DECODABLE IS NOT GROUNDED: A GROUNDING ARBITER FOR VLM SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial representations in vision–language models (VLMs) can support effective interventions while the resulting answer gains persist without the image. Across fourteen VLMs on ViewSpatial-Bench, thirteen score below chance on depth and twelve perform better on horizontal relations than depth. Matched-protocol interventions in three VLMs reveal a shared reversed depth response: steering toward the positive relation pole lowers its answer probability, while reverse-sign oracle injection improves accuracy by 13.7–22.2 percentage points. Vertical direction removal instead has model-dependent effects. We introduce the *Grounding Arbiter* to examine what successful edits recover: compare matched, blank, and mismatched images, then repeat the same intervention with matched and blank inputs. In a detailed case study, vertical relations reach 86.9% probe accuracy under scene-grouped nested evaluation, yet an amplification operating point selected from an exploratory sweep gains 20.5 points with real images and 19.1 with blank images. We call this an *ablation-persistent intervention gain*. Further analyses show that decoding and controllability can peak at different layers, and cross-model task comparisons reveal large image benefits on other depth benchmarks. These findings connect the evaluation of spatial representations to two practical choices: where to intervene, and how to measure the resulting gain's dependence on visual evidence.

1 INTRODUCTION

A linear probe can recover information that a model fails to express in its answers (Gurnee & Tegmark, 2024; Belinkov, 2022). Activation steering then offers a way to change those answers by editing the representation (Rimsky et al., 2024; Zou et al., 2023). For vision–language models (VLMs), this sequence raises a further question: how much of the improvement depends on the image? An edit may improve the use of visual evidence, or amplify predictive information already available in the question. Measuring that dependence helps explain what a spatial intervention actually fixes.

Across fourteen VLMs, depth reasoning on ViewSpatial-Bench is markedly weaker than horizontal reasoning. Targeted interventions and image controls connect this broad pattern to a separation between decodable information, control over answers, and dependence on visual evidence.

We introduce the *Grounding Arbiter* to connect intervention accuracy gains to image dependence. It compares the same questions under matched, blank, and mismatched images, then repeats the key intervention with matched and blank inputs. This extends the question-prior controls used in VQA (Goyal et al., 2017; Agrawal et al., 2018) to representation interventions. Repeating the edit after image removal measures how much of its benefit survives without the matched visual input.

Three findings organize the paper. First, *depth responses recur across models, while vertical control varies*. Thirteen of fourteen VLMs have observed depth accuracy below 50%, and twelve perform better on horizontal relations. In Qwen2.5-VL-7B, Pixtral-12B, and InternVL3-8B, steering toward the positive depth pole lowers its answer probability. Reverse-sign oracle injection improves accuracy by 13.7–22.2 points. Vertical direction removal produces a marked drop in Qwen and little

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.

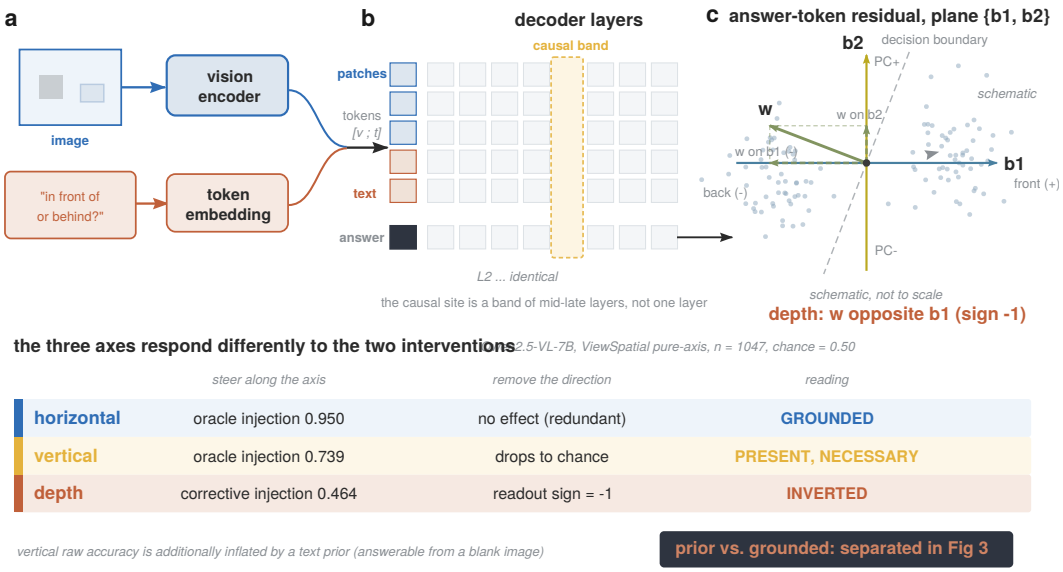


Figure 1: **Axis-specific behavior motivates paired intervention and image controls.** (a) Camera-relative VQA input. (b) Intervention sites in mid-to-late decoder layers. (c) Axis-direction and readout geometry in the b_1, b_2 plane. The original Qwen summary shows different removal responses across axes; Fig. 3 compares image replacements. Updated grouped probes and three-model interventions appear in Tables 1 and 2. $n = 1047$; binary chance $= 50\%$.

change in the other two models (§4.2). Paired tests on What'sUp and 3DSRBench reveal large matched-image benefits on other depth tasks (§5).

Second, *most of the intervention-induced accuracy gain survives image removal*. Scene-grouped nested probing recovers vertical relations at 86.9% accuracy in the detailed case study, making this axis a useful test of accessible information and weak answers. At $\gamma = 2$ in the exploratory amplification sweep, vertical accuracy gains 20.5 points with real images and 19.1 with blank images. We call this an *ablation-persistent intervention gain*, exposing an *intervention–grounding dissociation* (§4.3).

Third, *a shared reversed depth response calls for model-specific operators and edit sites*. Rotation gains 55.7 points in Qwen3-VL-8B, but the best tested rotations gain only 15.5 and 2.4 in Qwen2.5-VL-7B and Pixtral; trained edits gain 42.9 and 32.5 in the latter two. In Qwen3-VL-8B, controllability peaks at L24 and probe accuracy at L26. These peaks differ in three of seven models (§6). Operator and layer selection therefore require measuring behavioral response alongside decodability.

These findings guide two practical choices: behavioral layer profiles help select where to edit, and the Grounding Arbiter measures the resulting gain's dependence on visual evidence.

2 Related Work

Spatial reasoning benchmarks. What'sUp (Kamath et al., 2023) and VSR (Liu et al., 2023) test object relations in images. SpatialBot (Cai et al., 2025) and EmbSpatial-Bench (Du et al., 2024) extend evaluation to spatial understanding and embodied tasks. More recent benchmarks cover video, multiple images, and perspective changes: VSI-Bench (Yang et al., 2025), MMSI-Bench (Yang et al., 2026), OmniSpatial (Jia et al., 2026), and ViewSpatial-Bench (Li et al., 2026). Training-based approaches improve these capabilities using spatial supervision and reinforcement learning (Chen et al., 2024; Liu et al., 2026a; Daxberger et al., 2025). Our main analysis instead asks what existing models' scores and representations establish about their use of images.

Spatial representations and their use. Gurnee and Tegmark (Gurnee & Tegmark, 2024) recover spatial and temporal structure from language-model representations. Tehenan et al. (2025) study linear spatial world models, including causal interventions; Kang et al. (2026) identify spatial IDs that mediate VLM predictions through language tokens, using mirrored-input activation swaps, targeted

steering, and image-masking diagnosis. Their layer analyses trace spatial information across token types. Our paired evaluation follows the same edit across matched and blank inputs, measuring how much of its behavioral gain remains after image removal.

Several studies locate specific obstacles to spatial reasoning. Wang & Gao (2026) isolate latent scene topology and improve its use through regularization. Qi et al. (2025) examine how token-norm imbalance suppresses positional cues. Savietto et al. (2026) relate concept-vector geometry to visual errors and test these directions with steering. Min et al. (2026) identify vertical–distance entanglement and introduce SpatialTunnel to remove natural-image correlations. These works characterize what spatial information is represented and how it can be manipulated. We instead measure whether an intervention benefit survives replacement of the image that should supply the information. A structured comparison is provided in Appendix B.

Probing, steering, and input controls. Control tasks and methodological analyses have established that probe accuracy alone does not demonstrate use of a representation (Hewitt & Liang, 2019; Belinkov, 2022). Work on model editing and circuit analysis provides methods for testing causal effects (Meng et al., 2022; Wang et al., 2023), while contrastive activation addition and representation engineering change behavior through latent directions (Rimsky et al., 2024; Zou et al., 2023). Basu et al. (2026) find that highly accurate probes need not yield reliable error correction; Liu et al. (2026b) examine a disconnect between visual attention and answer correctness. Our focus is complementary: whether a successful intervention remains successful after the relevant image is removed. This connects representation analysis to the question-prior controls used in VQA (Goyal et al., 2017; Agrawal et al., 2018).

3 SETUP

Tasks and scoring. The three benchmarks serve complementary roles. ViewSpatial-Bench (Li et al., 2026) provides the primary mechanistic task: its camera-relative direction questions support a common horizontal/vertical/depth decomposition for probing, intervention, and image controls. We use its 1773 camera-perspective relative-direction questions. Each native question offers four choices drawn from 18 directional labels, giving 25% chance. We group questions by the physical axis of the gold answer: horizontal, vertical, depth, or composite. The primary analyses use the 1047 pure-axis items (Table 1). The binary protocol replaces the native choices with the two poles of one axis, with option order randomized per item. Raw accuracy weights questions equally; balanced accuracy averages the two pole-specific recalls, with 50% chance.

What'sUp (Kamath et al., 2023) tests whether the patterns persist in object-relation tasks expressed as left/right, on/under, and front/behind. 3DSRBench (Ma et al., 2025) tests height and closer-to-camera relations, providing another formulation of vertical and depth reasoning. Matched, blank, and mismatched-image comparisons on these tasks locate the benchmark boundary of the ViewSpatial depth inversion (Table 11).

Probing. A logistic probe predicts the within-axis pole from residual activations, using standardization and PCA before classification. The primary estimate uses five outer folds grouped by source scene, with layer and token-locus selection nested within the outer training partition. Shuffled-label scores accompany each axis. The selected vertical locus is text-pooled in all five outer folds. Appendix A distinguishes this evaluation from the earlier question-level layer sweeps.

Image ablation. For axis a, let A_a^{real} and A_a^{blank} denote balanced accuracy with a matched image and a same-size gray blank. With questions, options, and decoding fixed, we measure

$$\Delta_a = A_a^{\text{real}} - A_a^{\text{blank}}, \qquad L_a = A_a^{\text{blank}} - \tfrac{1}{2}. \tag{1}$$

Δ_a measures whether image content helps or harms accuracy, and L_a measures blank performance above chance. We call below-chance real accuracy with negative Δ_a an *behavioral inversion pattern*. Small gaps with positive L_a identify performance that largely survives image removal. Additional controls use black, noise, patch-scrambled, and unrelated same-axis images. The mismatched-image control substitutes a real image from another question to test whether accuracy depends on the correct image–question correspondence.

Interventions. Fixed-direction steering adds an axis direction, $h^{(\ell)} \leftarrow h^{(\ell)} + \alpha \hat{d}_a^{(\ell)}$; directional ablation removes it, $h^{(\ell)} \leftarrow h^{(\ell)} - (h^{(\ell)\top} \hat{d}_a^{(\ell)}) \hat{d}_a^{(\ell)}$. We compare these edits in Qwen2.5-VL-7B,

Pixtral-12B, and InternVL3-8B under a common protocol. For each model, $\hat{d}$ is oriented toward the positive pole using labelled calibration examples. Increasing α tests the sign of the behavioral response; a corrective oracle uses the item's gold pole and reverses the injection sign when the response is negative. Random-direction removal provides a control. The comparison uses the same evaluation procedure, seed 0, and full axis sets across the three models. Each gain is computed against its own run's unedited baseline. A separate activation-amplification experiment uses selected neuron activations to scale a write along a relation direction at MLP down-projection outputs. The multiplier γ controls this write (Appendix A); $\gamma = 2$ is selected from the evaluation sweep. These edits leave model parameters fixed; the correction comparison also includes trained low-rank edits.

Study design and coverage. The fourteen-model behavioral comparison establishes the scope of the axis contrast. The three-model intervention study tests whether the directional responses recur under a common protocol. Grouped probing and paired amplification provide a detailed analysis on Qwen2.5-VL-7B. Four models have the complete paired depth comparison across What'sUp-B and 3DSRBench in Table 4; Qwen also supplies the horizontal and vertical task controls. Within-family size ladders examine the behavioral contrast across model sizes. Eight-model correction and seven-model layer analyses examine where and how depth answers can be changed. Appendix F defines the measurement terms; earlier eight-model probe summaries appear in Appendix C.

4 Intervention Gains That Survive Image Removal

Vertical relations provide a clear case of accessible information, effective intervention, and gain that persists without the image. We connect the scene-grouped probe result to directional control, then compare the amplification gain under matched and blank inputs.

4.1 Vertical information remains decodable despite weak answers

The scene-grouped nested evaluation gives vertical probe accuracy of 86.9%, compared with 61.4% under shuffled labels (Table 1). The probe selects text-pooled activations in all five outer folds. Vertical relation information thus remains accessible when training and test folds contain different scenes. The corresponding behavioral summary reports 58.2% binary answer accuracy, making vertical the strongest behavior–probe contrast under the updated protocol.

Table 1: Behavior and scene-grouped nested probe accuracy on Qwen2.5-VL-7B (%). Behavior columns retain the capability run; probe and shuffled-label scores use the updated grouped protocol. Native questions have four choices; binary questions compare two poles. Intervention baselines are reported within their respective experiments.

Axis	n	Native	Binary	Grouped probe	Shuffle
Horizontal	518	74.9	81.3	49.1	52.7
Vertical	376	26.9	58.2	86.9	61.4
Depth	153	16.3	30.7	57.1	48.9

The other axes show different patterns. Depth probe accuracy is 57.1%, with a shuffled score of 48.9%, while binary answers score 30.7%. Horizontal answers reach 81.3%, but the grouped probe reaches 49.1% against a shuffled score of 52.7%. High answer accuracy and high probe accuracy therefore give different axis rankings. For horizontal relations, this probe setup provides little predictive signal despite strong behavioral performance. The earlier item-level layer sweeps are presented separately in Appendix B.

Native four-choice accuracy is 45.4% on all 1773 questions. Depth also remains below binary chance under two camera-relative phrasings (Appendix Table 6). The next experiment asks how relation directions affect answers, before testing the image dependence of an effective edit.

4.2 Depth response replicates while vertical removal effects vary

Three VLMs show the same reversed depth response under the matched protocol (Table 2). As steering moves toward the positive depth pole, its answer probability decreases. Reverse-sign, gold-conditioned oracle injection raises accuracy by 13.7 points in Qwen2.5-VL-7B, 15.0 in Pixtral-12B, and 22.2 in InternVL3-8B. The common response sign thus accompanies a consistent benefit from

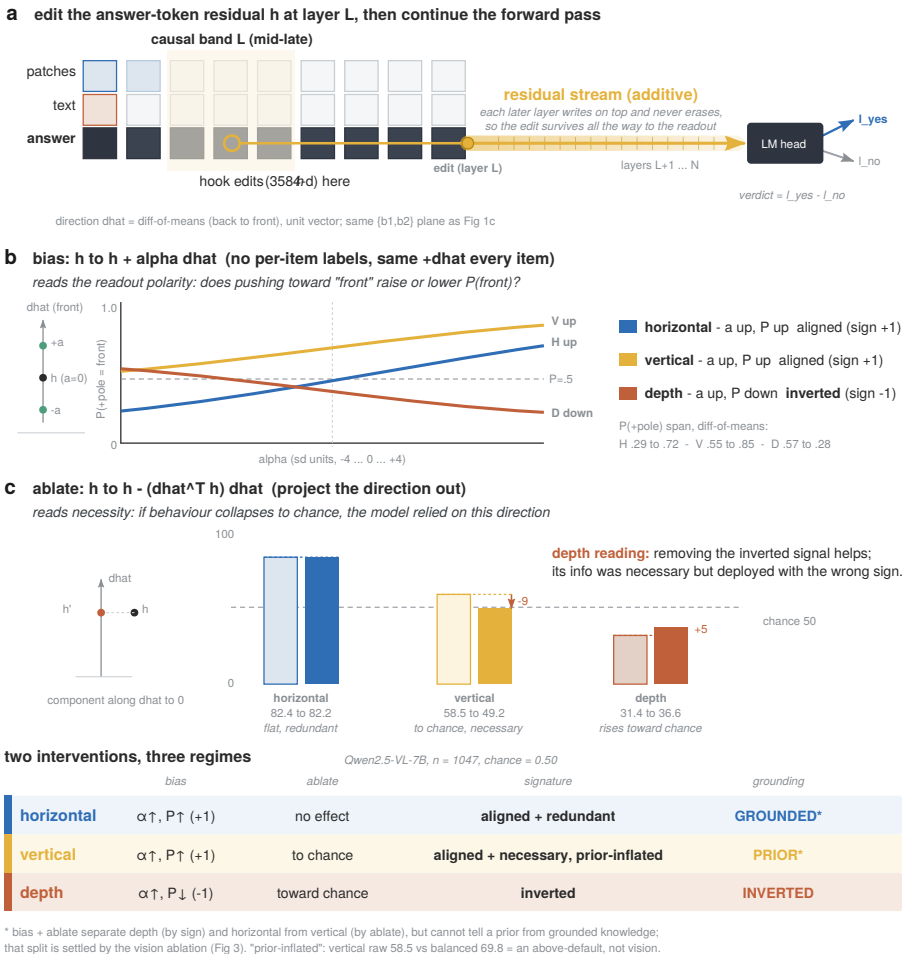


Figure 2: **Axis-specific responses to residual-stream interventions.** (a) Edit site at the answer-token residual. (b) Adding $+\hat{d}$ moves horizontal and vertical responses with the direction, and depth against it. (c) Removing $\hat{d}$ leaves horizontal accuracy nearly unchanged, lowers vertical toward chance, and raises depth toward chance. Original Qwen artwork and case-study labels are retained; Table 2 reports the updated three-model run. Chance $= 50\%$.

reversing the injection. Figure 2 illustrates the edits in the original Qwen study; the three-model summary is visualized in Appendix Fig. 9.

Vertical removal has model-dependent effects: its accuracy relative to random-direction removal is -9.0 points in Qwen, $+0.2$ in Pixtral, and -0.4 in InternVL3-8B. Qwen falls from 58.5% to 49.7%, whereas the other two models change little. Horizontal removal-minus-random gaps are -0.9, -3.1, and $+0.2$ points, respectively. Full vertical scores appear in Appendix B.3.

Table 2: Reversed depth response in three VLMs under the seed-0 S3 protocol ($n = 153$). Positive-pole answer probabilities decrease across the steering sweep; corrective-oracle injection uses gold poles. Accuracy and gains are reported as percentages and percentage points.

Model	Baseline	$P(+)$ sweep	Oracle	Gain
Qwen2.5-VL-7B	32.7	56.2 → 28.8	46.4	+13.7
Pixtral-12B	41.2	55.6 → 21.6	56.2	+15.0
InternVL3-8B	30.7	59.5 → 23.5	52.9	+22.2

The comparison separates a shared depth response from model-dependent vertical control. It also motivates the paired amplification test: after finding a direction that changes answers, we test whether the resulting improvement depends on the image.

4.3 MOST OF THE VERTICAL INTERVENTION GAIN SURVIVES IMAGE REMOVAL

Most of the vertical amplification gain remains after replacing the matched image with a blank. To measure this directly, the Grounding Arbiter evaluates both input conditions before and after the same edit. Algorithm 1 specifies the paired evaluation; Table 3 reports the raw and balanced scores. For a fixed scoring rule, define

$$G_c = A_c^{\text{edit}} - A_c^{\text{base}}, \qquad c \in \{\text{real}, \text{blank}\}. \tag{2}$$

The contrast $G_{\text{real}} - G_{\text{blank}}$ reports how much larger the measured gain is with the matched image. Each condition uses its own unedited baseline.

Baseline behavior. Image removal reduces horizontal balanced accuracy by 31.9 points, improves depth by 14.7, and changes vertical by only 0.5 (Table 3). The model retains its average vertical performance without an informative image, while horizontal answers strongly benefit from the image.

Table 3: Paired image controls on Qwen2.5-VL-7B. Accuracy is reported both raw and balanced by gold pole. Amplification gains are raw percentage-point changes. Vertical gains use the recorded sweep; horizontal and depth gains retain the source summary precision (Appendix A).

	Raw accuracy (%)		Balanced accuracy (%)		Gain (pp)	
Axis	Real	Blank	Real	Blank	Real	Blank
Horizontal	82.4	49.4	82.6	50.7	+3	~ 0
Vertical	58.5	59.6	69.8	70.3	+20.5	+19.1
Depth	31.4	45.1	31.6	46.3	≤ 0	~ 0

The contrast extends beyond a gray blank. All five image replacements reduce horizontal accuracy and improve depth relative to matched images (Appendix Table 9). The horizontal loss ranges from 28.5 to 35.2 points, while depth improves by 11.1 to 17.3 points. Vertical is more sensitive to the replacement: gray and black preserve its score, whereas patch scrambling lowers it from 69.8% to 60.5%. The controls reveal which image changes affect unedited answers; the paired edit then measures the image dependence of the gain itself.

Ablation-persistent intervention gain. At $\gamma = 2$, activation amplification raises vertical accuracy from 58.5% to 79.0% with real images, a gain of 20.5 points. The same edit gains 19.1 points with gray images; the corresponding balanced gains are 13.6 and 12.2 points. Most of the apparent recovery therefore persists without the matched image. An off-band control shows no comparable gain. We call the retained benefit an *ablation-persistent intervention gain*, and the separation between intervention success and image dependence an *intervention–grounding dissociation*.

The raw gain contrast is 1.4 points, and the balanced gain contrast is also 1.4 points. The large matched-image improvement thus accompanies a much smaller difference between the two gains. The full strength sweep appears in Appendix Table 8.

Algorithm 1 Grounding Arbiter: paired input and intervention evaluation

Require: model M; items (q_i, I_i, g_i, y_i); fixed edit E; mismatched images J_i

Define $\text{Bal}(z) = \frac{1}{2} \sum_{p \in \{-,+\}} \text{mean}_{i:y_i=p}\, z_i$
for $c \in \{\text{real}, \text{blank}, \text{mismatch}\}$ **do**
  $I_i^c \leftarrow I_i$, $\textsc{Gray}(I_i)$, or J_i, respectively
  $z_i^c \leftarrow \mathbf{1}[M(q_i, I_i^c) = g_i]$ for each item i
  $A_c \leftarrow \text{Bal}(z^c)$
end for
for $c \in \{\text{real}, \text{blank}\}$ **do**
  $e_i^c \leftarrow \mathbf{1}[M_E(q_i, I_i^c) = g_i]$ using the same edit settings
  $G_c \leftarrow \text{Bal}(e^c) - A_c$
end for
return $\{A_c\}$, $\Delta = A_{\text{real}} - A_{\text{blank}}$, $L = A_{\text{blank}} - \frac{1}{2}$, G_{real}, G_{blank}, $G_{\text{real}} - G_{\text{blank}}$

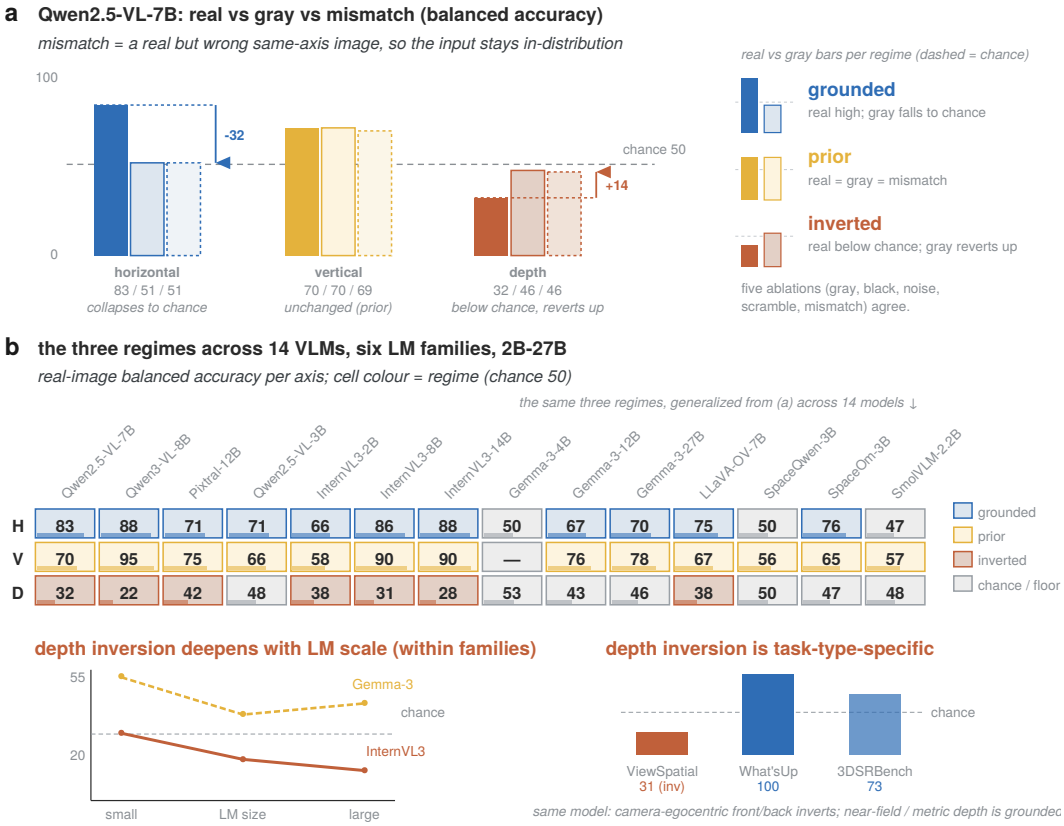


Figure 3: **Vision ablation reveals three contrasting axis-level patterns.** **(a)** Qwen2.5-VL-7B balanced accuracy with the matched image, a same-size gray blank, and a mismatch (a real but wrong same-axis image). Horizontal shows visual support, vertical retains above-chance performance after replacement, and the below-chance depth error weakens without the matched image. Additional gray, black, noise, patch-scramble, and mismatch controls are reported in Table 9. **(b)** The panel summarizes reported patterns across 14 VLMs from six LM families, together with within-family scale series and the cross-benchmark boundary (§5). The displayed regime labels are qualitative summaries from post-hoc thresholds. The text uses continuous score differences to compare model and task effects; chance is 50%.

Paired outcomes behind the average. For each question, paired image conditions distinguish real-only correct, blank-only correct, both correct, and neither correct outcomes. For raw accuracy over n questions,

$$A^{\text{real}} - A^{\text{blank}} = \frac{n_{\text{real only}} - n_{\text{blank only}}}{n}. \tag{3}$$

The accuracy gap is the net balance of the two opposing correctness transitions. Similar average scores can therefore accompany many changed outcomes. Computing these transitions before and after an edit locates which questions benefit under each image condition. For balanced accuracy, the decomposition is applied within each gold pole and then averaged (Appendix B.5).

5 Recurring Depth Deficits and Their Task Boundaries

The reversed depth response occurs across three models on ViewSpatial-Bench. Broader behavioral comparisons show a recurring depth deficit, while paired benchmark tests locate tasks on which the same models benefit strongly from the image.

5.1 Horizontal–Depth Asymmetry Recurs Across Model Families

Horizontal balanced accuracy exceeds depth accuracy in twelve of fourteen models, and thirteen have depth accuracy below 50% (Appendix Table 10). The ordering appears across different lan-

guage and vision component combinations (Zhu et al., 2025; Li et al., 2024). Gemma-3-4B and SmolVLM-2.2B are the two exceptions to the horizontal–depth ordering; both have horizontal scores near chance. Figure 3 summarizes the original axis-level patterns.

Larger models do not consistently resolve the deficit. Depth accuracy decreases across the available InternVL3 sizes (38.1%, 30.9%, 28.2%) and Qwen2.5-VL sizes (47.9%, 31.6%), while Gemma-3 is non-monotonic (52.6%, 43.0%, 45.6%; Fig. 14). The within-family comparisons show persistent, and sometimes stronger, depth errors. This deficit changes with the task: the same models gain substantially from matched images on the depth relations tested by What'sUp and 3DSRBench.

5.2 Matched images support depth on other tasks across models

The task boundary extends beyond the primary model. Qwen2.5-VL-7B, Qwen3-VL-8B, Pixtral-12B, and SpaceQwen-3B reach 98–100% on What'sUp-B front/behind, with accuracy falling to 45.1–49.0% under mismatched images (Table 4). Their weak ViewSpatial depth performance thus coexists with large image benefits on a different depth task.

Table 4: Depth accuracy (%) with real and mismatched images across two benchmarks. What'sUp-B tests front/behind ($n = 204$ per axis); the primary 3DSRBench adapter has 1028 depth items. Entries retain each model's recorded task run.

	What'sUp-B		3DSRBench	
Model	Real	Mismatch	Real	Mismatch
Qwen2.5-VL-7B	99.5	48.5	72.8	49.4
Qwen3-VL-8B	100.0	45.1	79.2	53.1
Pixtral-12B	98.0	45.6	69.1	54.3
SpaceQwen-3B	100.0	49.0	54.7	50.0

On 3DSRBench closer-to-camera, Qwen2.5-VL-7B, Qwen3-VL-8B, and Pixtral-12B also show positive matched-image effects: 23.4, 26.1, and 14.8 points. SpaceQwen-3B has a smaller 4.7-point gap and accuracy near chance. The magnitude of the image benefit varies across both model and task.

Vertical exhibits a similar task contrast in the primary model. What'sUp on/under scores 99.5% with real images, 48.5% with gray, and 47.6% with mismatch. On 3DSRBench vertical tasks, the corresponding scores are 56.1, 56.2, and 54.1% (Appendix Table 11). These paired comparisons explain why axis names alone give an incomplete account of image use: the model's dependence on visual evidence changes with the task formulation.

6 Depth Controllability and Model-Dependent Correction

The shared depth response admits model-dependent corrections. We compare recalibration, reflection, rotation (Vu & Nguyen, 2025), activation and feature steering (Qiu et al., 2024; Wang et al., 2025; Li et al., 2023; Pach et al., 2025), and trained low-rank edits (Hu et al., 2022; Wu et al., 2024) to identify effective operators and layers.

6.1 Opposing directional responses recur across models

Seven of eight models have a positive probe-response slope and a negative behavioral-response slope along the depth direction (Fig. 17a, Table 12). Moving toward the decoded pole therefore moves behavior away from the corresponding benchmark answer. Qwen2.5-VL-3B has a near-zero positive behavioral slope. This pattern spans five VLM series and four language backbone families: Qwen2, Qwen2.5, Qwen3, and Mistral.

6.2 Rotation gains vary sharply by model

On Qwen3-VL-8B, a norm-preserving rotation raises depth accuracy from 21.4% to 77.1% (+55.7 points), approaching the trained-edit gain of +57.1 in the same comparison (Table 13). The separate three-seed summary appears in Appendix D. The contrast is sharper on Qwen2.5-VL-7B and Pixtral-12B: their best tested rotations gain 15.5 and 2.4 points, while trained edits gain 42.9 and 32.5. Rotation can therefore produce a large correction, but its effectiveness varies substantially by

model. Table 13 summarizes the tested method families; Appendix E describes the scope of these comparisons.

6.3 Decoding and controllability peak at different layers

The layer that best exposes a relation to a probe can differ from the layer where an edit has the largest behavioral effect. On Qwen3-VL-8B, the recorded depth profile reaches its decoding peak at L26, while one-sided controllability at $\beta = 4$ peaks at L24 (Fig. 4). Controllability falls sharply after L24 even as the probe remains accurate in later layers. The two profiles therefore select different intervention sites.

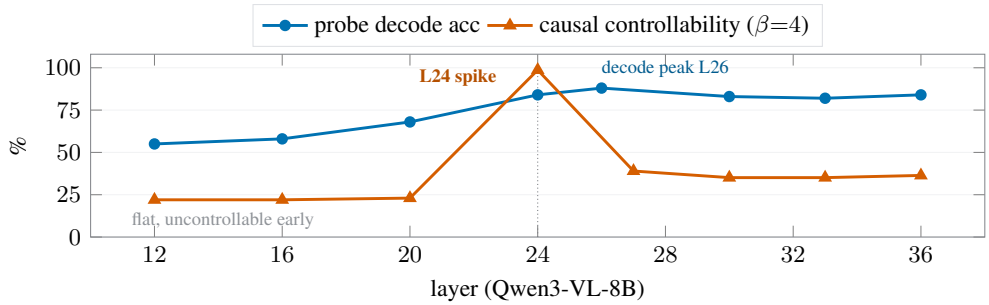


Figure 4: **Controllability and decoding peak at different layers on Qwen3-VL-8B.** The archived layer-sweep probe accuracy peaks at L26; this profile uses the earlier probing protocol (Appendix A). The one-sided controllability measure at $\beta = 4$ peaks at L24 (98.7%), then drops in later layers. The detailed trained-edit comparison includes L30 (Table 14).

Across seven models, peak controllability lies at layers 18–36 in the tested sweeps (Appendix Fig. 19). The reported probe and controllability peaks coincide in four models and differ in three. Qwen3-VL-8B's controllability peak precedes its probe peak, whereas the recorded peaks for Qwen2.5-VL-7B and InternVL3-14B have the opposite ordering. Layer indices are local to each model.

Direction choice matters too: rank-1 ReFT and four-component ICA reflection outperform several probe-derived edits (Table 14). Layer selection changes the objective of an intervention search. Probing locates where a relation is accessible, whereas intervention sweeps locate where modifying that representation changes answers. Paired image controls then establish how the measured improvement varies with the available visual input.

7 Discussion and Conclusion

The central finding is an intervention–grounding dissociation: the vertical amplification gain is 20.5 points with real images and 19.1 with blank images. Most of this edit's accuracy gain thus survives removal of the matched image.

A broad depth deficit and three-model reversed responses coexist with model-dependent correction operators and layer profiles. Across tasks, weak ViewSpatial depth performance coexists with large matched-image benefits elsewhere. Together, these findings motivate selecting edit sites by their behavioral response and evaluating the resulting gains under paired image controls.

Paired gains guide the interpretation of a correction. Gains with both matched and blank images show that the edit also helps without the matched input. Repeating the edit with mismatched images tests its dependence on the correct image–question correspondence. Section 4.3 gives a transition decomposition for examining changes hidden by aggregate scores. Independent amplification validation and grouped probing across more models would extend the evidence. The Grounding Arbiter makes the evaluation target explicit: measure both answer improvement and its dependence on visual evidence. This evaluation can guide the selection of interventions whose accuracy gains depend on the matched image, particularly when several edits achieve similar task performance.

Ethics Statement

This study evaluates publicly released VLMs on existing spatial-reasoning benchmarks. It introduces no human-subject study or new personal-data collection. The depth results concern behavior under the evaluated label conventions; they do not establish that the benchmark annotations are incorrect. Spatial errors may matter in embodied and assistive applications, but we do not evaluate deployment safety in those settings.

Reproducibility Statement

The benchmark decomposition, scoring, grouped probe evaluation, directional edits, and paired image controls are described in §3 and Appendix A. Each experiment retains its own baseline. Appendix E records evaluation selection, protocol versions, and the scope of the reported comparisons.

AI Use Statement

Generative AI tools assisted with code development and manuscript drafting and editing. The authors take responsibility for the experiments, results, references, and final text.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Théophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. URL https://arxiv.org/abs/2410.07073.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

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

Sanjay Basu, Sadiq Y. Patel, Parth Sheth, Bhairavi Muralidharan, Namrata Elamaran, Aakriti Kinra, John Morgan, and Rajaie Batniji. Interpretability without actionability: Mechanistic methods cannot correct language model errors despite near-perfect internal representations. *arXiv preprint arXiv:2603.18353*, 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498, 2025. doi: 10.1109/ICRA55743.2025.11128671. arXiv:2406.13642.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024. doi: 10.1109/CVPR52733.2024.01370.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7395–7408, 2025. arXiv:2503.13111.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024. doi: 10.18653/v1/2024.acl-short.33. URL https://aclanthology.org/2024.acl-short.33/. arXiv:2406.05756.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *ICLR*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *ICLR*, 2026. arXiv:2506.03135.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *EMNLP*, 2023. arXiv:2310.19785; benchmark commonly cited as "What'sUp".

Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2601.12626.

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. *arXiv preprint arXiv:2408.03326*, 2024.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. In *European Conference on Computer Vision (ECCV)*, pp. 95–111, 2026. arXiv:2505.21500.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.03341.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9570–9581, 2026a. arXiv:2510.27606.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2510.17771.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M. de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2412.07825.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. URL https://arxiv.org/abs/2504.05299.

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

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. In *European Conference on Computer Vision (ECCV)*, pp. 22–39, 2026. arXiv:2605.30161.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.02821.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. Spectral editing of activations for large language model alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.09719.

Remyx AI. SpaceOm. Hugging Face model card, 2025a. URL https://huggingface.co/remyxai/SpaceOm.

Remyx AI. SpaceQwen2.5-VL-3B-Instruct. Hugging Face model card, 2025b. URL https://huggingface.co/remyxai/SpaceQwen2.5-VL-3B-Instruct.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15504–15522, 2024. arXiv:2312.06681.

Daniele Savietto, Declan Campbell, André Panisson, Marco Nurisso, Giovanni Petri, Jonathan D. Cohen, and Alan Perotti. The geometry of representational failures in vision language models. *arXiv preprint arXiv:2602.07025*, 2026.

Matthieu Tehenan, Christian Bolivar Moya, Tenghai Long, and Guang Lin. Linear spatial world models emerge in large language models. *arXiv preprint arXiv:2506.02996*, 2025.

Minh Hieu Vu and Tan M. Nguyen. Angular steering: Behavior control via rotation in activation space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.26243.

Haoming Wang and Wei Gao. Uncovering and shaping the latent representation of 3d scene topology in vision-language models. *arXiv preprint arXiv:2605.07148*, 2026.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2211.00593.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.12299.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *NeurIPS*, 2024.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025. doi: 10.1109/CVPR52734.2025.00994. VSI-Bench.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *ICLR*, 2026. arXiv:2505.23764.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL `https://arxiv.org/abs/2504.10479`.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A EXPERIMENTAL PROTOCOL

Model and decoding. The primary model is Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), evaluated with 4-bit NF4 weights on a single RTX 4090, SDPA attention, and greedy decoding with at most eight new tokens. Its language tower has 28 decoder layers and hidden dimension 3584. In the recorded implementation, `hidden_states[i]` is the output of decoder layer $i-1$, with index zero reserved for the embedding. The recorded Transformers version is 4.57.

Answer reading and scoring. The evaluator accepts answer-letter variants with and without a leading space. Binary option order is randomized per item; the gold-pole-to-letter map must be used consistently for scoring and intervention targets. Raw accuracy weights items equally, whereas balanced accuracy averages recalls for the two gold poles. The image-control run has raw H/V/D baselines of 82.4/58.5/31.4. The behavior–probe summary and correction experiments report different baselines; we keep these summaries separate (Appendix E).

Probes. The updated evaluation groups items by source scene using five outer GroupKFold partitions and nests layer/locus selection within training data. The grouped evaluation reports H/V/D accuracies of $49.1/86.9/57.1\%$, with grouped shuffled-label controls of $52.7/61.4/48.9\%$. Vertical selects text-pooled activations in all five outer folds. The earlier pipeline caches all 29 hidden-state indices on the 1,047 pure-axis questions and fits a StandardScaler–PCA-50–logistic-regression pipeline inside each question-level fold. It reports the maximum over layers and loci. Its plots are retained as exploratory profiles; the scene-grouped estimates in Table 1 provide the primary probe result. The rerun's item-level reference scores are $57.5/89.1/69.9\%$, which differ from the original $84.9/94.1/77.2\%$. These are separate evaluations, not a paired estimate of leakage magnitude. The sparse-neuron localizer ranks standardized class-mean differences, $|d'|$.

Steering scale. Fixed-direction steering is scaled by the standard deviation of activations projected onto the unit direction, so α is measured in those units. The reported sweep uses $|\alpha| \leq 4$. Larger magnitudes can produce substantial changes to the residual norm and are not used to support the main interpretation.

Activation amplification. Activation amplification applies a forward hook to each selected MLP's `down_proj` output. It adds

$$\gamma \left(\sum_{i \in S_l} \operatorname{sign}(d'_i) a_i(x) \right) \hat{d}_l, \qquad |S_l| = 64. \tag{4}$$

The write magnitude depends on the input's neuron activations. Labelled training items supply the per-layer difference-of-means direction and neuron selection. The real and gray arms use the same direction, layers, strength, and evaluation items. The code defaults to an 80/20 train/eval split with seed 0, and also supports full-set evaluation. The sweep tests $\gamma \in \{0.5, 1, 2, 4, 8\}$ on the reported evaluation set; $\gamma = 2$ is its selected point. The reported bands include L20–24 overall, L21–25 for vertical, L19–23 for depth, and an L2–6 off-band control. The sweep is reported as an exploratory comparison; its available configuration record does not establish the exact sample counts and run flags.

Image controls. A same-resolution gray blank passes through the same image preprocessor as the real image. Additional controls use black pixels, noise, patch scrambling, and a mismatched image from the same axis. The mismatch preserves natural image content but disrupts its correspondence with the question. The configuration record omits the noise and scrambling parameters and the mismatch-pairing seed.

B Evidence for the Intervention–Grounding Dissociation

The supporting analyses connect the behavior–probe gap to directional control and paired image replacement. The related-work comparison situates this question among spatial representation studies (Table 5).

Table 5: Related spatial representation studies and the question each addresses.

Study	Relation to the present question
Qi et al. (2025)	Spatial information in VLM representations.
Wang & Gao (2026)	Representation and modification of 3D scene topology.
Savietto et al. (2026)	Geometry of representational failures.
Min et al. (2026)	Relations between vertical and distance representations.
Kang et al. (2026)	Linear mechanisms and interventions for spatial and temporal reasoning.
Basu et al. (2026)	Whether decodable information can be used to correct errors.
This study	Per-axis accuracy changes under image replacement, alongside probing and intervention results.

B.1 Weak answers coexist with accessible axis information

The archived item-level probe analysis shows larger behavior–probe gaps on vertical and depth (Fig. 5). The primary grouped nested estimates appear in Table 1. Depth remains below chance under both tested camera-relative phrasings (Table 6).

Table 6: Depth accuracy under two camera-relative phrasings ($n = 153$). Both remain below 50% chance. Units are percent and percentage points.

Phrasing	Accuracy	Difference from chance
“front / back”	28.1	−21.9
“closer / farther”	32.7	−17.3

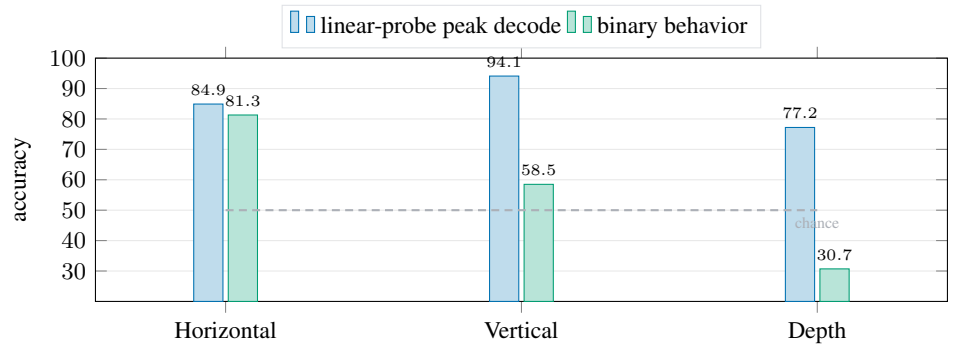


Figure 5: **Vertical and depth show large behavior–probe gaps.** Peak probe accuracy and binary answer accuracy differ by approximately 4, 36, and 47 percentage points on horizontal, vertical, and depth relations. Vertical has the highest probe score; depth has the lowest behavioral score. Values are percentages. These are archived item-level probe estimates; the primary scene-grouped nested results are in Table 1.

B.2 Decodability and intervention responses rise in mid-to-late layers

In the exploratory item-level profiles, decision- and text-token readouts rise above image-pooled scores in mid-to-late layers (Fig. 6). The decision-token profiles show vertical reaching the highest peak (Fig. 7). Gold-signed injection and direction removal produce different axis-level responses in the layer sweep (Fig. 8); the band results are reported in Table 7.

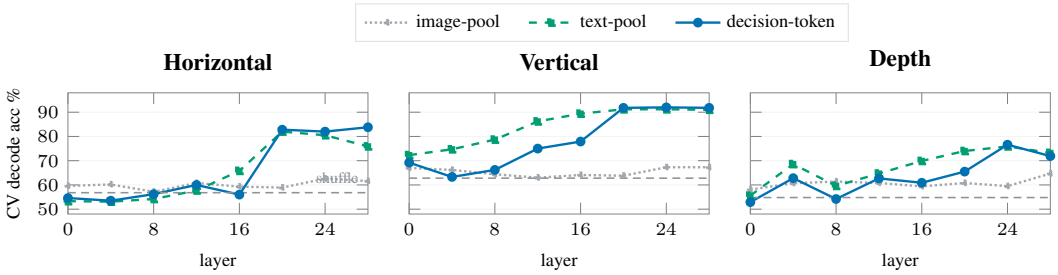


Figure 6: **Decision- and text-token readouts exceed image-pooled readouts.** Layer-wise, five-fold probe accuracy on Qwen2.5-VL-7B compares decision-token residuals, mean-pooled text tokens, and mean-pooled image tokens. Decision/text scores rise in mid-to-late layers while image-pooled scores remain lower. Dashed lines show shuffled-label scores. Values are percentages. These are archived item-level probe estimates; the primary scene-grouped nested results are in Table 1.

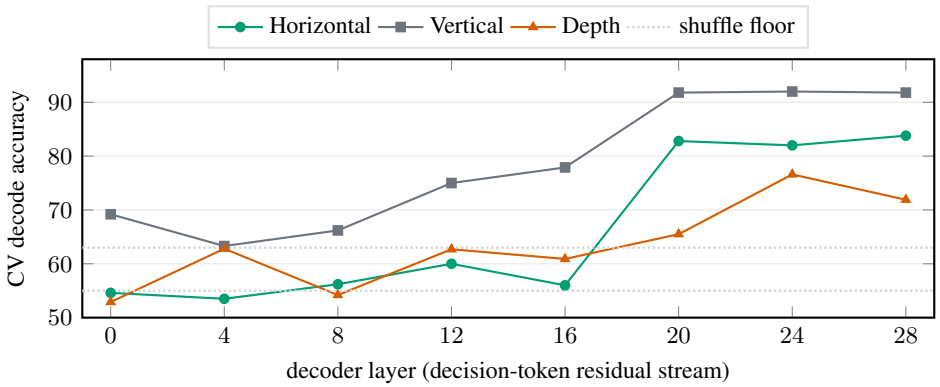


Figure 7: **Vertical reaches the highest decision-token probe accuracy.** Five-fold logistic-probe accuracy uses PCA-50 features at each layer. Scores rise at L16–24; vertical peaks near 94%, horizontal near 85%, and depth near 77%. Dotted lines show shuffled-label scores of 55–63%. These are archived item-level probe estimates; the primary scene-grouped nested results are in Table 1.

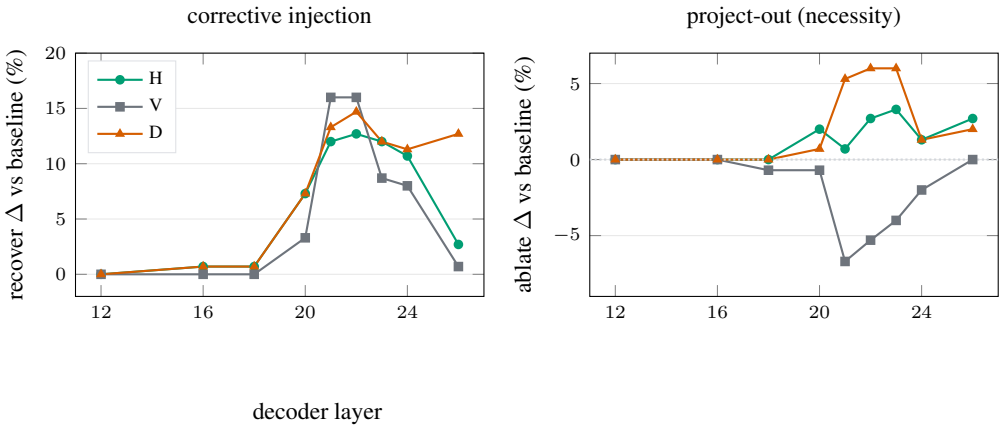


Figure 8: **Injection and removal have different layer-wise effects across axes.** On the $n = 150$ subsample, gold-signed injection gains peak around L21–22 (left). Direction removal reduces vertical accuracy around L21–24, changes horizontal little, and improves depth (right). Multi-layer results appear in Table 7. Gains are percentage points; injection uses the gold pole.

Table 7: Multi-layer interventions on Qwen2.5-VL-7B. Injection uses the gold pole for each item and is an oracle control, not test-time recovery. Values are raw accuracies (%); parenthesized changes preserve the source rounding.

Axis	Band	Baseline	Oracle injection	Removal
Horizontal	L19–23	82.4	98.1 (+15.6)	82.2 (−0.2)
Vertical	L21–25	58.5	70.2 (+11.7)	49.2 (−9.3)
Depth	L19–23	31.4	62.7 (+31.4)	36.6 (+5.2)

B.3 Directional control and blank-image gain

Fixed-direction steering moves horizontal and vertical positive-pole probabilities upward and depth downward (Fig. 10). The amplification sweep then separates behavioral gain from the presence of a matched image: the vertical gain remains large at the measured blank-image operating point (Fig. 11). For vertical, Qwen changes from 58.5% to 49.7% after removal (random: 58.7%), Pixtral from 66.8% to 67.3% (random: 67.1%), and InternVL3-8B from 86.2% to 85.9% (random: 86.3%). Horizontal removal scores are 81.7, 67.2, and 86.1%, compared with random scores of 82.6, 70.3, and 85.9%.

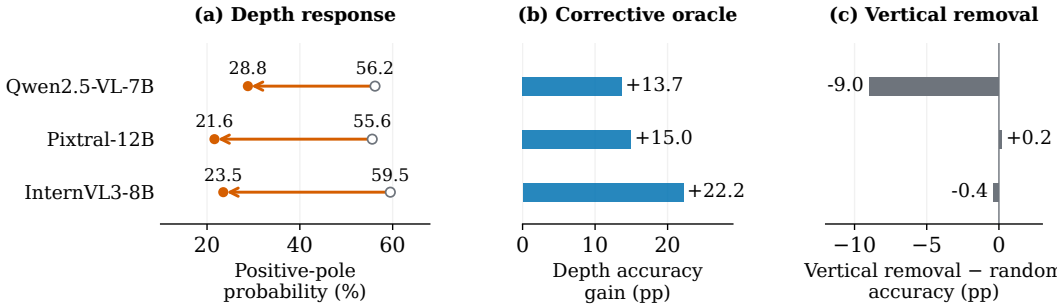


Figure 9: **Depth responses recur across three models, while vertical removal effects differ.** (a) Recorded positive-pole probability endpoints as steering moves toward the positive depth pole; arrows indicate endpoint changes, not interpolated sweeps. (b) Depth accuracy gain from reverse-sign, gold-conditioned oracle injection. (c) Vertical accuracy after axis-direction removal minus accuracy after random-direction removal. All models use the same evaluation procedure with seed 0 and full axis sets ($n_{\text{depth}} = 153$, $n_{\text{vertical}} = 376$). Endpoint scores appear in Table 2.

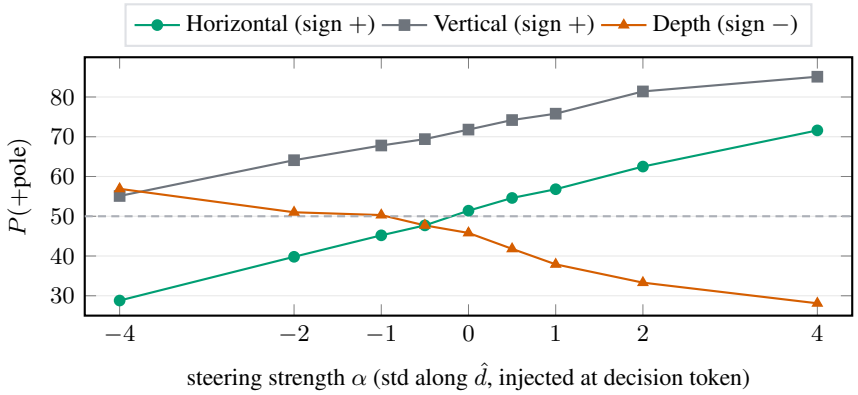


Figure 10: **The three axes respond differently to the same steering rule.** Adding each axis's difference-of-means direction at the decision token increases horizontal and vertical positive-pole probability but decreases depth probability. Qwen2.5-VL-7B, $n = 1047$; probabilities are percentages.

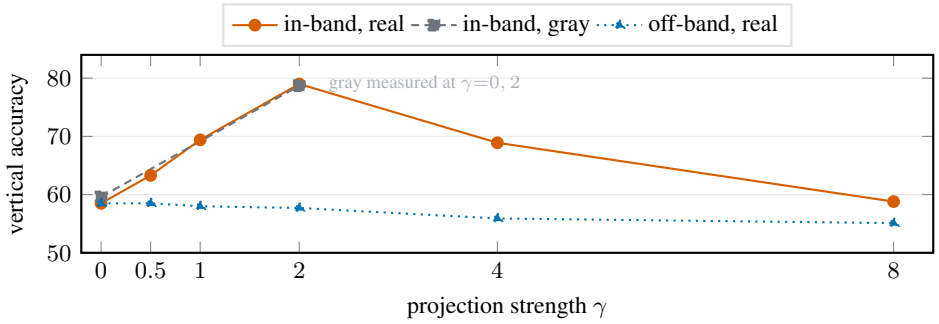


Figure 11: **Most of the vertical amplification gain survives image removal.** At $\gamma = 2$, real-image accuracy rises from 58.5% to 79.0% ($+20.5$ points); gray-image accuracy rises from 59.6% to 78.7% ($+19.1$). The original plot shows gray-image points at $\gamma = 0, 2$; its dashed segment connects these anchors. Table 8 reports the full gray sweep recorded in the run log. The L2–6 off-band control shows no comparable improvement.

Table 8: Recorded vertical amplification sweep on Qwen2.5-VL-7B (raw accuracy, %). Real and gray use the same edit at each strength. The real-image sweep selects $\gamma = 2$; all recorded operating points are shown to expose the strength dependence.

γ	Real	Gray	Real gain	Gray gain	Off-band real
0	58.5	59.6	0.0	0.0	58.5
0.5	63.3	62.5	+4.8	+2.9	58.5
1	69.4	71.0	+10.9	+11.4	58.0
2	79.0	78.7	+20.5	+19.1	57.7
4	68.9	64.4	+10.4	+4.8	55.9
8	58.8	53.2	+0.3	-6.4	55.1

B.4 Image controls preserve the horizontal–depth contrast

All five replacement controls reduce horizontal accuracy and improve depth relative to matched images (Table 9). Vertical stays above chance under every control, with the largest drop under patch scrambling: $69.8\% \rightarrow 60.5\%$. The controls thus locate different net effects of image content across the three axes.

Table 9: Unedited balanced accuracy (%) under five image-replacement controls on Qwen2.5-VL-7B. All replacements reduce horizontal accuracy and improve depth relative to matched images. Vertical's largest drop is under patch scrambling: 9.3 percentage points.

Axis	Real	Gray	Black	Noise	Scramble	Mismatch
Horizontal	82.6	50.7	49.6	47.4	54.1	50.7
Vertical	69.8	70.3	70.1	66.5	60.5	68.6
Depth	31.6	46.3	44.2	42.7	48.9	45.7

B.5 Paired outcomes explain the aggregate image effect

Equation 3 decomposes the raw image-effect gap into real-only and blank-only correct outcomes. For balanced accuracy, let n_p be the number of items with gold pole $p \in \{-, +\}$. Applying the same identity within each pole gives

$$A_{\text{bal}}^{\text{real}} - A_{\text{bal}}^{\text{blank}} = \frac{1}{2} \sum_{p \in \{-,+\}} \frac{n_{\text{real only},p} - n_{\text{blank only},p}}{n_p}. \tag{5}$$

Both-correct and neither-correct items contribute zero to the gap. The two remaining counts separate losses from gains under image removal, while their difference gives the aggregate image effect. The identity applies separately to the unedited and edited predictions.

C Cross-Model and Cross-Task Boundaries

The model and benchmark comparisons test two dimensions of scope: which patterns recur across released VLMs, and which change when the task changes.

C.1 Axis asymmetry across fourteen models

Table 10: Real-image balanced accuracy (%) for fourteen VLMs spanning six LM families. H/V/D denote horizontal, vertical, and depth. A dash marks an unreported value. Model names give nominal sizes; language backbones are listed separately.

Model	LM backbone/family	H	V	D
InternVL3-2B (Zhu et al., 2025)	Qwen2.5-1.5B	65.8	57.5	**38.1**
InternVL3-8B	Qwen2.5-7B	86	89.5	30.9
InternVL3-14B	Qwen2.5-14B	88	90	**28.2**
Gemma-3-4B (Gemma Team et al., 2025)	Gemma 3	49.5	—	52.6
Gemma-3-12B	Gemma 3	67.3	75.9	**43**
Gemma-3-27B	Gemma 3	69.8	78.4	45.6
Qwen2.5-VL-3B (Bai et al., 2025b)	Qwen2.5	70.7	65.9	**47.9**
Qwen2.5-VL-7B	Qwen2.5	82.6	69.8	31.6
LLaVA-OV-7B (Li et al., 2024)	Qwen2-7B	75.3	66.6	37.8
Qwen3-VL-8B (Bai et al., 2025a)	Qwen3	87.7	94.6	**22.4**
Pixtral-12B (Agrawal et al., 2024)	Mistral	70.8	74.5	42.2
SpaceOm (Remyx AI, 2025a)	Qwen2.5 (sp-FT)	75.6	65.3	47.3
SpaceQwen-3B (Remyx AI, 2025b)	Qwen2.5 (sp-FT)	*50.2*	56.2	*49.9*
SmolVLM-2.2B (Marafioti et al., 2025)	SmolLM	*47.4*	57.4	47.8

Observed depth accuracy is below 50% in thirteen models, while horizontal exceeds depth in twelve. Table 10 contains the full roster; Fig. 12 visualizes ten models. For the primary model's native composite class ($n = 726$), overall accuracy is 39.9%. Items containing a left/right component score 52.1% ($n = 520$), versus 9.2% for the other items ($n = 206$), extending the descriptive horizontal contrast to composite questions.

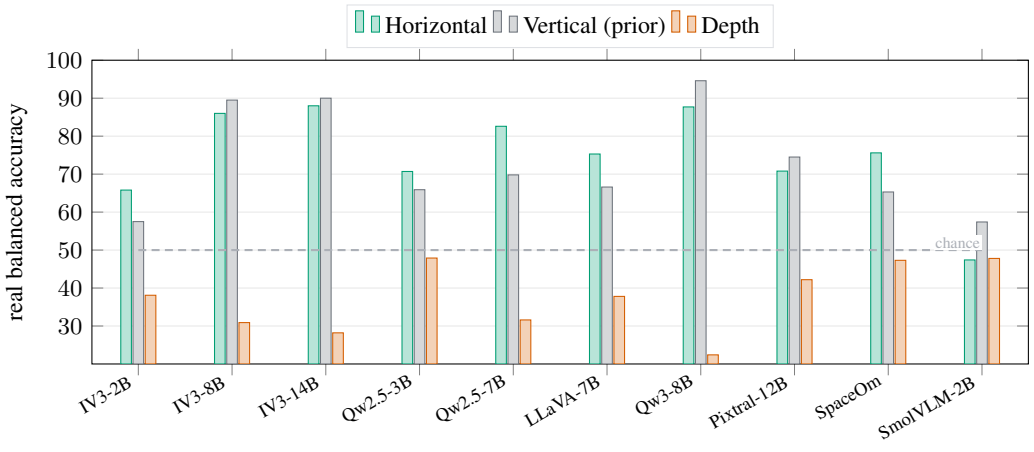


Figure 12: **Horizontal–depth asymmetry recurs across models.** Real-image balanced accuracy for ten of the fourteen VLMs in Table 10. The full table has horizontal accuracy above depth in twelve models and depth below 50% in thirteen. Values are percentages.

C.2 Archived item-level probe and image-effect summaries

The original eight-model scatter pairs item-level probe estimates with image effects (Fig. 13). Its layer-selected probe scores use the earlier protocol; the grouped nested rerun applies to the primary model in Table 1.

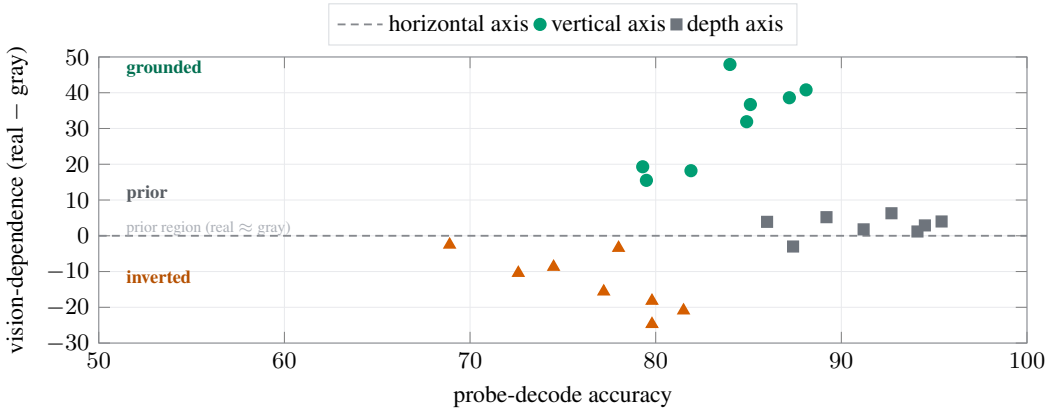


Figure 13: **Archived exploratory probe and image-effect summary.** Each point is a model–axis pair from eight VLMs. Probe scores use the earlier item-level layer-selection protocol, separate from the grouped nested estimate in Table 1. The horizontal axis shows probe accuracy; the vertical axis shows real-minus-gray accuracy in percentage points. Horizontal relations have positive image effects, vertical effects lie near zero, and depth effects are negative. This cross-model summary uses separate runs from Table 3; the primary model's vertical/depth gaps are $+1.2/-15.6$ here and $-0.5/-14.7$ there (Appendix E).

C.3 Size ladders retain depth deficits and above-chance blank performance

Depth decreases with sampled model size in InternVL3 and Qwen2.5-VL and partly recovers in the largest Gemma-3 model (Fig. 14). The blank-score summary separately describes vertical performance above chance across the sampled models (Fig. 15).

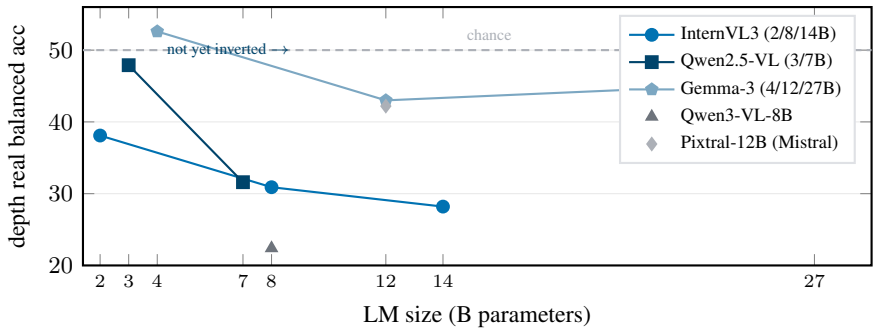


Figure 14: **Larger sampled models do not consistently improve depth accuracy.** Depth decreases across the available InternVL3 sizes (38.1, 30.9, 28.2) and Qwen2.5-VL sizes (47.9, 31.6). Gemma-3 is non-monotonic (52.6, 43.0, 45.6). The horizontal axis uses nominal model sizes. Accuracies are percentages.

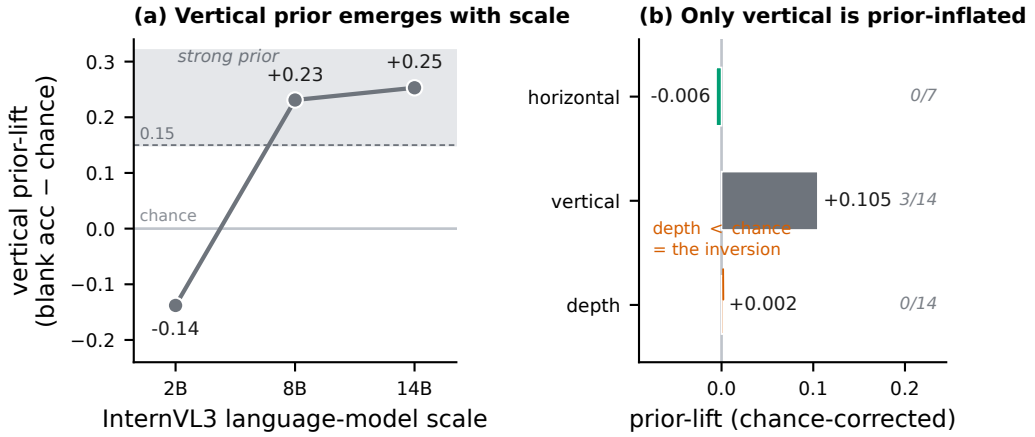


Figure 15: **Above-chance blank performance across model sizes and axes.** (a) Vertical prior-lift, $\text{acc}^{\text{blank}} - \frac{1}{2}$, across the sampled InternVL3 sizes. (b) The seven-model summary compares axes, using lift ≥ 0.15 to count strong cells. This exploratory panel uses a separate condition summary from the paired controls (Appendix E).

C.4 Changing the task changes depth and vertical image dependence

The primary model's depth accuracy is high and image-dependent on What'sUp front/behind and 3DSRBench closer-to-camera, contrasting with ViewSpatial-Bench (Table 11, Fig. 16). Table 4 extends the paired depth comparison to four models. Vertical also shifts from weak image effects on ViewSpatial to a large mismatch effect on What'sUp on/under.

Table 11: Cross-benchmark balanced accuracy (%) for Qwen2.5-VL-7B. Matched-image depth accuracy drops sharply under mismatch on What'sUp front/behind and 3DSRBench closer-to-camera.

Benchmark	Task	Real	Blank	Mismatch
What'sUp-A	horizontal (left/right)	99	51.5	51.5
What'sUp-A	vertical (on/under)	99.5	48.5	47.6
What'sUp-B	horizontal	100	52.5	47.1
What'sUp-B	depth (front/behind)	**99.5**	50	48.5
3DSRBench	vertical (height)	56.1	56.2	54.1
3DSRBench	depth (closer-to-cam)	**72.8**	56.6	49.4

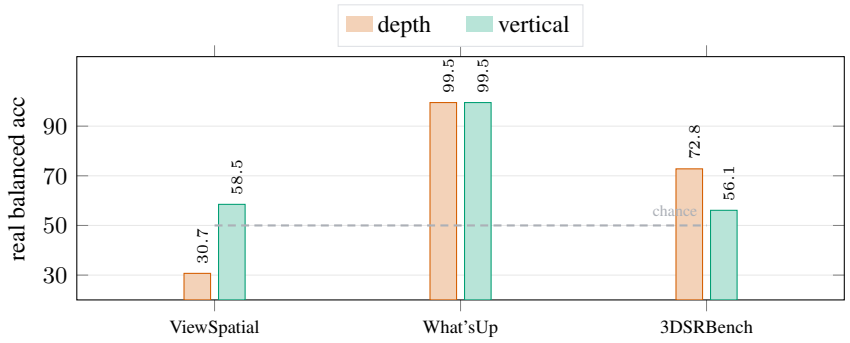


Figure 16: **Depth and vertical performance change across task formulations.** Qwen2.5-VL-7B has low depth accuracy on ViewSpatial-Bench but reaches 99.5% on What'sUp front/behind and 72.8% on 3DSRBench closer-to-camera. The mismatch controls in Table 11 show large depth losses on the latter two tasks. What'sUp on/under also has a large vertical mismatch effect.

D CORRECTION ANALYSES

Correction results separate a recurring directional mismatch from the model-specific edits that improve depth accuracy.

D.1 OPPOSING SLOPES AND MODEL-DEPENDENT CORRECTION GAINS

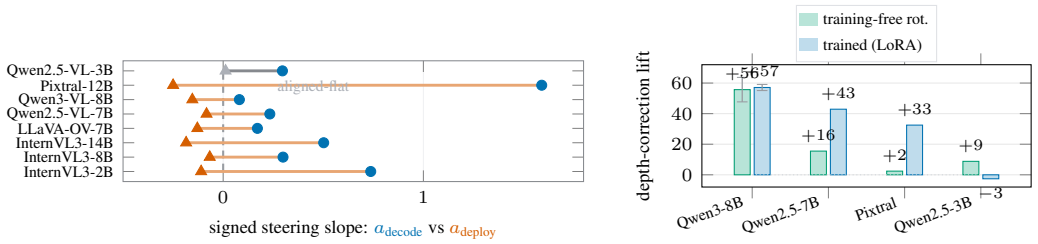
Seven of eight models show opposing probe and behavioral slopes (Table 12). The method comparison shows the largest rotation gains on Qwen3-VL-8B and much larger trained-edit gains on Qwen2.5-VL-7B and Pixtral-12B (Table 13, Fig. 17). A separate three-seed summary reports $+49 \pm 8$ for rotation and $+55 \pm 2$ for the trained edit on Qwen3-VL-8B, with greater reported variation for rotation.

Table 12: Depth steering slopes and best reported correction gains for eight models (single-seed summary). a_{dec} and a_{dep} are probe-output and behavioral slopes. Gains are percentage points. Seven models have negative behavioral slopes, across five VLM series and four LM families. The Qwen3 recalibration entry differs from Table 13 (Appendix E).

Model	Base (%)	a_{dec}	a_{dep}	Rotation	Recalib.
InternVL3-2B (Zhu et al., 2025)	38.3	+.74	−.11	−2.1	+1.4
InternVL3-8B	26.4	+.30	−.07	+8.1	+2.6
InternVL3-14B	31.4	+.50	−.19	−1.0	**+6.2**
LLaVA-OV-7B (Li et al., 2024)	36.2	+.17	−.13	**+13.8**	+7.9
Qwen3-VL-8B	21.4	+.08	−.15	**+55.7**	+6.0
Qwen2.5-VL-7B	26.2	+.23	−.08	+15.5	+11.2
Pixtral-12B (Agrawal et al., 2024)	39.8	+1.59	−.25	+2.4	+1.0
Qwen2.5-VL-3B (Bai et al., 2025b)	48.1	+.30	+.01	+8.8	+0.0

Table 13: Exploratory depth-correction gains (pp). Baselines are 26.2/21.4/39.8/48.1% in column order. Rows retain the recorded method groupings; selection and implementation details are collected in Appendix E.

Correction	Qwen2.5 7B	Qwen3 8B	Pixtral 12B	Qwen2.5 3B
recalibrate (1-D scalar)	+11.2	+0	+1	+0
Logit-lens reflection	n/a	+0	+0.7	−2.7
Single-layer rotation (Vu & Nguyen, 2025)	+8.8	**+55.7**	+0.7	+8.8
Multi-layer rotation (Vu & Nguyen, 2025)	+15.5	**+54.8**	+2.4	+1.9
SEA (Qiu et al., 2024) / SADI (Wang et al., 2025) edits	+10	+6	+1	+1.4
ITI head reflection (Li et al., 2023)	+5.5	+1.4	+0	+0.2
SAE feature steering (Pach et al., 2025)	+2.9	+1.2	+0.2	−1.2
LoRA (Hu et al., 2022) / ReFT (Wu et al., 2024) (trained)	**+42.9**	**+57.1**	**+32.5**	−2.6



(a) Opposing probe and behavioral slopes in seven of eight models.

(b) Rotation and trained-edit gains differ by model.

Figure 17: **A recurring directional mismatch accompanies model-dependent correction.** Left: depth probe-response slopes are positive in all eight models, while behavioral slopes are negative in seven, across five VLM series and four LM families. Right: rotation has a large gain on Qwen3-VL-8B, while trained edits are stronger on Qwen2.5-VL-7B and Pixtral-12B. Bars retain the single-summary values; the Qwen3 three-seed summary reports rotation $+49 \pm 8$ and trained edit $+55 \pm 2$.

D.2 THE ROTATION ACTS IN A TWO-DIMENSIONAL PLANE

Let b_1 be the normalized difference of class means and b_2 the leading principal component after removing the b_1 component. For $c_j = h^\top b_j$, the operator is

$$\mathcal{R}_\theta(h) = h + (c_1' - c_1)b_1 + (c_2' - c_2)b_2, \qquad \begin{pmatrix} c_1' \\ c_2' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}. \tag{6}$$

It preserves the norm and changes only the in-plane component. At $\theta = \pi$, both in-plane coordinates reverse sign. The one-dimensional Householder reflection $h - 2(h^\top b_1)b_1$ reverses only the b_1 component. The supervised rotation uses labels to estimate b_1. Figure 18 illustrates the geometry schematically. A preliminary top-PC-direction result with fixed $\theta = \pi$ reports approximately 56 points of gain on Qwen3-VL-8B. The available result specifies the direction type and angle, but omits the centering procedure, subspace construction, and selection split.

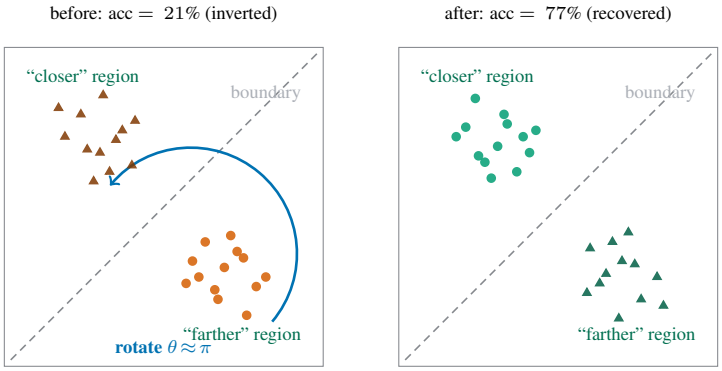


Figure 18: **A two-dimensional rotation changes the alignment of a readout.** The schematic illustrates a norm-preserving rotation in the b_1, b_2 plane; the points depict the operator's geometry. The displayed 21.4% and 77.1% accuracies come from the Qwen3-VL-8B correction summary in Table 13.

D.3 CORRECTION AND DECODING PEAKS CAN DIFFER

Peak controllability lies at layers 18–36 across the seven plotted models (Fig. 19). Qwen3-VL-8B peaks at L24 for controllability and L26 for decoding (Fig. 4); its detailed correction comparison includes effective ReFT and ICA edits at L30 (Table 14).

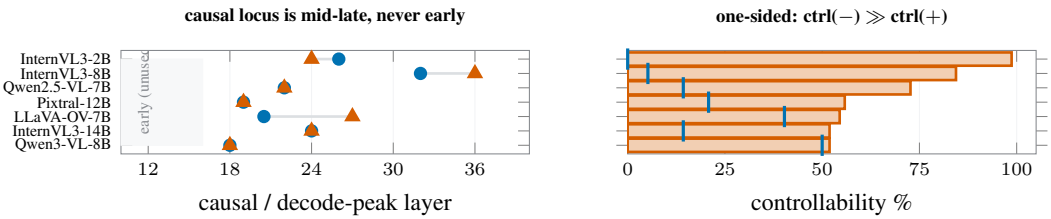


Figure 19: **Controllability peaks and steering asymmetry across seven models.** Left: reported probe and controllability peak layers. Right: response magnitudes for the two steering signs; six of seven models respond more strongly to the negative steer. Layers are indexed within each model; controllability is reported in percent.

Table 14: Exploratory Qwen3-VL-8B depth corrections (baseline 21.4%). Accuracies are percentages and gains are percentage points. Values preserve the original rounded summary, so displayed differences can vary by 0.1 point. The summary does not specify search budgets or data splits.

Correction	Best setting	Accuracy	Gain
ReFT (Wu et al., 2024) (trained)	rank 1, L30	**79.2**	**+57.8**
ICA-subspace reflect	k=4, L30	**81.0**	**+59.5**
probe-direction reflect	k=1. . .8	21.4	+0.0
GEVD subspace (closed-form)	k=1, L30	21.7	+0.2

D.4 Direction choice and neuron localization

Difference-of-means and probe directions produce different steering and removal responses (Table 15). These measurements directly compare how two direction estimators affect behavior. A

Table 15: Probe and difference-of-means directions produce different intervention responses. Bias range is positive-pole response probability across the sweep; oracle is the best gold-signed injection result; removal projects out the direction, with random-direction controls in parentheses. Entries use percentage units.

Axis	**direction**	**bias range**	**oracle**	**ablate (rand)**
Horizontal	diff-of-means	29 → 72	**95.0**	82.0 (82.4)
	probe	47 → 57	87.8	82.8 (82.4)
Vertical	diff-of-means	55 → 85	**73.9**	**50.0** (58.6)
	probe	64 → 77	65.4	57.7 (58.6)
Depth	diff-of-means	57 → 28	46.4	36.6 (31.4)
	probe	48 → 43	34.6	31.4 (31.4)

preliminary vertical-axis experiment compared 300 tuning steps on $|d'|$-localized neurons, random neurons in the same layer band, and the full band. Localized and random-neuron tuning produced similar vertical recovery; the localized run recorded less horizontal degradation. This comparison provides a qualitative observation; the available summary omits the optimization settings and per-run estimates needed for a quantitative comparison.

E Evaluation Scope and Limitations

Image dependence. Real-minus-blank accuracy is a net effect of image replacement. Item-level transition analysis can reveal prediction changes hidden by that average. Blank, noise, and mismatched images perturb inputs differently: vertical balanced accuracy, for example, falls by 9.3 points under patch scrambling. Grounding Arbiter results are therefore reported through paired scores and intervention gains for each condition.

Probe scope. The primary grouped nested estimate is available for Qwen2.5-VL-7B. The archived layer profiles and eight-model scatter use the earlier item-level protocol, in which folds can share scenes and the reported layer/locus is selected by its CV score. These profiles remain exploratory. The rerun selects text-pooled vertical features; their decodability can reflect both question and image information. Grouped evaluation does not by itself identify which input supplies the decoded signal. The reported probe results omit per-fold predictions, confidence intervals, and complete inner-selection settings.

Intervention selection. The amplification operator and real/gray pairing are specified in Appendix A. Its strength was selected on the reported evaluation set, so the paired gains describe an exploratory operating point. The three-model directional intervention comparison uses one pinned seed; its corrective oracle is gold-conditioned. The available intervention summaries omit the exact calibration-item membership and complete sweep configurations. Correction comparisons retain their own baselines and selection procedures. The grouped SEA/SADI and LoRA/ReFT rows do not separate the constituent implementations.

Run-specific baselines. The capability summary reports H/V/D binary scores of $81.3/58.2/30.7\%$. The paired amplification summary uses raw $82.4/58.5/31.4\%$, and its balanced image-control scores are $82.6/69.8/31.6\%$. The seed-0 directional-intervention depth baselines are $32.7/41.2/30.7\%$ for Qwen, Pixtral, and InternVL3-8B. The Qwen correction baseline is 26.2%; a separate InternVL3-8B correction run starts at 26.4%. Qwen's separate vertical correction baseline is 69.0%. Each entry belongs to a distinct evaluation. Gains are interpreted relative to the baseline of that evaluation; raw and balanced scores also use different weighting rules. The 28.1% front/back phrasing score also uses a different answer-reading evaluation.

Archived comparisons. The original probe–image scatter uses separate summaries: its Qwen vertical/depth gaps are $+1.2/-15.6$ points, versus $-0.5/-14.7$ in the paired table. The prior-lift panel uses a separate condition summary and serves as an exploratory visualization. The Qwen3 recalibration gains of 6.0 and 0.0 in Tables 12 and 13 correspond to separate summaries; the run log distinguishes strengths -2.5 and -2. Method and seed summaries should be compared within their own runs.

Correction selection and uncertainty. The correction-selection rule selects the least complex eligible rung within 0.05 of the best lift, subject to a 0.05 gain floor and an inversion gate. The rule was revised during analysis and is retrospective. Under the final rule, InternVL3-8B selects none: its 0.048 rotation gain is below the floor. It is a reproducible selection among tested candidates. Most correction results are single-seed; the Qwen3 three-seed summary reports greater variation for rotation than for trained editing.

Coverage and task scope. Fourteen-model behavior, three-model directional control, four-model paired depth task comparisons, and the single-model grouped probe and amplification study provide different coverage. The original figure's regime labels summarize exploratory patterns using post-hoc thresholds. Quantitative claims use the reported scores and contrasts. Benchmarks differ in scenes, questions, and relation definitions; their comparisons establish task dependence. Annotation and perspective effects require matched examples to separate. The evaluations use 4-bit weights, and within-family size trends are observational.

F TERMINOLOGY AND MEASUREMENT GUIDE

The paper combines representation, intervention, and input measurements. Table 16 collects their definitions. In particular, image ablation changes the input, while direction ablation changes an internal activation. Behavioral inversion and reversed directional response refer to separate measurements that can be compared on the same task.

Table 16: Terms and operational definitions used in this paper.

Term	Operational meaning
Decodability	Prediction accuracy of a specified probe under its stated split, preprocessing, and layer/locus selection protocol.
Directional control	Change in answer probabilities or accuracy produced by a specified internal direction edit.
Controllability	The recorded behavioral response measure for a specified sign, strength, and layer; the layer study reports a one-sided measure at $\beta = 4$.
Image dependence	The change in scores under matched versus replaced images, holding questions, options, and decoding fixed.
Image ablation	Replacing the matched image with a blank; other replacement controls test different changes to image content.
Direction ablation	Removing the component of an activation along a unit direction: $h \leftarrow h - (h^\top \hat{d})\hat{d}$.
Behavioral inversion pattern	Below-chance matched-image accuracy together with $\Delta = A_{\text{real}} - A_{\text{blank}} < 0$.
Reversed directional response	A decrease in the positive-pole answer probability as steering moves toward a calibrated positive-pole direction.
Blank lift	$L = A_{\text{blank}} - 1/2$ for balanced binary accuracy.
Ablation-persistent intervention gain	A positive gain G_{blank} retained when the same intervention is evaluated with blank input; report it beside G_{real}.
Intervention–grounding dissociation	The observed separation between an edit's behavioral success and its dependence on the matched image, quantified through the paired gains.
Corrective oracle	Gold-conditioned injection that uses the item's true pole and the measured response sign to choose a correction direction.
Grounding Arbiter	Paired input evaluation plus repeated intervention under matched and blank images (Algorithm 1).