Same Reward, Different Ability: A Single-Variable Intervention Shows Task Formulation Governs What Spatial RLVR Learns

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning with verifiable rewards (RLVR) is the standard recipe for improving spatial reasoning in vision–language models, and a growing diagnostic literature shows that such models often answer without looking. We ask the next question: what happens if the shortcut is removed before training? Nearly all point-level spatial benchmarks name the queried location as text coordinates, which the ground-plane relation makes informative about depth — a ridge regression that never sees the image beats a 7B vision–language model (VLM) on absolute depth. Holding the model, algorithm, reward, data volume, step count and the geometric distribution of training samples fixed, we vary only how the question refers to a pixel: drawn markers instead of coordinates. Across three paired seeds per arm, the marker formulation more than doubles metric accuracy and lifts every depth bucket off zero, while on these backbones the coordinate formulation leaves the hardest bucket at exactly zero under both RLVR and supervised fine-tuning. A decomposition arm shows the effect is the sum of two sub-interventions rather than one. The learned ability survives an adversarial set of off-ground points where the shortcut is void by construction, but does not transfer indoors, and ordinal accuracy stays below chance on a matched adversarial set — training removes the metric shortcut, not the ordinal position prior. Repeating the audit on two further backbones, including one from a different vendor, reproduces the diagnosis. Repeating the intervention shows that backbone capacity decides whether the formulation governs existence or merely degree — and that on the second vendor's model the effect does not appear at all, because that backbone learns no measurable metric ability here to begin with.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become the default way to improve spatial reasoning in vision–language models: a verifier checks a numeric answer against geometric ground truth, the reward is unambiguous, and a steady stream of methods report substantial gains on point-level depth, distance and ordering. The implicit consensus is that what remains is an optimisation problem — better advantage estimation, denser rewards, longer rollouts.

We examine a premise underneath that consensus. Almost every point-level spatial benchmark names the queried location by writing its pixel coordinates into the prompt — "what is the distance to the point at $[527, 437]$". For a point on the ground plane the depth follows Eq. 2, so the pixel row already carries most of the answer, and the prompt therefore contains a shortcut that requires no perception. Appendix Fig. 10 draws the geometry. The shortcut is not small: a ridge regression trained on prompt text alone — never shown the image — reaches 0.2951 on absolute depth and 0.8550 on depth ordering in driving scenes, roughly $5\times$ what a 7B VLM scores on the same items (Figure 1). Rejection sampling to decorrelate moves $\mathrm{corr}(v, \log z)$ only from -0.78 to -0.62, because the relation is geometric rather than incidental. Any reward computed on such a benchmark is collectible without looking at the image, and a policy optimiser will find that out.

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.

Diagnosis is not the gap. A growing body of work already establishes that spatial VLMs lean on non-visual routes: replacing the image with a gray blank exposes errors that probing and steering miss (Liao et al., 2026); cross-view consistency can be high while answers are wrong (Bhat & Yamasaki, 2026); vertical image position is entangled with distance (Min et al., 2026); pass@k at large k can measure trials rather than reasoning (Dragoi et al., 2025). We reproduce several of these and say so where we do. What none of these four does is intervene — they diagnose trained models rather than changing the task formulation and retraining to see what the optimiser then learns. (Two other lines of work do intervene, on the reward rather than the formulation; we place ourselves against them in Sec. 2.) The distinction is not one of degree. A diagnosis tells you what a trained model currently relies on; it cannot tell you what the same recipe would learn if the shortcut were absent, because that is a counterfactual about the optimisation rather than a property of the checkpoint. Yet the counterfactual is what a benchmark designer or a training pipeline actually needs to decide. That intervention is this paper.

What we do. We ask how much of the reported progress survives when the shortcut is removed, and what exactly about the formulation matters. We hold the model, algorithm, reward function, data volume, step count and the geometric distribution of training samples ($\mathrm{corr}(v, \log z) = -0.600$ vs. -0.590) fixed, and vary only how the question refers to a pixel: coordinates in the prompt, or a marker drawn on the image. Metric accuracy on a marker-based benchmark (Figure 1) moves from 0.1336 with a worst-bucket accuracy of exactly zero — the signature of a constant predictor — to 0.3050 ± 0.0484 over three seeds, worst bucket 0.1546, every bucket non-zero in every seed. Figure 2 shows what that looks like on individual queries, next to the per-bucket breakdown.

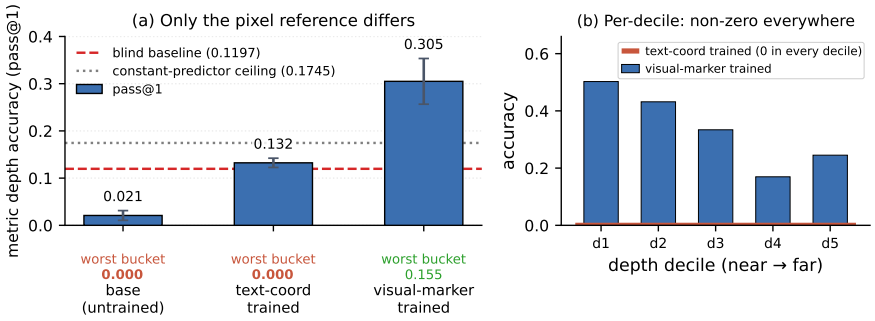


Figure 1: Same reward, same algorithm, same data volume — only the pixel reference differs. (a) Metric-depth pass@1 on a marker-based benchmark. Dashed line: the blind baseline (a text-only regressor that never sees the image); dotted line: the best-constant ceiling. Both are references for pass@1 only; worst-bucket accuracy is printed below each bar because its criterion differs (> 0, since a constant predictor scores exactly zero there). (b) Per-decile accuracy: the text-coordinate arm is zero in every decile despite a competitive aggregate score — the signature of distribution matching rather than geometry. Bars in (a) are three-seed means with std error bars; (b) shows one representative seed's per-decile profile.

What we mean by "governs", and what the ability is. We do not claim a prompt-format effect. The two arms differ in what information the question carries about the answer: coordinates expose a row from which depth is partly predictable, a drawn marker does not. Nor do we claim the formulation creates the ability outright — Sec. 4.4 shows that at lower capacity neither formulation suffices and at higher capacity both partly succeed, so what the formulation governs is how much of the reward pressure lands on looking at the image. As for the ability itself: ordinal accuracy sits far above a blind baseline in every model we tested, yet metric accuracy sits below it, and the arm we train remains below chance on a matched ordinal adversarial set. Training removes the metric shortcut; it does not touch the ordinal position prior, and the metric competence it produces does not transfer to indoor scenes.

Contributions. We state these against the diagnostic literature discussed above, which establishes that spatial VLMs often answer without looking; our contribution is not that finding but what follows from intervening on it.

- A single-variable intervention on the pixel reference with three paired seeds per arm — to our knowledge the first to manipulate the task formulation as the single variable, rather than measuring its correlates or intervening on the reward. Appendix Fig. 9 states the manipulated variable and the design around it.
- A decomposition showing the aggregate effect is the sum of two sub-interventions, so that "add markers" is not a single knob.
- Evidence that the direction of the effect is stable across two backbone generations of one family, while its form — whether the formulation governs existence or degree — depends on capacity; on a second vendor's backbone the same intervention does not reproduce, and we report that as a boundary rather than smoothing it over.
- A two-part reporting criterion — margin over a text-only blind baseline, and a non-zero worst bucket — that separates learning the relation from learning the answer distribution, with the task suite and evaluation code.

What we do not claim. That spatial VLMs answer without looking (Liao et al., 2026); that probes overstate grounding (Liao et al., 2026); that outputs collapse onto priors (Bhat & Yamasaki, 2026); that image position is entangled with distance (Min et al., 2026); that pass@k at large k measures trials (Dragoi et al., 2025); that visual markers beat text coordinates in a supervised setting (Cai et al., 2025). We reproduce all six and cite them at each point of contact.

2 Related Work

Metric depth from VLMs, and visual prompting. DepthLM (Cai et al., 2025) shows a VLM fine-tuned with visual prompting attains competitive metric depth, and its Finding 1 already ablates visual markers against text coordinates in a supervised setting. We do not claim that comparison as ours. Our contribution is to decompose the intervention to a single variable (Sec. 5) and measure what it does to reinforcement learning, where the reward can be collected from a shortcut. Their Table 1 also contains a same-model two-formulation control (eight-dataset mean rising $0.243 \rightarrow 0.400$), which we cite as third-party evidence for the direction of our effect. Concurrent work continues to crowd this space (Yu et al., 2026).

Shortcut audits and spatial representation probing. Fitting a model on non-visual inputs to expose exploitable shortcuts is due to Brown et al. (2025); we claim no originality for the protocol, only for applying it to RLVR training dynamics and ordinal tasks, and for the finding that under the text-coordinate formulation no arm we tried surpassed its own blind baseline. Separately, Why Far Looks Up (Min et al., 2026) reports a vertical–distance entanglement in VLM embeddings and a gap between perspective-consistent and counter-heuristic items on a synthetic benchmark. Our ground-plane analysis overlaps with that phenomenon and we treat it as confirmatory rather than new; what differs is the domain (absolute metric depth on real LiDAR ground truth) and the intervention axis — we show the metric and ordinal shortcuts respond differently to the same training (Appendix Tab. 3).

Reward design and visual shortcuts in multimodal RLVR. AP-GRPO (Jiao et al., 2026) diagnoses advantage collapse for near-miss samples under GRPO (group-relative policy optimisation) and densifies the reward with a parameterised sigmoid; our formalisation (App. A.1) is a criterion for when that failure occurs rather than a remedy, and we supply the boundary condition that density does not help when the formulation leaks. Xu (2026) treats grounding-penalty strength as a control knob and characterises shortcut formation and reversal over training time in video–language models; the control variable there is the regulariser, here it is the task formulation, which we argue is more basic because it governs how much — and at intermediate model capacity, whether a shortcut exists to be penalised. That work reports its effect robust across seeds, which prompted our multi-seed replication.

When RLVR does expand the frontier. Ariadne (Shen et al., 2025) shows RLVR succeeding where the base model's pass@k is identically zero, which is incompatible with any blanket claim that RLVR cannot teach spatial reasoning — and we make none. Our two criteria (Sec. 7) reconcile the observations: their setting has no shortcut and a workable q; ours has a strong shortcut and, for metric targets, $q \approx 0.02$.

3 Auditing Protocol

These seven quantities are the instruments the intervention is measured with, and we report all of them before any claim about spatial ability. The protocol is a consolidation rather than an invention — five of the seven items are adapted from prior work, and we mark each — but consolidating them into one entry point (src/audit_protocol.py) is what makes every number in this paper auditable against the same set of trivial baselines.
The seven quantities are listed in full, with per-item attribution, in App. C: a text-only blind baseline; a constant predictor and a cheating upper bound (the best constant found on the validation set); a random baseline under the same tolerance, including for pass@k; worst-bucket accuracy over five ground-truth quantile buckets; an answer-collapse rate; an adversarial subset on which the position heuristic is wrong by construction; and a cross-domain check. Appendix Figures 11 and 12 show examples from the marker arm and from that subset. Five of the seven are adapted from prior work and we mark each as such; what is ours is the pairing of the blind baseline with the worst bucket into a two-part criterion, and the adversarial construction on real LiDAR ground truth.

Evaluation noise, and why the worst bucket matters more at low scores. Generation uses sampling, so repeated evaluations of the same checkpoint differ. We measured this: four repeats of the base model on the same benchmark give 0.0083–0.0293 (mean 0.0209±0.0103), while two repeats of a trained checkpoint give identical values for two arms and differ by 0.0089 for a third. The absolute noise (≤ 0.021) is an order of magnitude below the main effect (0.1727) and below the seed-to-seed spread (0.0484), so it does not affect any comparison we make. But in relative terms it is large where scores are low (std/mean = 0.49 for the base model), which is a second reason to judge low-scoring arms by the worst bucket rather than by pass@1: the worst bucket was exactly 0.0000 in every repeat.

Two criteria, not one. A single aggregate score is not sufficient to distinguish geometry from distribution matching. Our threshold-reward arm trained on text-coordinate data reaches pass@1 = 0.1888, which exceeds both the blind baseline (0.1197) and the cheating upper bound (0.1745), yet its worst bucket is exactly 0.0000. We therefore require

$$\text{pass@1} > \text{blind} \quad \wedge \quad \text{worst-bucket} > 0, \tag{1}$$

and Appendix Fig. 13 shows only the marker-based arms satisfy both.

4 Controlled Experiment: Only the Pixel Reference Changes

4.1 The shortcut the intervention removes

Mechanism. For a point on the ground plane at pixel row v, the depth follows

$$z \approx \frac{f_y \, h_{\text{cam}}}{v - c_y}, \tag{2}$$

so the pixel row alone carries most of the depth information in driving scenes. Any prompt that states (u, v) therefore leaks the answer to a model that never looks at the image.

Magnitude. Appendix Tab. 7 quantifies the leak. A text-only model reaches 0.2951 on absolute depth in driving scenes — about 5× what a 7B VLM scores on the same items — and 0.8550 on depth ordering. The effect is not specific to driving: the same protocol on indoor scenes gives 0.2355 and 0.8000.
Replacing the coordinates with a drawn marker cuts the text-only score from 0.2951 to **0.1197** in driving scenes and from 0.2355 to 0.2101 indoors (Appendix Tab. 7, last row) — the single number that motivates every controlled comparison below.

Decorrelation is not enough. Rejection sampling to break the correlation moves $\text{corr}(v, \log z)$ only from -0.78 to -0.62; the shortcut survives because the relation is geometric, not statistical.

Why the intervention targets the prompt rather than perception. Prior diagnostic work localises this failure downstream of perception, and we reproduce it: a small probe decodes depth from the frozen representation at 2.2× the model's own behavioural accuracy under the same criterion, and an RLVR-trained model's probe is indistinguishable from the base model's (App. A). Since the information is present but not emitted, we intervene on how the question is posed rather than on the visual encoder.

4.2 Design and what is held fixed

Both arms use the same base model (Qwen2.5-VL-7B-Instruct), RL algorithm (GRPO advantages with a sequence-level ratio in the style of group sequence policy optimisation, GSPO, and filtering in the style of decoupled clip and dynamic sampling policy optimisation, DAPO), reward function, number of training samples, and number of optimisation steps. Two further confounds deserve explicit treatment.

The geometric sample distribution is also controlled. One might worry the marker-based training set is geometrically easier. It is not: $\mathrm{corr}(v, \log z) = -0.6002$ for the marker arm and -0.5902 for the text-coordinate arm, a difference of 0.010; both validation sets are likewise ordinary samples (-0.6190, -0.5310). A decorrelating sampler exists and works (-0.4405 with it versus -0.6466 without, same seed and family), but the main experiment does not use it. The observed effect therefore cannot be attributed to sample distribution.

The composite intervention is decomposed. Switching to markers removes (u, v) from the prompt and alters image pixels, possibly providing an attentional anchor. We separate them with an isolation arm in which the point is both drawn and named. Its blind baseline is 0.2935, within 0.0016 of the pure text-coordinate formulation (0.2951) and 0.1738 from the marker formulation (0.1197). Shortcut strength is therefore governed entirely by the prompt at the blind level; Section 5 shows this does not extend to the ability level.

4.3 Result

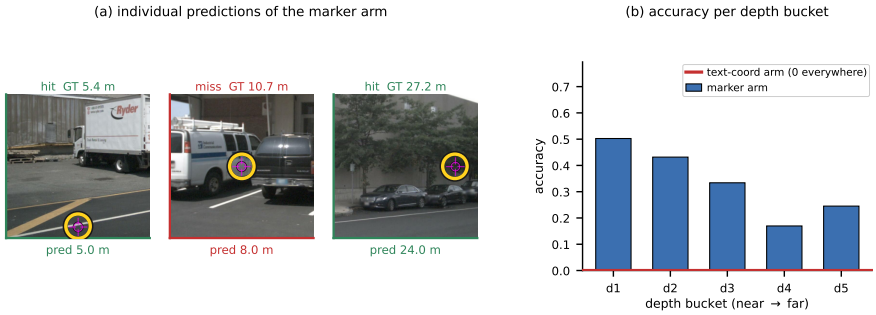


Figure 2: What the numbers look like on real queries. Three predictions of the marker-trained arm, one outside the tolerance band and two inside, with the LiDAR ground truth and the prediction printed under each crop. Right: per-bucket accuracy of the same arm; the text-coordinate arm is at zero in every bucket, which is the signature the second criterion is designed to catch. The yellow ring is added here for legibility only; the model's input contains the magenta cross alone, which is what the prompt names.

Table 1: Main controlled experiment, three paired seeds per arm. All arms on the same benchmark ($n = 241$; blind = 0.1197, cheating bound = 0.1745, random = 0.0601). The text-coordinate arm's worst bucket is exactly zero in all three seeds, across two paradigms. Paired t-test over seeds. †Base accuracy is mean±std over four repeated evaluations (sampling noise is large in relative terms at this level: range 0.0083–0.0293; small in absolute terms, 0.021, versus a main effect of 0.1727). Paired difference: $+\mathbf{0.1727 \pm 0.0575}$, $t = 5.20$, $\mathbf{p = 0.036}$. Per-seed differences: +0.1888, +0.1088, +0.2204 (all positive). The worst-bucket criterion is re-tested at five times the per-bucket count in Sec. 4.4, where the text-coordinate arm remains at exactly 0 out of 240.

Arm	pass@1	worst-bucket
Base (untrained)†	0.0209±0.0103	0.0000
Text coords, thresh. (1 sd)	0.1888	0.0000
Text coords (3 sd)	0.1324±0.0098	0.0000 (all)
Markers (3 sd)	**0.3050±0.0484**	**0.1546±0.0541**

Table 1 lists all arms on the same benchmark. With three paired seeds per arm, the paired difference is $+0.1727 \pm 0.0575$ ($t = 5.20$, $p = 0.036$), and all three per-seed differences are positive. On this backbone the text-coordinate arm's worst bucket is exactly zero in every

seed and under both paradigms (RLVR and SFT), and its highest pass@1 (0.1415) has a 95% confidence interval (CI) that contains the blind baseline — no seed surpassed it. Two facts make the conclusion robust to the modest sample size. Even the weakest marker seed (0.2503) exceeds the cheating upper bound (0.1745), so "beats the distribution-matching ceiling" does not rest on selecting a favourable run. And on the worst bucket the text-coordinate arms score 0/48 (Wilson upper bound 0.0741) while the marker arm's CI lower bound is 0.0870 — the two intervals do not overlap.

Is this just distribution matching? The sharpest objection is that the marker arm is evaluated in the formulation it was trained in. Three answers. By construction of the metric: a constant predictor scores exactly zero on the worst bucket, so a non-zero worst bucket cannot come from matching the answer distribution — and the text-coordinate arms sit at 0.0000 in all three seeds while the marker arms do not. By controlling sample geometry: the two training sets differ in $\mathrm{corr}(v, \log z)$ by 0.010. Decisively, by voiding the shortcut: on an adversarial subset of off-ground points the marker arm still reaches 0.1837 with a CI lower bound above both the blind baseline and the cheating bound (Sec. 6) — a distribution matcher cannot survive that subset. A fourth signature is distributional rather than accuracy-based: the text-coordinate arm concentrates far more of its outputs on a single value than the marker arm does; App. I shows the distributions and the seed-averaged rates. The same contrast appears under an SFT carrier, so the effect is not specific to the RL algorithm; App. H gives the numbers.

4.4 How far the result holds

Seeds. Both arms are covered at three seeds each. Appendix Tab. 8 lists them individually. The marker arm gives 0.3050 ± 0.0484 and the text-coordinate arm 0.1324 ± 0.0098, whose worst bucket is exactly zero at every seed and under both the RLVR and the SFT carrier. The paired difference over the three seed pairs is $+0.1727 \pm 0.0575$ ($t = 5.20$, $p = 0.036$). Seed variance is not negligible (marker range 0.0921), so we report intervals rather than single numbers; even the weakest marker seed (0.2503) exceeds the cheating upper bound 0.1745, and pre-training baselines agree across seeds (0.0293/0.0293/0.0228), so the spread comes from training rather than from evaluation noise.

Scale, and what capacity changes. Appendix Fig. 5a and Tab. 4 run the same intervention on three backbones. The direction is positive in all three, with comparable magnitude; what changes is the form. At 3B neither formulation lifts the worst bucket off zero, and the marker arm does not clear the trivial baselines — margin over blind +0.0539 is not significant and the paired difference is only marginal ($p = 0.058$), so at this scale the formulation is necessary but not sufficient. At 7B the marker arm clears them and the text-coordinate arm does not: this is the regime in which the formulation can be said to decide whether the ability appears. On the newer 8B backbone both arms reach a non-zero worst bucket, and the effect survives as a difference in degree. Over three paired seeds the pass@1 gap is $+0.1801 \pm 0.0313$ ($t = 9.96$, $p = 0.0099$) and the worst-bucket gap is $+0.1484 \pm 0.0250$ ($t = 10.30$, $p = 0.0093$); the text-coordinate arm's worst bucket is 0.2109, 0.1699 and 0.2439, whose Wilson intervals all exclude zero. We therefore state the scope explicitly: that a text-coordinate formulation leaves the worst bucket at exactly zero is a property of the capacities we measured it on, not a backbone-independent law. What holds across all three is the sign and rough size of the pass@1 gap.

A newer, spatially-optimised backbone. Appendix Fig. 5b puts the audit on three backbones side by side. Qwen3-VL-8B, released nine months after the 2.5 generation and advertising improved 3D grounding, shows the same dissociation before any training: ordinal 0.8069 (95% CI $[0.7545, 0.8504]$) against a blind baseline of 0.4883, but metric 0.0125 whose CI upper bound (0.0360) still falls below the blind baseline, with a worst bucket of zero. InternVL3_5-8B, from a different lab and with a different vision tower, agrees (0.7683 ordinal; 0.0585 metric with upper bound 0.0951). On the text-coordinate version of the benchmark all four question families (depth order, height order, metric depth, metric gap) fall below their blind baselines, for both of these backbones. The advertised upgrade therefore does not translate into absolute metric competence as measured under a blind-baseline audit.

Sample size, and the resolution of the worst-bucket test. Because the criterion concerns a worst bucket, its resolution is set by the per-bucket count. At 48 per bucket a Wilson interval has half-width ±0.0996, which is why the worst-bucket paired difference of Sec. 5 is directionally clear but not significant. Rebuilding the benchmark at 240 per bucket — with every reference line recomputed on the new set (blind 0.1104, cheating bound 0.1545) — leaves the text-coordinate arm's worst bucket at exactly 0 out of 240, tightening the Wilson upper bound from 0.0741 to **0.0158**; the marker arm reaches 0.1533 there, a difference of +0.1542 with 95% CI $[+0.1110, +0.2053]$. Absolute scores are lower on this harder set (its cheating bound is lower too), so all readings are taken against reference lines computed within the same set.

5 Decomposing the Intervention

Switching to visual markers changes two things at once: it removes (u, v) from the prompt, and it alters the image. We separate them with an arm in which the point is both drawn on the image and named by coordinates.

At the shortcut level, only the prompt matters. The blind baseline of this arm is 0.2935, within 0.0016 of the pure text-coordinate formulation (0.2951) and 0.1738 away from the marker formulation (0.1197). Drawing a marker neither introduces nor removes the textual shortcut.

At the ability level, both matter — and the prompt effect is measurable in isolation. Because this arm and the marker-based SFT arm are trained on identical images for the same number of steps, their difference isolates the prompt (Tab. 2; Appendix Fig. 6a).

Table 2: The cleanest single-variable contrast in this paper, over three seeds. Both arms are SFT, 300 steps, trained on the same images (markers drawn) and evaluated on the same benchmark ($n = 241$); the only difference is whether the prompt also states (u, v). Paired t-test over seeds. The pass@1 effect is significant; the worst-bucket effect is not, driven by seed 43 where the two arms nearly coincide. Paired difference on pass@1: $+\mathbf{0.0854} \pm \mathbf{0.0322}$, $t = 4.59$, $\mathbf{p = 0.045}$. On the worst bucket: $+0.0970 \pm 0.0791$, $t = 2.12$, $p = 0.168$ (n.s.).

	markers		+ coords	
Seed	pass@1	worst-bucket	pass@1	worst-bucket
42	0.3224	0.1836	0.2394	0.0447
43	0.2503	0.0921	0.1958	0.0863
44	0.3424	0.1880	0.2236	0.0417
mean	0.3050	0.1546	0.2196	0.0576

Over three seeds the paired difference is $+0.0854 \pm 0.0322$ on pass@1 ($t = 4.59$, $p = 0.045$), with all three seeds positive; removing the coordinates therefore has a significant effect on pass@1 that is not attributable to seed noise. The worst-bucket difference ($+0.0970 \pm 0.0791$, $p = 0.168$) is not significant — on seed 43 the two arms nearly coincide (0.0921 vs. 0.0863) — so we claim the pass@1 effect only. Since the aggregate effect in Table 1 is larger, that number should be read as two interventions summed, not one.

An important negative: coordinates do not abolish learning. The decomposition arm reaches 0.3861 on its own coordinate-bearing validation set (95% CI $[0.3444, 0.4294]$, whose lower bound exceeds its own blind baseline 0.2935), with all five buckets non-zero. Transferred to the coordinate-free benchmark it still scores 0.2394 with worst-bucket = 0.0447, satisfying both criteria. Hence removing the coordinates is not an on/off switch for geometric ability but a factor governing its magnitude; a worst bucket of exactly 0.0000 occurs only when the prompt gives coordinates and the image carries no marker.

6 Dissociation of Ordinal and Metric Ability

6.1 The two abilities separate on the same benchmark

On the marker-based benchmark (blind 0.4883 ordinal, 0.1197 metric), the base model and a spatial-RLVR model both score far above blind on ordering (0.7467/0.8267) and both below it on metric depth (0.0151/0.0126; Appendix Fig. 15). The pattern reproduces

on BLINK (Fu et al., 2024), a third-party benchmark that already uses visual markers: Relative_Depth improves 0.6935 → 0.8548 (blind 0.5161) and Spatial_Relation 0.8462 → 0.9091 (blind 0.5385). It is not a scale artefact: at 3B the same model family reaches 0.7883 on ordinal but 0.0017 on metric.
On an adversarial ordinal subset where the position heuristic is wrong by construction, both models fall significantly below chance, and a mixture estimate puts their reliance on it at 0.41 and 0.33; App. G has the details.

6.2 Training removes the metric shortcut but not the ordinal one

This is the sharpest form of the dissociation, and it is visible within a single trained model. We evaluate the marker-trained arm on two adversarial sets built with the same ground-plane logic, one metric and one ordinal (Appendix Tab. 3). On the metric subset its CI lower bound 0.1544 exceeds both the blind baseline 0.0906 and the cheating bound 0.1358, its worst bucket is non-zero, and the gain over the untrained base is +0.1517 (95% CI $[+0.1180, +0.1865]$). Since the ground-plane relation is unusable on off-ground points, this is the strongest single piece of evidence here that the marker formulation yields real metric competence. On the ordinal subset the same arm scores 0.2733 — still below chance, and statistically indistinguishable from the untrained model's 0.2967 (difference −0.0234, 95% CI $[-0.028, +0.074]$). Training on markers therefore removed the metric shortcut while leaving the ordinal position prior in place: the two shortcuts are not one mechanism, and an intervention that voids one need not touch the other.

7 Discussion

Why the formulation acts before the algorithm and the reward. A verifiable reward can only select among behaviours the policy already produces. When the pixel is named as text, a policy that regresses depth from the row satisfies the reward on most of the distribution, so the reward has no gradient to distinguish it from one that inspects the image; the shortcut is not merely available but sufficient. Drawing the marker removes that sufficiency, and the same reward now separates the two behaviours. This is why the effect appears before any choice of algorithm, and why supervised fine-tuning reproduces it at a fifth of the compute: nothing in the mechanism is specific to policy-gradient optimisation. It also explains why the question could not be settled by measurement alone. Every benchmark in this space couples the pixel reference to a great deal else, so no purely observational comparison isolates it; only an intervention that holds the rest fixed can attribute the difference to the formulation.

Two measurable conditions. Before claiming that RLVR improved a spatial ability we suggest reporting the margin of a text-only blind baseline over chance, and whether the worst bucket is non-zero; App. B explains why each matters.

Metric and ordinal are different targets, and neither transfers freely. Ordering two depths needs only a monotone cue, and the pixel row supplies one; naming a distance in metres needs a calibrated map. This is why the two abilities dissociate on the same images and why training moves one without the other: our arm's ordinal accuracy stays below chance on a matched adversarial set, indicating the position prior is untouched. It also predicts the domain result. The ground-plane relation our driving-scene arm must learn is a scene-geometry regularity, not a general one, so an arm trained only on driving markers has no reason to calibrate indoors — and indeed does not, falling below the indoor blind baseline. The transferable part is the ordinal comparison, which is exactly the part that a monotone cue already supports.

On evaluation: pass@k needs a random baseline, and on shortcuts we did not find. For a continuous target with a tolerance band, drawing k samples raises the chance that one lands inside the band even without depth information, so a random predictor has a non-trivial pass@k that grows with k — on our hardest bucket the base model's pass@8 falls below random search (Appendix Fig. 4). Reporting pass@k without that anchor can present coverage of the answer space as reasoning. Two shortcuts we expected are absent: marker colour and size do not predict depth, and accuracy is unchanged when markers are re-rendered at a different radius — which is what makes the positive result interpretable.

8 Limitations

Two views of the same mechanism: the text-coordinate arm places 51.2% of its answers on the single value that is the depth distribution's median — the best constant guess, obtainable without the image — while the marker arm spreads over three times as many distinct values; and both arms' loss curves and gradient norms stay in one band with no sign of a broken run, so the formulation changes the ability the optimisation reaches, not whether it converges (Appendix Figs. 7, 8, Tab. 5).

Model coverage and the version ceiling. Our training interventions cover two backbone generations of one family (Qwen2.5-VL-7B and Qwen3-VL-8B) plus a 3B contrast, with one RL implementation (EasyR1/GRPO); we did not vary the RL algorithm, so "formulation matters more than algorithm" is supported for this family of methods rather than for RLVR in general. The diagnosis is not vendor-specific: repeating the audit on InternVL3_5-8B — a different lab, and an InternViT rather than a Qwen vision tower — reproduces the dissociation (ordinal 0.7683 vs. blind 0.4883; metric 0.0585 with CI upper bound 0.0951 still below blind; all four text-coordinate families 0.127 to 0.185 below their blind baselines). The intervention, however, does not carry over. We ported the two-arm contrast to InternVL3_5-8B, first verifying that its image channel participates in the loss and that training is healthy (loss 3.78 $\rightarrow$ 0.21, matching the Qwen runs), and found no effect: the marker arm reaches 0.0303 against the text-coordinate arm's 0.0580, a difference of -0.0277 whose interval contains zero. Both arms sit far below that benchmark's blind baseline of 0.1197, so this is not a reversed effect but a backbone that learns no measurable metric ability here at all; the difference between two arms that both fail is not interpretable. Its ordinal accuracy, by contrast, is 0.78–0.80 in both arms, far above the ordinal blind baseline but essentially unmoved by training — that ability was there before. Its parameter count matches Qwen3-VL-8B, so capacity in the sense of Sec. 4.4 does not explain the failure. Neither does a weaker vision tower: probed identically, this backbone's frozen features are the most decodable of the three we tested ($\delta < 1.25$ of 0.3275 under an MLP probe against 0.3075 and 0.2225 for the two Qwen backbones; this three-way comparison uses its own probe fit, so its absolute values are not comparable with Sec. A.4's), so the information is present but unused. What fails is the conversion of that information into output behaviour — the same gap Sec. A.4 documents within a single model, here in a more extreme form: more decodable representation, worse behaviour. The formulation effect is therefore established for one model family across two generations, and tested and not reproduced on a second vendor. Separately, an infrastructure ceiling rather than a choice kept us off the newest releases; App. L gives the driver and Python constraints and what they excluded.

Five limitations deferred for space. App. E records a failure that produced no error and a plausible-looking loss curve when we ported the intervention to a newer backbone, the limits of the shortcut analysis itself, a pre-registered cross-domain hypothesis that failed, a planned pass@k test whose metric turned out not to discriminate, and our sample sizes.

Corrections to our own earlier analysis. Six earlier mis-readings of our own data were each caught by one of the seven metrics we advocate; App. D lists them. That they were caught is the strongest argument we can give for reporting all seven.

9 Conclusion

We asked what a spatial RLVR recipe learns when the shortcut it usually leans on is removed before training rather than diagnosed afterwards. Holding the model, algorithm, reward, data volume, step count and sample geometry fixed and changing only how the prompt refers to a queried pixel, metric accuracy more than doubles and every depth bucket comes off zero — while under the coordinate formulation the hardest bucket stays at exactly zero across seeds and under both carriers.

Two boundaries are as informative as the effect: once capacity is sufficient the formulation governs how much rather than whether, and on a second vendor's backbone the intervention does not reproduce at all. Both point at the same place — the bottleneck sits downstream of perception, and a reward can only select among behaviours the formulation has made distinguishable. Before claiming that RLVR improved a spatial ability, report the blind baseline's margin over chance and whether the worst bucket is non-zero.

AI Use Statement

We made limited use of a generative AI assistant throughout this project. Its role was confined to text polishing, grammar checking, result consistency verification, and literature search. Specifically, it was used for: (i) polishing the text and correcting grammar and syntax errors in this manuscript, including compressing sections to meet the page limit; (ii) running consistency checks between the numbers in the paper and the registry that records which run produced each one; and (iii) searching the literature for prior work that anticipates our findings.

Every empirical claim rests on a run we launched and whose output we inspected. The number-to-run correspondence is machine-checked: each measured value in the text carries a provenance tag, and a script we wrote verifies that no value appears in the paper that is absent from the registry while also performing consistency checks. That check currently reports zero orphans. Literature found with AI assistance was read and verified by hand; author lists and titles in the bibliography were checked against the arXiv listing rather than transcribed, which corrected three titles we had recorded incorrectly or incompletely.

Several conclusions in this paper are weaker than our initial hypotheses, and in each case the weakening was forced by a measurement rather than suggested by the assistant. We take full responsibility for the content of this work, including the text polished with AI assistance.

Ethics Statement

This work uses only nuScenes, a public autonomous-driving dataset released for research use, and publicly released vision–language model checkpoints. No new data was collected, no human subjects were involved, and no personally identifying information was annotated or released; the query targets are LiDAR points projected into camera frames, not people.

Two ethical considerations are worth stating explicitly. First, the paper's central claim is that a spatial ability which looks present under one task formulation can be almost absent under another. Depth estimation is safety-relevant, and a benchmark number that is inflated by a formulation shortcut could be read as evidence of a competence a deployed system does not have. We regard reporting that gap as the socially useful direction, and the two checks we recommend (a blind-baseline margin and a non-zero worst bucket) are cheap enough to adopt.

Second, we report a negative result: the intervention does not reproduce on one of the three backbones we tested. We report it in full rather than restricting the paper to the backbones where it works, because selective reporting of which models an effect holds on is itself a form of overclaiming.

Reproducibility Statement

Appendix J states the data construction (splits, projection and visibility filtering, and exact per-split counts), the verifier that serves both as the RLVR reward and as the evaluation metric, and the full training configuration for both arms in Table 6; every value there is read from the shipped configuration files rather than transcribed. The auditing protocol that all numbers are computed under is given as a seven-item procedure in Appendix C, and the blind-baseline fits it depends on are tabulated per dataset. Appendix F gives the probing setup, and Appendix M documents the diagnostic that distinguishes a genuinely negative result from a silently broken run — the one failure mode we consider most likely to mislead a reader attempting to reproduce this work on a different backbone. Appendix L records the inference-stack version ceiling our hardware imposes, because results are not reproducible on a newer stack without also changing the driver.

References

S Divakar Bhat and Toshihiko Yamasaki. Consistent yet wrong: Evidence insensitivity in spatial vision-language models. arXiv preprint arXiv:2606.02742, 2026. v2, 29 Jun 2026. Shows leading VLMs give view-invariant yet incorrect answers on metric distance queries, and evaluates metric accuracy, distributional concentration and internal collapse (via a latent feature probe) on the ViewDiag protocol. Concludes that stable predictions may reflect prior-driven collapse rather than evidence-sensitive reasoning. Our answer-collapse

statistic and our probe result are confirmations of that phenomenon; our contribution is the causal link from task formulation to the degree of collapse.

Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should "train on the test set" to expose exploitable non-visual shortcuts. arXiv preprint arXiv:2511.04655, 2025. Source of the blind-baseline auditing methodology used here. Principle: "if a benchmark can be gamed, it will be.".

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, et al. DepthLM: Metric depth from vision language models. arXiv preprint arXiv:2509.25413, 2025. ICLR 2026 Oral. Finding 1 ablates visual markers vs. text coordinates in a supervised setting; Table 1 contains a same-model two-formulation control (8-dataset mean 0.243 -> 0.400). 3B reaches $\delta_1 = 0.799$ on nuScenes, 7B reaches 0.865.

Marius Dragoi, Ioana Pintilie, Florin Gogianu, and Florin Brad. Beyond Pass@k: Breadth-depth metrics for reasoning boundaries. arXiv preprint arXiv:2510.08325, 2025. v2, 21 Oct 2025. Argues that on tasks with discrete answer spaces Pass@k at large k reflects the chance of success in the limit of trials rather than genuine reasoning, and proposes Cover@tau (fraction of problems where at least a tau proportion of completions are correct). Our pass@k observations are a confirmation of this point in a different regime (continuous targets with a tolerance band); our worst-bucket criterion thresholds on sample difficulty rather than on completion consistency.

Karim Elmaaroufi, Liheng Lai, Justin Svegliato, Yutong Bai, et al. GRAID: Enhancing spatial reasoning of VLMs through high-fidelity data generation. arXiv preprint arXiv:2510.22118, 2025. Reports a human pass rate of 0.576 for mainstream spatial data pipelines; used here as the comparison point for our 0.9714.

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. arXiv preprint arXiv:2404.12390, 2024. ECCV 2024. 14 visual perception tasks; the spatial subset uses drawn visual markers, which is why its blind baseline is near chance.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, et al. SURDS: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. arXiv preprint arXiv:2411.13112, 2024. Discretises depth into buckets and converts distance into pairwise comparison. Its Table 2 reports Random = 34.27 on the depth bucket, above GPT-4o (2.49), Gemini-1.5-pro (22.70) and Qwen2.5-VL-3B (27.68), without identifying this as a problem.

Siwen Jiao, Tianxiong Lv, Kangan Qian, Chenxu Zhao, Xiuyuan Zhu, Tianlun Li, Xiaolong Cheng, Jinyu Li, Zhihao Liao, and Yang Cai. Smooth operator: Smooth verifiable reward activates spatial reasoning ability of vision-language model. arXiv preprint arXiv:2601.07695, 2026. v2, 15 Jan 2026. Diagnoses near-miss advantage collapse under GRPO; proposes SNRA (dynamic sigmoid dense reward) and AP-GRPO; releases Numerical3D-50k.

Chih-Ting Liao, Fei Shen, Xin Cao, and Tat-Seng Chua. Decodable is not grounded: A vision-ablation arbiter for VLM spatial reasoning. arXiv preprint arXiv:2606.31257, 2026. 30 Jun 2026. Shows a linear probe plus steering recovery can systematically overstate what a VLM grounds in the image, and that a one-line causal control — replacing the image with a gray blank — exposes the error. Identifies three grounding regimes (grounded / prior / inverted) across 14 models spanning six LM families and 2B–27B; finds horizontal grounded, vertical a prior, and depth inverted (below chance). Our blind baseline is the text-only analogue of their blank-image arbiter; our probe results and our below-chance ordinal finding are confirmations within their taxonomy. What we add is a training intervention: we change the formulation and retrain, which none of these diagnostic works do.

Cheolhong Min, Jaeyun Jung, Daeun Lee, Hyeonseong Jeon, Yu Su, Jonathan Tremblay, Chan Hee Song, and Jaesik Park. Why far looks up: Probing spatial representation in

vision-language models. arXiv preprint arXiv:2605.30161, 2026. Submitted 28 May 2026. Reports vertical-distance entanglement, a perspective-consistent vs. counter-heuristic accuracy gap that intensifies under data scaling, and releases the SpatialTunnel synthetic benchmark.

Minghe Shen, Zhuo Zhi, Chonghan Liu, Shuo Xing, et al. Does RLVR extend reasoning boundaries? investigating capability expansion in vision-language models. arXiv preprint arXiv:2511.00710, 2025. v4. Shows RLVR succeeding on synthetic mazes where base pass@k is identically zero. Cited here as the counterexample our two criteria must reconcile.

Zekun Xu. When does a video-language model stop watching? reward strength controls the formation and reversal of visual shortcuts in multimodal RLVR. arXiv preprint arXiv:2606.22043, 2026. Submitted 20 Jun 2026. Grounding-penalty strength as a control knob; reports sharp onset robust across random seeds, monotone dose-response, and a critical intervention window.

Hanxun Yu, Xuan Qu, Yuxin Wang, Jianke Zhu, and Lei Ke. Unlocking dense metric depth estimation in VLMs. arXiv preprint arXiv:2605.15876, 2026. Cited only to indicate the density of concurrent work on metric depth from VLMs.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, et al. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? arXiv preprint arXiv:2504.13837, 2025. Argues that RLVR mainly redistributes existing capability (pass@1 up, pass@k flat or down) rather than expanding the reasoning boundary. Cited as the language-domain precedent for our probe result.

A Supplementary Figures and Tables

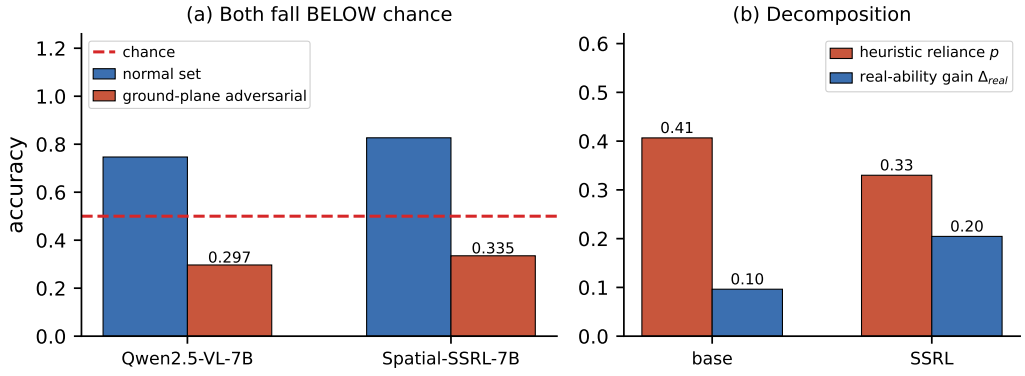


Figure 3: Ordinal accuracy falls below chance when the position heuristic is wrong by construction. (a) Both models score under 0.5 on the adversarial subset, with Wilson CI upper bounds below chance. (b) Decomposition of Eq. 4 into heuristic compliance and residual real ability; the compliance term dominates in both models.

A.1 Why sparse verifiable rewards fail on continuous targets

Let $q = P(\text{relative error} \leq 2\rho)$ be the probability that a sampled answer falls inside the tolerance band. For a group of G rollouts sharing one advantage normalisation,

$$P(\text{no gradient for the whole group}) \;\geq\; (1-q)^G. \tag{3}$$

Substituting into Eq. 3 the values measured on the base model, with ρ set to the verifier's own tolerance: $q = 0.0250$ for the metric-gap family, giving 81.67% gradient-free groups at the $G = 8$ we use; for depth ordering $q = 0.5625$, giving only 1.1% (Appendix Fig. 14).

Table 3: The same arm passes the metric adversarial test and fails the ordinal one. Both subsets void the ground-plane prior by construction. "Untrained" columns are the base VLM and a spatial-RLVR model. Off-ground subset: points whose ground-plane prediction errs by $\geq 2\times$; $n = 600$ each.

Adversarial subset	marker arm	untrained	blind / chance
Metric, off-ground pts	**0.1837**	0.031 / 0.012	0.0906
95% CI	[.154, .216]	—	cheat 0.1358
worst-bucket	0.0927	—	—
Ordinal, ground-plane	0.2733	0.2967 / 0.3350	0.5

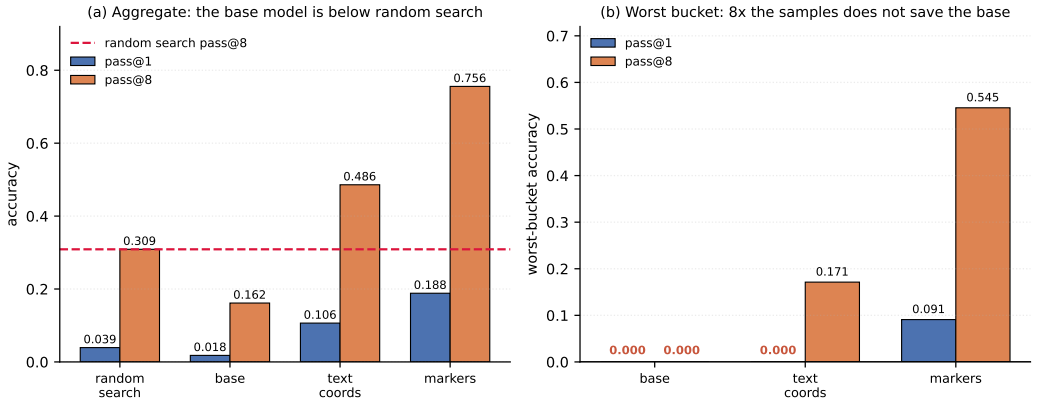


Figure 4: Why pass@k needs a random anchor, and why the worst bucket is more reliable. (a) Aggregate pass@k on metric depth ($n = 56$). The base model's pass@8 (0.1615) falls below random search (0.3089, dashed line, mean of 20 repetitions), so reading it without the anchor overstates its capability; both trained arms clear the anchor, and the marker arm exceeds the text-coordinate arm by $+0.2679$ (95% CI $[+0.0878, +0.4251]$). (b) The same runs restricted to the hardest depth bucket ($n \approx 11$). The base model scores exactly zero at $k = 8$ as well — eight samples, none inside the tolerance — so a zero in the worst bucket cannot be removed by spending more samples. The gap between the two trained arms is not significant at this per-bucket sample size.

Relation to prior work, and the boundary we add. The near-miss advantage collapse in GRPO was diagnosed by Jiao et al. (2026), which densifies the reward with a dynamically parameterised sigmoid. Our formalisation is a criterion rather than a method. Crucially, our text-coordinate $\times$ shaped-reward arm is a counterexample to the hope that densification suffices: with shaping it still reaches only 0.1336 (CI $[0.0956, 0.1814]$, which contains the blind baseline 0.1197) and worst-bucket = 0.0000. Densifying the gradient does not help when the formulation leaks a shortcut — it merely finds the shortcut faster.

A.2 Where the failure lies: capability ladder and probes

Section 6 establishes that metric competence is absent in models not trained for it. This section localises the failure: it is not that the visual encoder lacks the information, but that the model cannot emit it as a number.

A.3 A capability ladder

We ask the same underlying geometric question in five answer formats of decreasing freedom, from free-form metres to five-way multiple choice, and compare each against its own chance level and its own cheating bound (Appendix Fig. 16). Under five-way multiple choice (chance 0.20) models reach 0.3800–0.3950 — roughly twice chance, and consistently so across four model/prompt combinations. Every free-numeric format, by contrast, falls below its own ceiling: ratio 0.098–0.209 against a bound of 0.3383, anchored 0.012–0.066 against 0.1559, absolute depth 0.005–0.022 against 0.1706. The discriminative signal is present; the generative one is not.

A.4 Probing the representation

A probe on the frozen visual features recovers depth far better than the model's own behavioural output ($\delta < 1.25$ of 0.3475 versus 0.1587), and a probe trained on the spatial-RLVR model scores the same as one on the base model: RLVR reshaped behaviour, not representation. App. F gives the numbers and the cross-vendor comparison.

B The Two Reporting Conditions

Two measurable conditions, and what they imply for benchmarks. Our results suggest two quantities worth reporting before claiming that RLVR improved a spatial ability. The first is the margin of the blind baseline over chance: when a text-only predictor already reaches most of the achievable score, training has little room to reward looking. The second is whether the worst bucket is non-zero, which distinguishes a policy that has learned the relation from one that has learned the answer distribution — a constant predictor scores exactly zero there by construction. Neither requires additional annotation, and both are cheap: the blind baseline is a ridge regression on the question text, and the bucket criterion is a re-partition of results one already has. Benchmarks that report only aggregate accuracy against chance cannot distinguish these cases, and our text-coordinate arm is an example of a model that would look competent under that reporting.

C The Auditing Protocol in Full

These are the seven instruments referenced in Sec. 3. The protocol is a consolidation rather than an invention: five of the seven items are adapted from prior work and we mark each. What consolidating them into one entry point (src/audit_protocol.py) buys is that every number in this paper is auditable against the same set of trivial baselines. The cost of omitting a blind baseline is visible in print: SURDS (Guo et al., 2024) reports a random baseline of 34.27 on its depth bucket, above GPT-4o's 2.49, without flagging it.

1. Blind baseline. A regressor/classifier trained on the prompt text only, never seeing the image. Adapted: the methodology is due to Brown et al. (2025), and it is the text-only analogue of the blank-image arbiter of Liao et al. (2026). Our use of it inside RLVR training dynamics and on ordinal tasks is what is new here.
2. Constant predictor (train-set median) and cheating upper bound (best constant found by grid search on the validation set) — the distribution-matching ceiling. Standard practice, stated for completeness.
3. Random baseline under the same tolerance — including for pass@k, where random search can exceed a base model (Sec. 7); the general point is due to Dragoi et al. (2025).
4. Worst-bucket accuracy over five ground-truth quantile buckets. A constant predictor scores exactly 0 here by construction. This pairing with the blind baseline, as a two-part criterion, is ours; a reliability-thresholded alternative that conditions on completion consistency rather than sample difficulty appears in Dragoi et al. (2025).
5. Answer-collapse rate: the frequency of the modal answer. Adapted: collapse onto priors in spatial VLMs is documented by Bhat & Yamasaki (2026); we use the rate as a per-arm statistic to compare formulations.
6. Adversarial subset on which the position heuristic is wrong by construction (chance = 0.5, heuristic = 0.0). Adapted: the counter-heuristic construction appears in Min et al. (2026) on synthetic images; ours is built on real LiDAR ground truth, and we add a metric (off-ground) counterpart.
7. Ground-truth validation: multi-frame reprojection consistency, point-in-box depth agreement, and triangle inequality. Our own, and reported because the adversarial subsets stand or fall on it.

D Corrections to Our Own Earlier Analysis

Six earlier mis-readings of our own data — that the gain persisted across domains, that general ability degraded, that sparse rewards were simply the wrong tool, that ordinal accuracy reflected real geometry, that drawing a marker is irrelevant to shortcut strength (true at the blind level, false at the ability level, Sec. 5), and that our marker training data was "debiased" in the sense of decorrelated sampling (it is not: the sampling is ordinary

and the debiasing is entirely in the formulation) — were each caught by one of the seven metrics we advocate. That is the strongest argument we can give for reporting all of them.

Table 4: The same intervention on three backbones. The direction of the formulation effect is positive in all three; what changes is whether the text-coordinate arm reaches a non-zero worst bucket at all. Reference lines for this benchmark: blind 0.1197, cheating bound 0.1745. All three rows are means over three paired seeds; the 8B pass@1 gap is $+0.1801 \pm 0.0313$ ($p = 0.0099$).

Backbone	Marker p@1 / worst-bucket	Text-coord. p@1 / worst-bucket	gap
Qwen2.5-3B	0.1738 / **0.0000**	0.0191 / **0.0000**	+0.1547
Qwen2.5-7B	0.3050 / 0.1546	0.1324 / **0.0000**	+0.1727
Qwen3-VL-8B	0.5353 / 0.3566	0.3552 / **0.2082**	+0.1801

The heuristic-reliance estimate. Writing a_{adv} for the heuristic's accuracy on the adversarial subset and assuming a model follows the heuristic with probability p and is otherwise at chance,

$$p = \frac{0.5 - \text{acc}_{\text{adv}}}{0.5 - a_{\text{adv}}}, \tag{4}$$

which gives $p = 0.41$ and 0.33 for the two models. This rests on a two-component mixture assumption and its output is derived rather than observed; we report it as an estimate.

E Further Limitations

These are deferred from Sec. 8 for space; none of them changes a result, but each bounds how a result should be read.

A pre-registered hypothesis that failed. We specified in advance that the learned metric ability should transfer across domains: an arm trained only on driving-scene markers should exceed the indoor blind baseline (0.2101) with a non-zero worst bucket. It does not — 0.0950 (95% CI $[0.0700, 0.1277]$, whose upper bound is below the blind baseline) with a worst bucket of 0.0000. It remains the best of the three models we measured there (4.2× the base VLM's 0.0227), but the criterion is not met and we report the hypothesis as falsified. Ordinal ability does transfer (0.8350, +0.3025 over blind). Sec. 7 argues this outcome is consistent with our mechanism rather than separate from it.

A planned test whose metric turned out not to discriminate. We planned a pass@k comparison to test whether the marker formulation establishes capability or merely elicits existing capability. The measurement itself succeeded — the base model goes from 0.0179 at $k = 1$ to 0.1615 at $k = 8$ on metric depth — but a control we ran alongside it invalidates the test: uniformly random answers drawn from the observed value range, at the same tolerance, reach pass@8 = 0.3089 (mean over 20 repetitions; 0.3937 under log-uniform sampling, which is closer to model output). The base model's pass@8 is therefore below random search. On a continuous target with a 12% relative tolerance, pass@k measures sampling coverage far more than capability, and the contamination grows with k (random reaches 0.4777 at $k = 16$, and 0.6929 under log-uniform sampling). We report this rather than a conclusion about elicitation. The ordinal task fails the same test for a different reason: with a binary answer, random guessing reaches $1 - 0.5^8 = 0.9961$, above the base model's 0.8864 (CI upper bound 0.9505). The worst bucket at $k = 1$ remains the stricter test, because random search scores zero there — it would have to hit the hardest depth decile as well. The underlying question (establish vs. elicit) therefore stays open; our probe result (App. A.2) is the only evidence we have, and it points toward elicitation.

Sample sizes, seeds, and derived quantities. The headline criterion concerns a worst bucket, so its resolution is set by the per-bucket count. We therefore rebuilt the metric benchmark at 240 examples per bucket and recomputed every reference line on it; the text-coordinate arm's worst bucket is exactly 0 out of 240, which tightens the Wilson upper bound from 0.0741 to 0.0158 (Sec. 4.4). Two quantities remain thinner than we would like: all three capacity rows now rest on three paired seeds each, and the human audit's non-adversarial pass rate (0.9714) has a CI lower bound of 0.8547, below the 0.90 we would want to assert. We report both as point estimates with intervals rather than as established values.

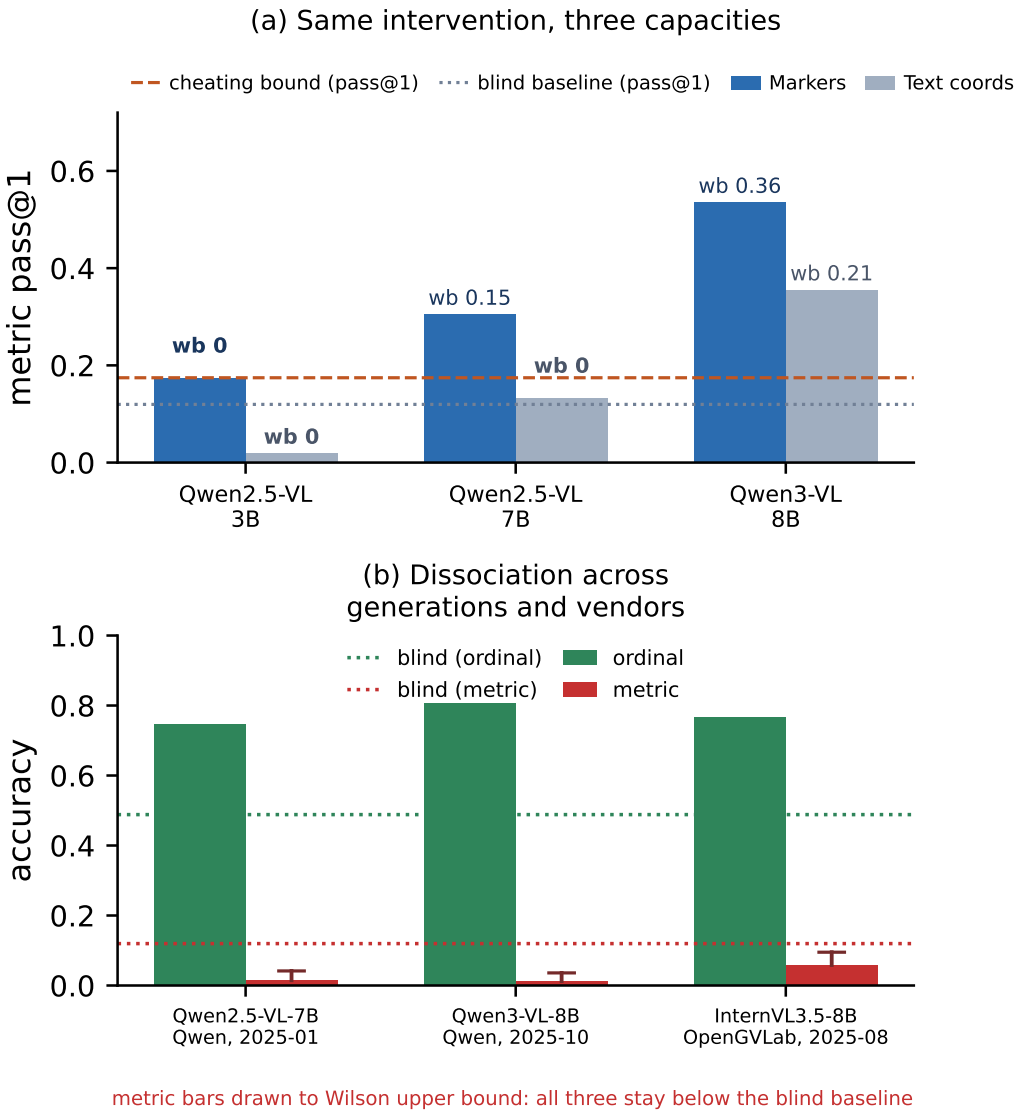


Figure 5: External validity of the two findings this paper rests on. (a) The same two-arm intervention on three backbones. The direction of the formulation effect is positive in all three and comparable in size; what changes is whether the text-coordinate arm reaches a non-zero worst bucket at all. Worst-bucket values (marked wb) are annotated rather than plotted, because the cheating bound and the blind baseline are references for pass@1 and are not comparable to a worst-bucket value. (b) The ordinal/metric dissociation before any training, on two model generations and two vendors (an InternViT rather than a Qwen vision tower). Ordinal accuracy sits far above its blind baseline in all three, while the metric bars — drawn up to their Wilson upper bounds — remain below the metric blind baseline. The newer Qwen backbone advertises improved 3D grounding; that does not change the pattern.

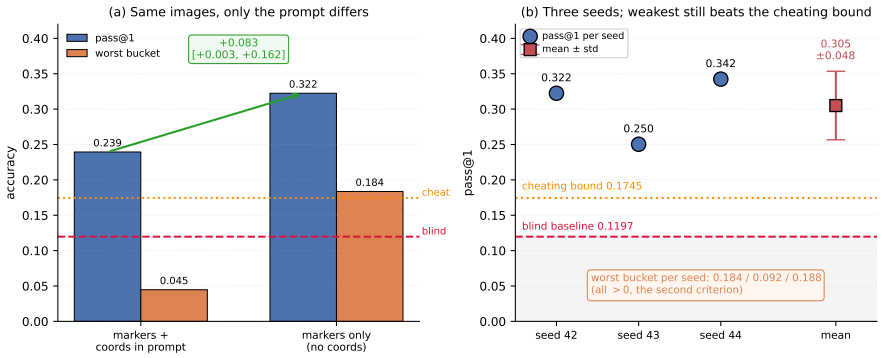


Figure 6: Two reinforcements of the main experiment. (a) The single-variable contrast: both arms train on identical images (markers drawn) for the same number of SFT steps; only the prompt differs. (b) Three seeds of the marker-based arm; every seed clears the blind baseline and the weakest still clears the cheating bound, with a non-zero worst bucket throughout. Worst-bucket values are annotated rather than plotted, because the cheating bound references pass@1 only.

A silent-failure mode worth reporting. Extending the intervention to a newer backbone hit a failure that produced no error and a plausible-looking loss curve: the training script instantiated the model with a hard-coded class for the older architecture, so on the newer checkpoint many weights silently failed to match and were re-initialised. Had we compared the resulting checkpoint without inspecting the loss scale, we would have reported a false negative — "the effect does not replicate on newer models." We now assert that no weight is missing at load time and calibrate the gradient-clipping threshold per backbone; App. M gives the diagnostic numbers. We report this because the failure is invisible in every artefact a reader would normally inspect, and because cross-generation comparisons are exactly where it strikes.

Scope of the shortcut analysis and of the tasks. The mechanism is specific to scenes with a dominant ground plane; it does not extend to aerial, underwater or micro-scale imagery. We enumerate three shortcut families (position prior, answer prior, coordinate leakage) and remove the third by construction; others may exist, and the blind baseline bounds their total strength without identifying them. Metric evaluation covers point-to-camera distance and point-to-point gap; size, volume, velocity and pose are untested.

F Probing the Representation

To ask whether depth is recoverable from the frozen visual representation, we train a small probe on pooled features. An MLP probe attains $\delta < 1.25$ of 0.3475 on the base model against a constant-prediction baseline of 0.1650 — a factor of 2.1 — and reduces AbsRel from 0.796 to 0.543; a linear probe adds only 0.075–0.085, so the information is present but not linearly available. Two comparisons matter. Scored under the same $\delta < 1.25$ criterion as the probes, the model's own behavioural output on point-level absolute depth reaches only 0.1587, i.e. a factor of 2.2 below what a small probe decodes from its own features — the bottleneck is downstream of perception. We quote the $\delta < 1.25$ figure so that probe and behaviour are scored by one criterion; under the paper's default tolerance band the same behavioural number is 0.0586, and the gap persists across the scan (4.3× at $\delta < 1.10$, 1.8× at $\delta < 1.50$). And the probe trained on the spatial-RLVR model scores 0.3525, indistinguishable from the base model's 0.3475: RLVR left the visual representation essentially unchanged and only reshaped output behaviour (Appendix Fig. 17). That RL post-training mainly redistributes existing capability has been argued in the language domain (Yue et al., 2025), and probing analyses of spatial representations appear in Min et al. (2026); Xu (2026); Bhat & Yamasaki (2026); we therefore treat this subsection as confirmatory for spatial RLVR specifically, with the base-vs-RLVR probe comparison as the added evidence.

G The Ordinal Position Prior

We build an adversarial ordinal subset in which the "lower in the image means closer" heuristic is wrong by construction (heuristic accuracy 0.0, chance 0.5). Both models fall significantly below chance — 0.2967 and 0.3350, with Wilson 95% CI upper bounds 0.3344 and 0.3737, both under 0.5 (Appendix Fig. 3). A two-component mixture estimate of how often each model follows the heuristic gives 0.41 and 0.33, with residual real increments of +0.096 and +0.205; the derivation and its assumptions are in App. D. Human judgement confirms the subset is genuinely hard rather than mislabelled: an expert marked 0.2857 of these items "cannot tell" versus 0.1026 elsewhere, and here the ground truth passes multi-frame consistency better than the full sample ($p_{50} = 0.0247$ vs. 0.0606 m).

H The SFT Carrier

Supervised fine-tuning on the same marker-based data reaches 0.3224 with worst-bucket = 0.1836 (from 0.0293 / 0.0000 before training) at roughly one fifth of the compute (63 min vs. 5.5 h). The precise reading is narrow: on this task, RL provides no measurable advantage over supervised training. It does not by itself establish that the formulation matters more than the algorithm, since changing the paradigm moves a different axis than changing the formulation. The formulation claim rests on the single-variable contrast of Section 5; this subsection only rules out "RL was doing the work".

I Same-Protocol Comparison and Training Dynamics

Table 5: Same-protocol comparison. Upper block: all measured by us on the identical de-biased benchmark, so the columns are commensurable and the blind baseline is the reference every row must clear. Lower block: numbers as published by Cai et al. (2025) under their own protocol — listed only to show that the direction of the formulation effect reproduces outside our implementation, not to be compared in magnitude with the upper block. worst-bucket is the accuracy of the worst depth bucket.

	ordinal	metric	worst-bucket
measured here, one benchmark, one verifier			
Blind text-only baseline	0.4883	0.1197	–
Qwen2.5-VL-7B (base)	0.7467	0.0151	0.0000
Spatial-SSRL-7B	0.8267	0.0126	0.0000
Ours, text-coordinate arm	–	0.1324	0.0000
Ours, marker arm	–	0.3050	0.1546
as published, different protocol (direction only)			
Seed1.5-VL, text-coord	—	0.243	—
Seed1.5-VL, marker	—	0.400	—

Table 5 puts every method we measured on one benchmark with one verifier, so the blind baseline is a line each row must clear rather than a number quoted from elsewhere. Two things are visible. A spatial-RLVR model improves the ordinal column by a real margin (0.7467 → 0.8267) while the metric column stays below the blind baseline, which is the dissociation the paper is about. And our marker arm is the only row whose worst bucket is non-zero.

J Reproducibility

Data. All queries are generated from nuScenes keyframes with LiDAR depth projected into the camera using the shipped intrinsics and extrinsics, followed by a visibility check that drops occluded points and points outside the frustum. Split sizes: 3000 marker-based training queries, 8000 text-coordinate training queries, 1200 marker-based validation queries (of which 241 are metric-depth and 259 ordinal after family filtering), 2000 text-coordinate validation queries, and two adversarial subsets of 600 each. The metric benchmark is additionally rebuilt at 240 examples per bucket (1200 total) for the sharpened worst-bucket test.

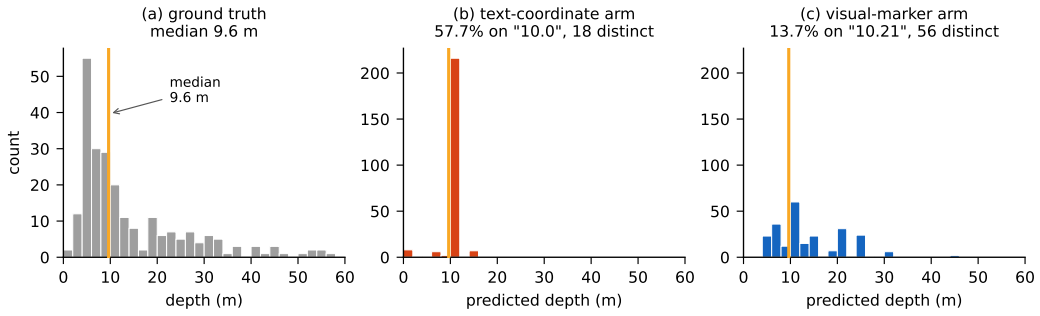


Figure 7: What the text-coordinate arm actually learned. All three panels are the same 241 metric queries. (a) ground-truth depths, median 9.6 m. (b) in the seed plotted here the text-coordinate arm emits one value for 57.7% of queries and only 18 distinct values in total, and that value sits on the distribution's median — the risk-minimising constant, which requires no image. (c) the marker arm spreads over 56 distinct values and tracks the distribution. Averaged over seeds the collapse rate is 51.2% against 14.5% for the marker arm. This is the same mechanism the worst-bucket zero reports, made visible.

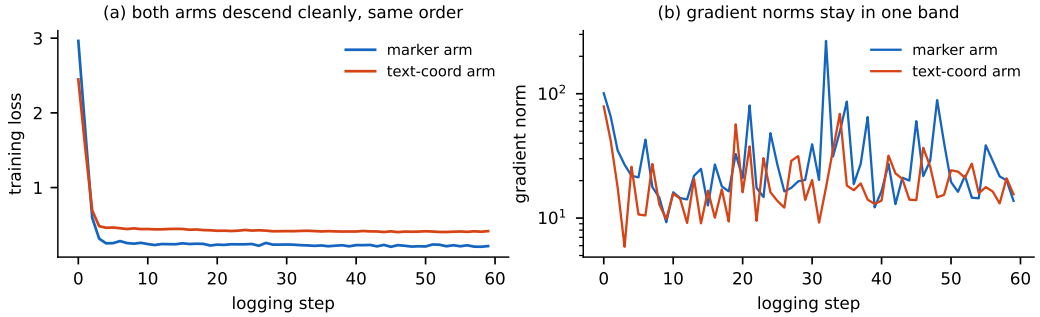


Figure 8: The difference is not an optimisation artefact. (a) both arms descend from the same starting order to the same final order (2.9633 → 0.2166 and 2.4480 → 0.4167; the two arms train on different data, so the achievable loss is not expected to match exactly). (b) gradient norms stay within one band (medians 21.4375 and 16.6875). Neither run shows the signature of a broken one; what differs is the ability the optimisation arrives at. A silently broken run looks nothing like this — see App. M.

Reward. A tolerance-band verifier: a scalar answer counts as correct when $|\hat{z} - z| \leq \max(\rho|z|, \alpha)$ with $\rho = 0.12$ and $\alpha = 0.15$ m; ordinal answers are exact-match on the letter. The reward is the correctness term with the format term weighted to zero for evaluation. The same verifier function serves as the RLVR reward and as the evaluation metric, so a training gain cannot come from a mismatch between the two.

Training. See Tab. 6. The RLVR arm uses GRPO advantages with a GSPO-style sequence-level importance ratio and DAPO-style filtering, KL disabled, 8 rollouts per prompt at temperature 1.0, and FSDP full sharding across 8 H20 GPUs; validation decoding uses temperature 0.6 with a single sample. The SFT arm computes loss on answer tokens only, masking the prompt. One image per prompt is enforced throughout.

What we release. The task suite (generators and the seven-item auditing protocol as a single entry point), the verifier, the evaluation and probing scripts, and the blind-baseline fits whose measured values are tabulated per dataset. Every number in the paper carries an internal identifier tying it to the run that produced it.

Known pitfalls when porting to another backbone. Three of them produce a checkpoint that trains without error while learning nothing useful, so we state them explicitly. (i) Instantiating the model with a hard-coded class for a different architecture silently re-initialises unmatched weights; assert that no weight is missing at load time. (ii) The gradient-clipping threshold must be calibrated per backbone — a value that is mild for one model compresses the effective step size by orders of magnitude for another. (iii) Verify that the image channel

Table 6: Training configuration. Both arms use identical settings; the only difference is the training data's pixel reference. The RLVR column is read from configs/config_geo_resolved.json, which is the configuration the trainer resolved at the start of the run and printed to its own log; the base file configs/config_geo.yaml is also shipped, but several keys are overridden at launch and it therefore does not record what ran. The SFT column is read from the defaults in src/train_sft.py. Neither column is transcribed by hand. The step counts and the hardware and wall-clock rows come from the launch environment.

	RLVR (GRPO)	SFT
learning rate	1×10^{-6}	1×10^{-5}
schedule	constant	cosine, warmup .05
optimiser	AdamW	AdamW
global batch	32	64 (1×8×8)
rollouts / prompt	8	—
rollout temp.	1.0	—
KL penalty	disabled	—
max prompt/resp.	2560/640	2560
grad-norm clip	1.0	1.0
precision	bf16, FSDP	bf16, ckpt
steps	100	300
hardware	8 × H20 96 GB	
wall-clock	5.5 h	63 min

participates in the loss at all: replacing the image with noise must raise the loss. A model whose vision path is silently disconnected otherwise looks like a negative result about the method.

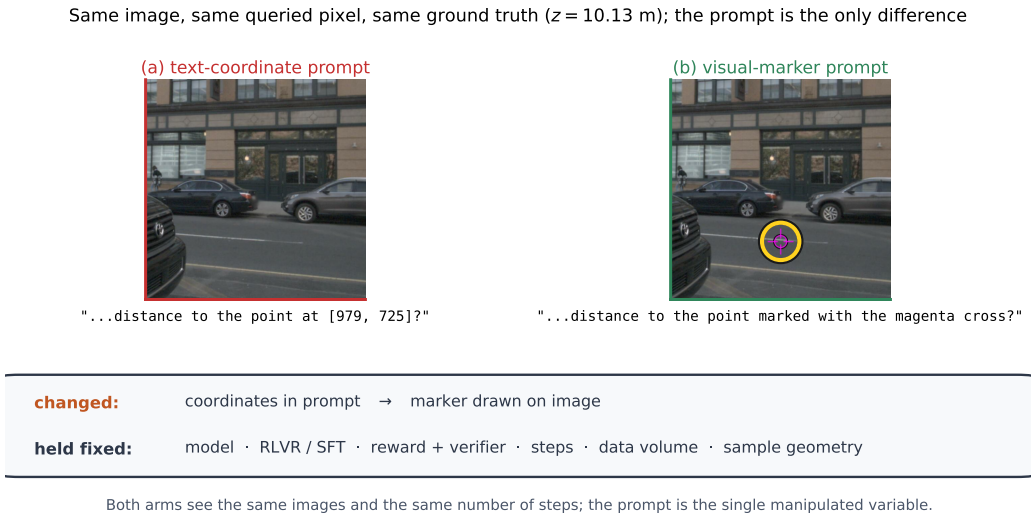


Figure 9: The manipulated variable and the design around it. Top: the same nuScenes image, the same queried pixel and the same LiDAR ground truth under the two prompts. With coordinates in the prompt the row v already predicts the answer; with a marker drawn instead, the location has to be recovered visually. Bottom: everything else is held fixed and both arms are compared under one anti-gaming protocol. The yellow ring is added here for legibility only; the model's input contains the magenta cross alone, which is what the prompt names.

K Additional Figures

Three figures referenced from the main text: the geometry that makes the text-coordinate shortcut work, qualitative examples of the marker-trained arm on real queries, and examples from the off-ground adversarial subset on which that shortcut is void by construction.

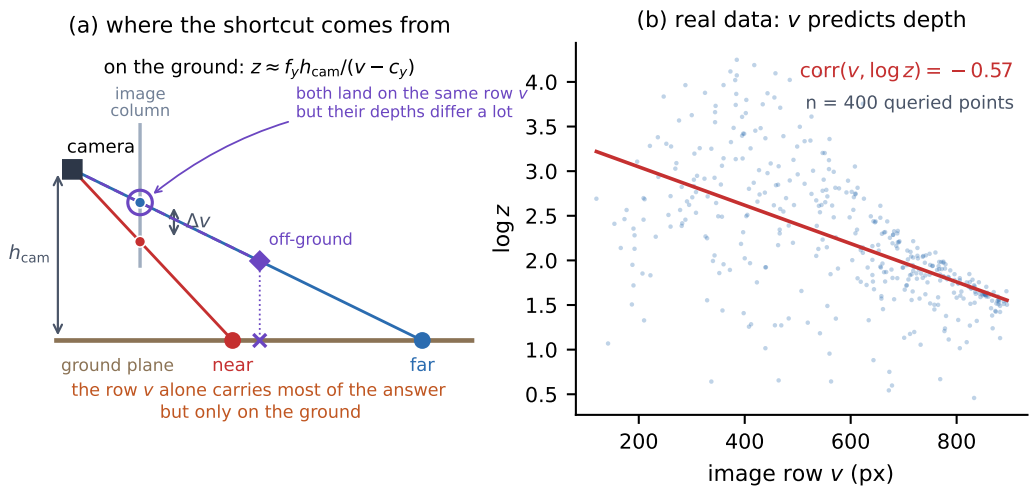


Figure 10: Where the text-coordinate shortcut comes from. (a) For a point on the ground plane, depth follows $z \approx f_y h_{\text{cam}}/(v - c_y)$, so the image row v is informative about the answer before any perception happens. The purple point shows why the adversarial subset defeats this: it lies off the ground and therefore shares an image row — one camera ray — with a much more distant ground point, so v no longer determines z. (b) The same relation in the data we query: $\text{corr}(v, \log z) = -0.57$ over the queried points of the driving validation split.

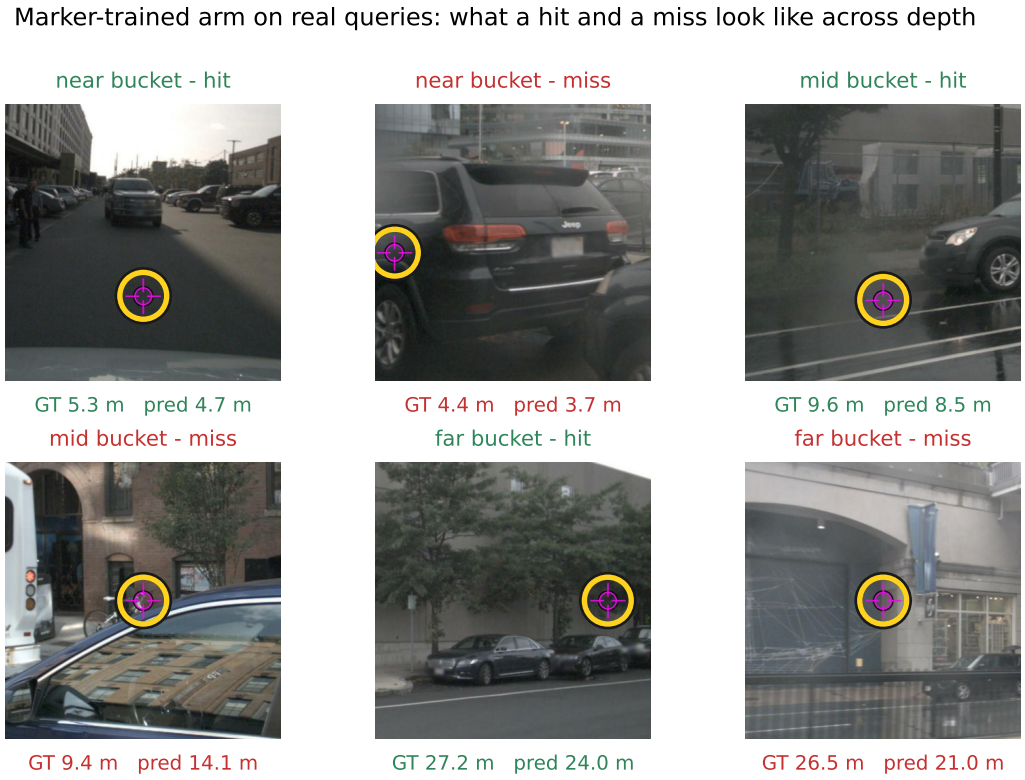


Figure 11: Marker-trained arm on real queries. Two examples from each of three depth ranges, one inside the tolerance band and one outside. Ground truth and prediction are printed under each crop. Misses are not concentrated in any one scene type; they are spread across ranges, which is what the worst-bucket criterion is designed to expose. The yellow ring is added here for legibility only; the model's input contains the magenta cross alone, which is what the prompt names.

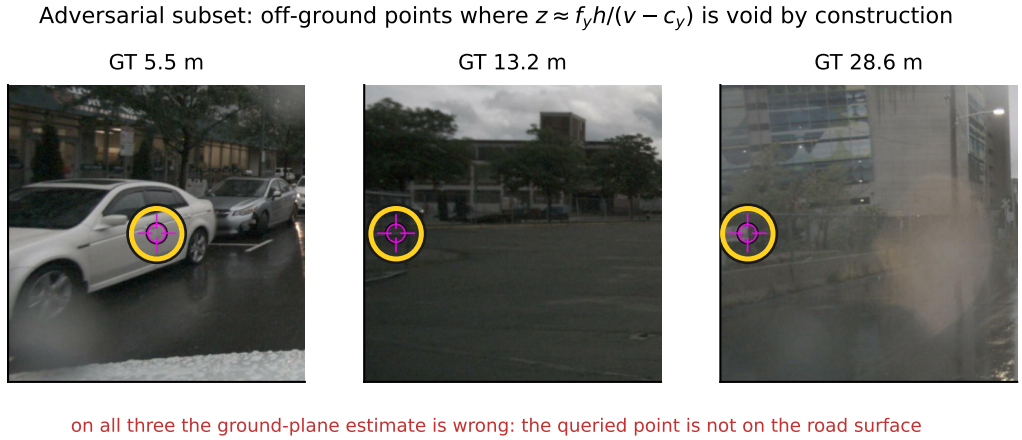


Figure 12: The off-ground adversarial subset. Queried points that do not lie on the road surface, so the ground-plane relation used by the text-coordinate shortcut is void by construction. This is the subset on which the marker arm's CI lower bound exceeds both the blind baseline and the cheating bound. The yellow ring is added here for legibility only; the model's input contains the magenta cross alone, which is what the prompt names.

L The Version Ceiling We Measured

Two independent constraints fixed which models we could run, and we state them because they determine the scope of our claims rather than merely inconveniencing us. First, the cluster driver is 535.261.03, which supports CUDA at most 12.2, whereas inference-engine releases from 0.29 onwards require torch built against CUDA 13 and hence a driver of at least 580. Second, the container ships Python 3.11, and we verified version by version that releases 0.12.0 through 0.20.2 publish no wheel for it, whereas 0.11.2 does. The effective ceiling is therefore 0.11.2. Under it, Qwen2.5-VL, Qwen3-VL and InternVL3_5 are all supported, while Qwen3.5 has no implementation at all. For Qwen3.5 we fall back to the HuggingFace transformers path, whose only hard dependencies are pure-Python packages and which therefore installs against the container's existing CUDA-12 build; it is five to ten times slower than the engine and so is usable for evaluation but not as a training rollout backend.

M Diagnosing a Silently Re-Initialised Model

The failure described in Sec. 8 presented as follows. Under an otherwise identical configuration and the same 300 steps, the older backbone went from loss 2.9633 to 0.2166 with gradient norms between 10 and 60, whereas the newer one went from 9.4306 to 2.0164 with gradient norms between 6.2×10^4 and 3.5×10^5 (median 1.17×10^5). An initial loss of 9.43 is close to the $\ln |V| \approx 11.9$ of uniform guessing. Because the default clipping threshold is 1.0, a gradient norm of 1.17×10^5 scales every step by roughly 8.5×10^{-6}, compressing the effective step size by five orders of magnitude. The cause was that the script requested Qwen2_5_VLForConditionalGeneration while the checkpoint declares Qwen3VLForConditionalGeneration. We ruled out three other candidates before finding it: both models use the same end-of-turn token, their generation prompts are byte-identical, and neither chat template contains thinking tags, so the label mask — which counts back from the end of the sequence — was correctly placed. After resolving the class from config.architectures, the newer backbone trains from 2.9245 to 0.2753 with a median gradient norm of 3.25, matching the older one.

N Two Supporting Results

A verifiable reward that needs no depth annotation. For two queried pixels with rays fixed by the intrinsics, the law of cosines $d^2 = r_A^2 + r_B^2 - 2 r_A r_B \cos\theta$ constrains the triple (r_A, r_B, d) without ground truth, because θ follows from pixel coordinates and intrinsics alone. Substituting ground truth gives residual $p_{90} = 0.00112$ with 100% of samples above

0.9. The construction is invariant to global scaling (multiplying all three by 1.2 still yields 0.9991), which we repair with a weak ground-plane anchor. At 80 steps the term raises the overall reward from 0.1898 to 0.3790 but leaves accuracy unchanged (0.00317 vs. 0.00318), so we claim only the constructive result — verifiability without annotation — and not a training gain. Derivation and the combined-reward ablation are in App. A.

Ground truth and human audit. Three independent geometric checks pass: multi-frame reprojection consistency (median deviation 0.0606 m), point-in-box depth agreement (0.9221 inside the expected range, with the bias 87.6% negative and correlated −0.5564 with half-extent — the signature of LiDAR seeing only the facing surface), and zero triangle-inequality violations in 250 triples. An expert annotator judged 60 ordinal items rendered with enlarged markers: among answerable items in the three non-adversarial families, 34/35 were correct, a pass rate of 0.9714 (Wilson 95% CI $[0.8547, 0.9949]$). The point estimate clears our 0.90 criterion but the CI lower bound does not; we report that rather than claim the criterion met. For comparison, GRAID (Elmaaroufi et al., 2025) reports 0.576 for mainstream pipelines. Details in App. A.

O A Ground-Truth-Free Verifiable Reward

For two queried pixels with rays $\hat{r}_A, \hat{r}_B$ determined by the intrinsics, the law of cosines

$$d^2 = r_A^2 + r_B^2 - 2 r_A r_B \cos\theta, \qquad \cos\theta = \hat{r}_A \cdot \hat{r}_B \tag{5}$$

constrains the triple (r_A, r_B, d) without any depth annotation, because θ follows from pixel coordinates and intrinsics alone. Substituting ground truth gives residual $p_{50} = 0.00029$, $p_{90} = 0.00112$, and 100% of samples above 0.9.

An inherent limitation, and the fix. Eq. 5 is invariant to global scaling: multiplying all three quantities by 1.2 still yields self-consistency 0.9991. We therefore add a weak ground-plane scale anchor; the combined reward (0.7 self-consistency + 0.3 anchor) gives 0.9347 on ground truth, 0.7894 when all quantities are doubled (the anchor term catching the scale error), and 0.3676 when only d is doubled (the self-consistency term catching the geometric error).

Honest scope. At 80 steps the self-consistency term raises the overall reward from 0.1898 to 0.3790 but leaves accuracy essentially unchanged (0.00317 vs. 0.00318). We therefore claim only the constructive result — verifiability without ground truth — and not a training gain.

P Ground-Truth Validation and Human Audit

Three automatic checks. Multi-frame reprojection consistency gives median world-coordinate deviation 0.0606 m; point-in-box depth agreement puts 0.9221 of points inside the expected range, with the bias being 87.6% negative and correlated −0.5564 with half-extent — the geometric signature of LiDAR seeing only the facing surface; the triangle inequality is violated in 0/250 cases; and the in-box point count correlates 1.0000 with the official num_lidar_pts.

Human audit. An expert annotator judged 60 ordinal items rendered with enlarged markers. Among answerable items in the three non-adversarial families, 34/35 were correct, a pass rate of 0.9714 with Wilson 95% CI $[0.8547, 0.9949]$. The point estimate clears our 0.90 criterion but the CI lower bound does not; we report this rather than claim the criterion is met. For comparison, GRAID (Elmaaroufi et al., 2025) reports 0.576 for mainstream pipelines.

The adversarial set is genuinely harder — for humans too. See Section 6: the "cannot tell" rate is 2.8× higher on the adversarial family, and the items judged incorrect are not ground-truth errors. We do not repeat those numbers here.

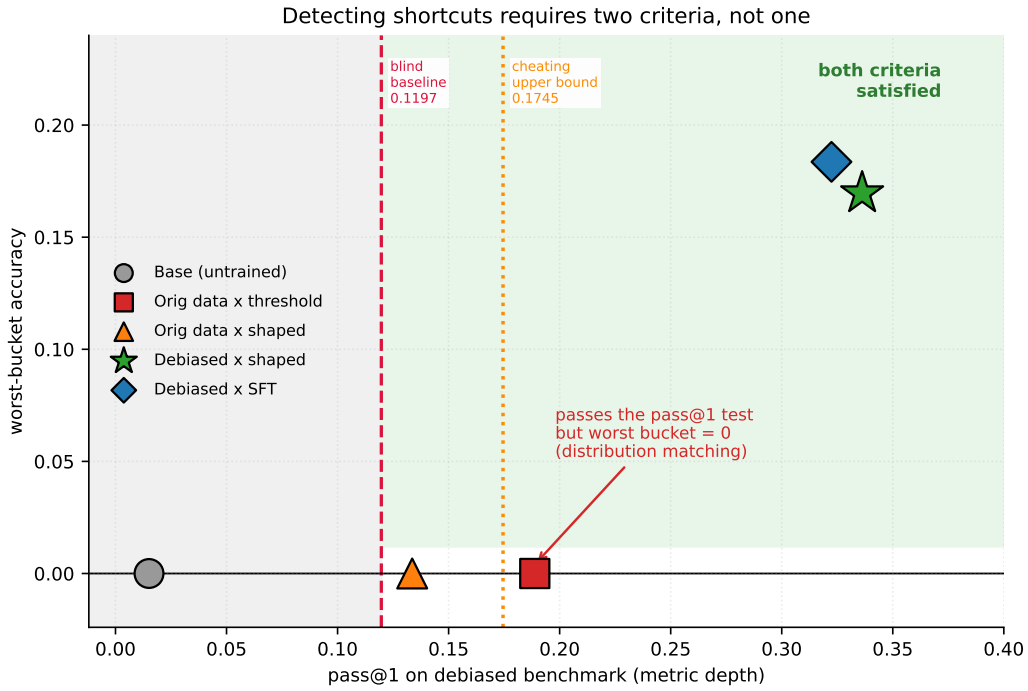


Figure 13: Detecting shortcuts requires two criteria. The text-coordinate threshold arm (red square) lies to the right of the cheating upper bound on the x-axis, yet sits exactly on $y = 0$ — outside the green region where both criteria of Eq. 1 hold.

Table 7: A text-only model, never shown the image, on text-coordinate benchmarks. L2 is a learned model on prompt features; it is the score obtainable with zero visual information.

Task family	driving	indoor
absolute depth (metric)	0.2951	0.2355
depth ordering	0.8550	0.8000
height ordering	0.9050	0.9820
pairwise gap (metric)	0.1577	0.2553
same benchmark, markers, no coords	0.1197	0.2101

Table 8: Three seeds, same conclusion. Marker-based SFT arm, 300 steps, evaluated on the same benchmark. Every seed satisfies both criteria of Eq. 1, and even the weakest exceeds the cheating bound.

Seed	pass@1	95% CI	worst-bucket
42	0.3224	[0.268, 0.385]	0.1836
43	0.2503	[0.199, 0.307]	0.0921
44	0.3424	[0.287, 0.406]	0.1880
mean ± std	0.3050±0.0484	—	0.1546±0.0541

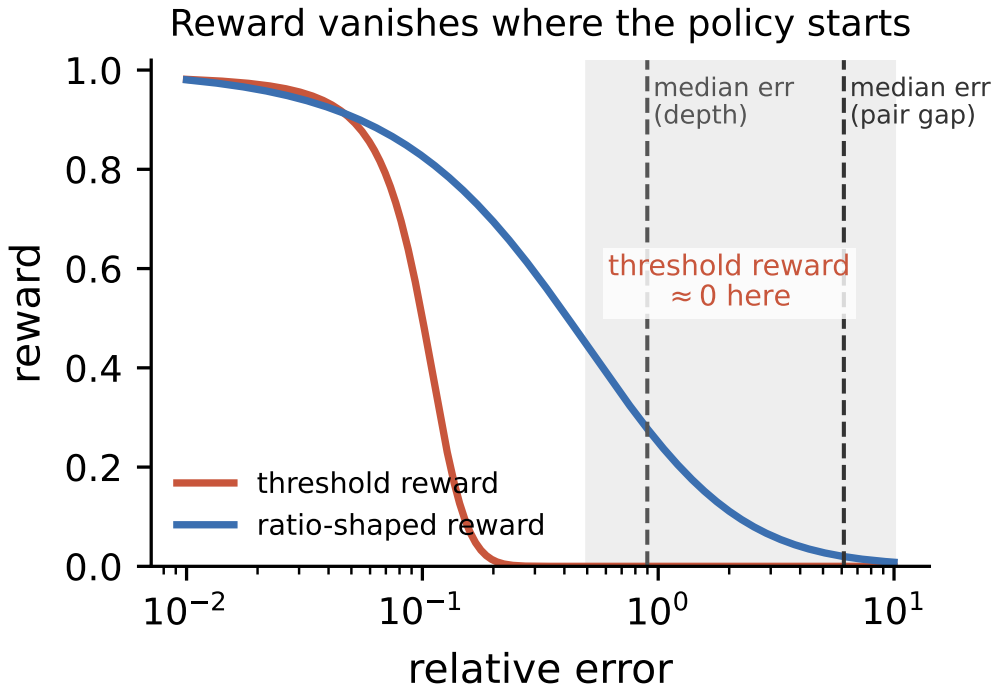


Figure 14: Why a threshold reward cannot start learning on continuous targets. With a tolerance band of $\pm 2\rho$ and an initial median error far outside it, almost every rollout receives the same zero reward, so the group's advantages vanish. A ratio-shaped reward retains a gradient outside the band. Note that shaping alone is not sufficient: Section 4 shows a shaped-reward arm still fails when the formulation leaks a shortcut.

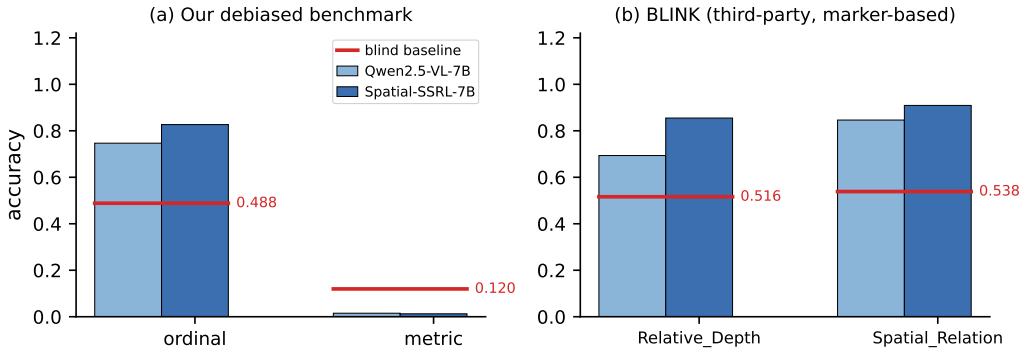


Figure 15: Double dissociation. On the same marker-based benchmark, ordinal accuracy is far above the blind baseline for both models, while metric accuracy is below it for both.

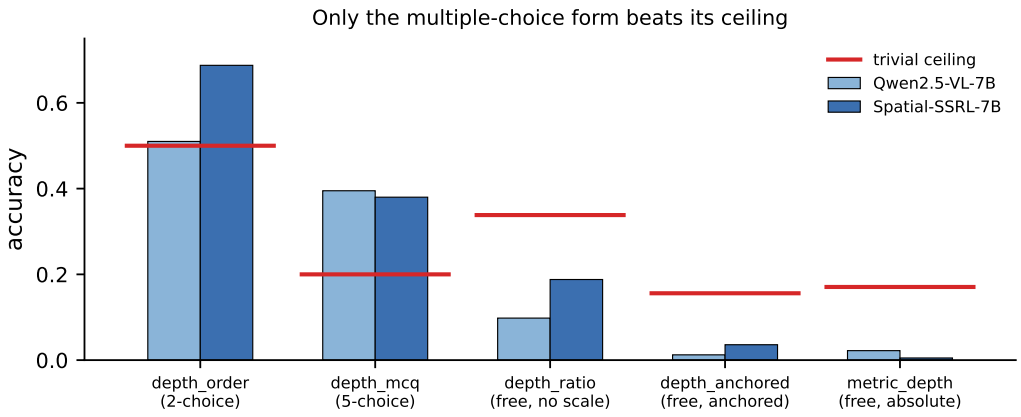


Figure 16: The deficit is in numeric emission, not perception. Each bar is normalised against the format's own chance level and cheating bound. Multiple choice clears chance by roughly 2×; every free-numeric format stays under its own ceiling.

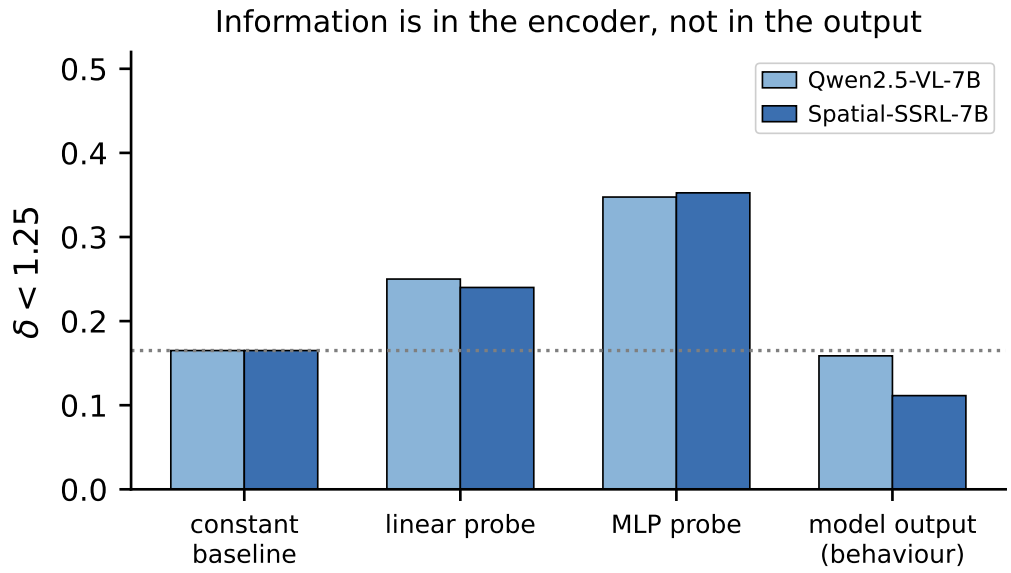


Figure 17: RLVR reshapes behaviour, not representation. Probe accuracy on the base and spatial-RLVR models is nearly identical, while both far exceed the models' own behavioural output. All four bars use the same $\delta < 1.25$ criterion, so the behavioural bar is directly comparable to the probe bars.