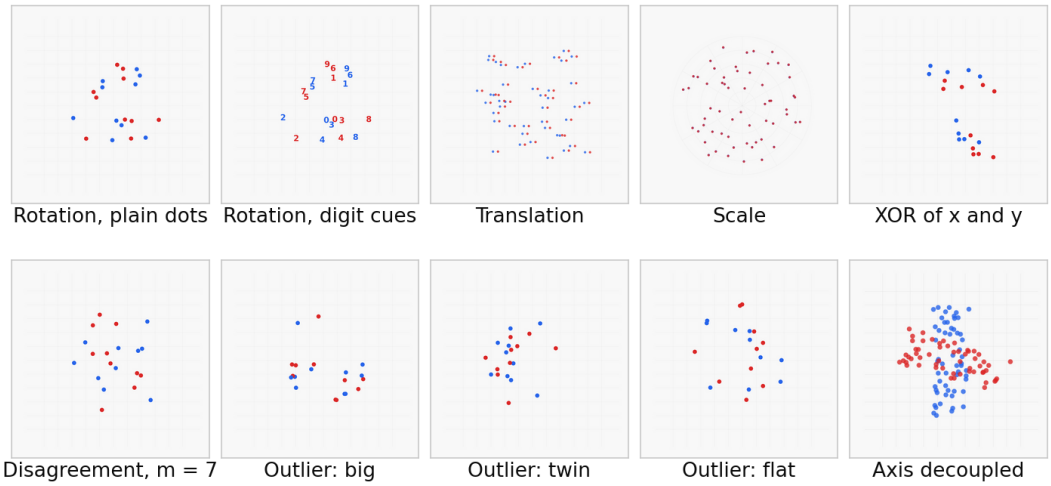

Figure 19: One scene from each battery, rendered as the model saw it: blue reference cloud, red transformed cloud. Top row: the plain-dot and digit-cued rotation scenes of Section 5, and the translation, scale, and XOR scenes of Sections 3 and 7. Bottom row: a disagreement-battery scene with seven of the ten points moved (Appendix H); the three outlier classes of Appendix F, one point far outside the cloud, two points at the same largest radius, and no point outside the cloud; and an axis-decoupled scene (Appendix E), in which the red cloud's principal axis turns by 90° while the true rotation stays inside the trained range. Round and elongated clouds are inset in Figure 1.

J PROMPTS AND HYPERPARAMETERS

Every question is a single user turn containing the image and the text below. The categorical questions end with the two answer words, whose order alternates across the pool so that neither position is favoured.

task	question
translation, x	Which direction should the red points move to align with the blue points? Answer with one of: right, left.
translation, y	Which direction should the red points move to align with the blue points? Answer with one of: down, up.
scale	Which direction should the red points move to align with the blue points? Answer with one of: inward, outward.
rotation direction	Which direction should the red points move over the shortest arc to align with the blue points? Answer with one of: clockwise, counterclockwise.
rotation angle	Estimate the rotation angle in degrees between the blue and red point sets. State your estimate in degrees as a number.
scale factor	By what factor are the red points radially scaled relative to the blue points? State your estimate as a numeric factor.
XOR	Consider the shortest move that aligns the red points to the blue points. Are its horizontal and vertical senses aligned (toward the same corner) or crossed? Answer with one of: aligned, crossed.

Table 6: The prompt for each task.

setting	value
base models	Qwen2.5-VL-7B-Instruct; LLaVA-OneVision-7B
precision, attention	bfloat16; scaled-dot-product attention
LoRA	$r = 32$, $\alpha = 64$, dropout 0.05, on the query, key, value, output, gate, up, and down projections of decoder layers 0–22; vision tower excluded
optimizer	AdamW, learning rate 5×10^{-5}, gradient accumulation 4, gradient-norm clip 1.0
image size	384 px for the supervision series and the wrap and anisotropy batteries; 768 px for the digit and plain-dot pools, the XOR pool, and the removal experiments
checkpoints	25, 50, 100, 150, 250, 400, 600, 900, 1,200, 1,800, 2,700, 4,000, 6,000, and 10,000 examples per task
seeds	99 for the supervision series and the removal runs; 301 for the digit and plain-dot pools; 7 for the replicates
transplant (Section 6)	activation matching on layers 0–22 with a per-channel standardized squared error, weight 1.0 against the answer loss
removal (Section 8)	removal loss weight 5.0 on layers 0–22; NPO with $\beta = 0.1$ and the pre-removal adapter as the reference; checkpoints at 200, 400, 800, 1,600, 3,200, and 6,400 steps

Table 7: Training settings shared by every run unless a section says otherwise.