COUNTERFACTUAL TESTS REVEAL LIMITED DEPTH-IMAGE USE IN RGB-D VLMS: A PROBING AND CAUSAL-PATCHING DIAGNOSIS ON METRIC-DISTANCE QUERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models are routinely given depth maps, but whether they condition on the depth input is unclear. We introduce **CF-Depth**, which holds the RGB image and question fixed while editing the depth map to flip the correct answer; RGB-only shortcuts then fail by design, so above-chance performance requires using depth. Across ordinary RGB-D fine-tuning and three off-the-shelf depth-VLMs, answers stay largely invariant to depth edits and shuffling. The representation is present but unused: linear probes recover metric depth from the depth-image tokens, yet activation patching shows these tokens causally affect answers only after counterfactual training, pointing to learned *read-out*, rather than perceptual encoding, as the bottleneck. Per-image depth normalization removes a non-redundant signal, absolute metric scale, and depth use then returns in stages. Fixed-scale depth with counterfactual training first elicits a *global* metric-scale read-out. A 2B VLM answers counterfactual distance queries well above chance (bin-logit 0.822 seven-seed mean vs. a 0.445 depth-blind floor, our primary read-out; free-generation 0.62 vs. 0.40 RGB-only, chance 0.33), while the controls stay at the depth-blind floor. A stricter local counterfactual that keeps the 256-bin depth histogram exactly fixed then gives evidence for a spatially localized read-out beyond global statistics: the local-CF model stays above chance, exceeding the global-statistics and RGB-only baselines, and cell-restricted patching supports a causal role for the edited cells. This local read-out is in-domain, while the coarser scale read-out transfers to held-out real sensor depth. For these metric-distance queries, a depth image is thus not sufficient by itself: depth use depends on scale-preserving encoding and a training signal that elicits read-out.

1 INTRODUCTION

A growing line of work augments vision–language models (VLMs) with a depth map to improve spatial reasoning. The implicit assumption is that, given depth, the model will *use* it. This assumption is rarely tested: standard evaluations report that adding depth improves accuracy, but such correlational gains can come entirely from the RGB image, whose monocular cues (perspective, familiar object size, vertical position) already encode much of the scene's spatial structure. A model can therefore appear "depth-aware" while ignoring the depth input.

We make this question testable. We introduce **CF-Depth**, a *counterfactual* benchmark in which an identical RGB image is paired with a depth edit that *flips* the ground-truth answer. Because RGB is held fixed while only depth determines the answer, RGB-only performance is bounded at chance by construction, and any above-chance accuracy must come from reading depth. A shuffle control (spatially permuting the depth) further tests that above-chance accuracy reflects depth content rather than an artifact.

A natural first test (shuffling or removing the depth) often leaves answers unchanged, which is suggestive of non-use but inconclusive on RGB-answerable questions; CF-Depth resolves the ambiguity, since RGB-only accuracy is bounded at chance and above-chance accuracy therefore requires

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.

depth. Yet a linear probe recovers per-cell metric depth from the depth-image tokens under leakage-free scene-grouped CV ($R^2 \approx 0.55$–0.70 at intermediate layers), *even in the very models that ignore depth*. Depth is decodable from the image tokens, but those tokens are causally inert for the answer under our patching tests; the failure is one of *use/read-out*, not representation. This is not an artifact of our own training: three published, purpose-built depth-VLMs (SpatialBot, RoboRefer, SD-VLM) are likewise invariant to the depth input on CF-Depth without fine-tuning, run outside their native eval distribution (§4).

We then ask why. On these indoor scenes, monocular depth structure is already decodable from the RGB tokens ($R^2 \approx 0.86$, scene-grouped within-dataset), so a depth map is largely *redundant* for *relative* judgments. What depth adds that RGB cannot supply (on tasks that need it) is *absolute metric scale*; the standard per-image normalization of depth maps discards exactly this. (We do not claim absolute scale is depth's *only* non-redundant signal; occlusion and per-object geometry also matter; it is the operative one for the absolute-distance family we study.) On an absolute-distance task, all encodings (RGB-only, normalized depth, even fixed-scale depth under ordinary training) collapse to a language prior over object identity and are invariant to the depth content.

Finally, a *scale-preserving* (fixed) encoding *together with* counterfactual training lets a 2B model read absolute distance above chance (bin-logit 0.822, seven-seed mean, vs. a 0.445 depth-blind floor; free generation 0.62 vs. 0.40 RGB-only, chance 0.33) while normalized-depth, RGB-only, and depth-removed controls stay at the depth-blind floor. In our runs, neither ingredient alone recovered this use.

Three results are central: the audit that ordinary training ignores the depth *image* (Table 1); the remedy that a scale-preserving encoding with counterfactual training exceeds chance (Table 4); and that this read-out becomes spatially localized under local counterfactuals (App. H). The probe, patching, shuffle, transfer, and seed analyses localize and stress-test these. Our claims are bounded to small (2–9B) VLMs and absolute-distance, metric-scale queries.

Contributions. (i) **CF-Depth**, a counterfactual, shuffle-controlled benchmark whose paired (base+counterfactual) metric is bounded at chance *by construction*, making "does the model use the depth input?" testable (and, in principle, modality-agnostic). (ii) A causal localization of the failure to *read-out* rather than perception (linear probe + activation patching), on our dual-image/4-channel variants and off-the-shelf depth-VLMs (Table 1), replicated on SmolVLM2; the behavioral audit further replicates on a trained InternVL3-2B. (iii) A diagnosis paired with a constructive result. Relative depth is RGB-redundant on these scenes, so for the absolute-distance task we study, the depth map's non-redundant contribution is absolute scale, which per-image normalization discards; a scale-preserving encoding + counterfactual training then recover depth use (seven seeds; neither ingredient alone suffices). (iv) Evidence that the recruited read-out's *scope* is controllable: global-rescale counterfactuals induce a global metric-scale anchor, while adding histogram-matched local counterfactuals induces an in-domain, spatially more localized read-out (with non-synthetic real-sensor corroboration, App. F).

2 Related Work

Depth-augmented VLMs. Many systems add depth to improve spatial reasoning, via a second image, early fusion, region depth tokens, a dedicated depth encoder, or depth-derived text prompts (Cai et al., 2025a; Cheng et al., 2024; Chen et al., 2025; Liu et al., 2025; Yu et al., 2026; Zhou et al., 2025; Chen et al., 2024; Daxberger et al., 2025; Roy et al., 2025; Zhang et al., 2026). These report gains but do not falsifiably establish that the model conditions on depth rather than correlated RGB cues: SD-VLM (Chen et al., 2025), e.g., adds depth positional encoding yet normalizes depth (discarding absolute scale) without separating perception from use. DepthLM (Cai et al., 2025b) shows VLMs reach strong *metric* depth estimation through text-based supervision alone, complementing our finding that a text pathway is used immediately (Table 1); that normalization discards scale is a standing theme in metric-depth estimation (Bhat et al., 2023); our contribution is the diagnosis that depth-VLM *input encodings* inherit it and that it causally gates whether the depth *image* is used. Depth maps are commonly encoded as HHA-like or per-image-normalized images (Gupta et al., 2014; Chen et al., 2025): whether absolute scale survives is encoding-dependent: HHA's

height-above-ground channel is gravity-referenced and absolute, whereas the per-image percentile normalization common in recent depth-VLMs discards it outright.

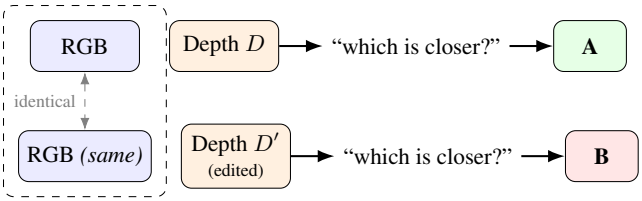
Spatial-reasoning benchmarks. Existing benchmarks (Liu et al., 2023; Shiri et al., 2024; Jia et al., 2026) measure spatial accuracy but allow monocular shortcuts; a counterfactual construction is what makes depth use *necessary*. Our paired design follows the logic of benchmarks that bound blind baselines at chance by construction (Agrawal et al., 2018; Li et al., 2024), applied here to the depth modality. We build on real RGB-D sources (Silberman et al., 2012; Armeni et al., 2017).

Privileged information and distillation. Learning from a privileged (depth) teacher relates to generalized distillation and LUPI (Lopez-Paz et al., 2016; Vapnik & Vashist, 2009; Hinton et al., 2015; Guo et al., 2025); output-level distillation from a depth-text teacher does not transfer depth use here (§8).

Probing, and read-out vs. representation. Linear probes test what representations encode; recent work probes depth-decodability from VLM tokens to motivate *better encoders* (Chen et al., 2025), aligns VLM features with 3D foundation models (Li et al., 2025a), and locates VLM bottlenecks in alignment or read-out rather than representation (Vompa et al., 2025; Nooralahzadeh et al., 2026); Fu et al. (2025) find VLMs underuse representations that stay decodable throughout (Tong et al., 2024); El Banani et al. (2024) probe the 3D awareness of vision encoders; and BLINK benchmarks fine-grained visual perception in MLLMs (Fu et al., 2024). Because probe decodability does not establish use (Belinkov, 2022), and following amnesic-counterfactual logic (Elazar et al., 2021), we use probes for the opposite purpose: to show depth is decodable *yet causally ignored*, localizing the failure to read-out. Concurrent work (Maheshwari et al., 2026) localizes a spatial deficit in a larger model (Qwen3-VL-8B) to the visual-token merger and patches in clean post-merger activations; our locus and remedy differ: with the vision encoder *and* merger frozen, a train-side counterfactual signal recovers depth use, placing the depth bottleneck in learned read-out, not an inference-time patch. Our mechanism experiments use Qwen3.5-2B (Qwen Team, 2026), with the headline remedy and its causal localization replicated on SmolVLM2 (Zohar et al., 2025), and the behavioral audit on an InternVL3-2B we *train* (Zhu et al., 2025) (depth-blind under ordinary training, depth-using after counterfactual training; App. G).

Counterfactual and intervention-based evaluation and training. Counterfactual edits that flip the answer while holding the rest of the input fixed have been used to test reliance on RGB priors and hallucination (Golovanevsky et al., 2025; Li et al., 2025c;b). We bring this paradigm to the *depth modality*: by pairing identical RGB with an answer-flipping depth edit, CF-Depth bounds RGB-only at chance *by construction*, a guarantee absent from existing spatial benchmarks that permit monocular shortcuts (Jia et al., 2026; Huang et al., 2026; Yu et al., 2025). ("Counterfactual depth" has separately named object-removal depth prediction (Issaranon et al., 2019), a different task from our depth-input edits.) Concurrent spatial counterfactuals instead edit *scene content* such as occlusion (Maheshwari et al., 2026); ours isolates the *modality* (identical scene, edited depth channel), a direct test of whether the depth *input* is read. On the training side, counterfactual fine-tuning has been used to *enhance* visual grounding (Zhang et al., 2025a), and inference-time steering can shift which modality a model relies on (Zhang et al., 2025b). For the depth modality, inference-time steering is complementary to counterfactual training and untested here.

Measuring and balancing modality use. Whether a model uses a given input is itself studied: MM-SHAP (Parcalabescu & Frank, 2023) and the perceptual score (Gat et al., 2021) quantify per-modality contribution by permutation/attribution, and blind-baseline and cross-modal-influence analyses show VL models often solve tasks from one stream (Frank et al., 2021); modality competition further explains why a stronger stream suppresses a weaker one (Wang et al., 2020; Peng et al., 2022), locating under-use in *optimization*, and Yu et al. (2025) discuss input-modality trade-offs for spatial reasoning. Beyond these metrics, our contribution is a *by-construction* chance bound specific to depth (CF-Depth), made causal by patching, which locates under-use in *read-out*, with the fix a *training signal* (counterfactuals), consistent with RoboRefer (Zhou et al., 2025), which obtains depth use via a reinforcement (forcing) signal.



Base and counterfactual share **identical RGB** but have **opposite** answers ⇒ any RGB-only predictor is wrong on one of each pair ⇒ **RGB-only ≤ chance by construction**. Above-chance accuracy implies depth use.

Figure 1: CF-Depth construction. Editing only the depth (so the near/far answer flips) while holding RGB and the question fixed bounds RGB-only at chance; a shuffle control further tests depth-content dependence.

3 CF-DEPTH: A COUNTERFACTUAL TEST OF DEPTH USE

Construction. We build CF-Depth from indoor RGB-D scenes with real sensor depth (NYU Depth v2 and 2D-3DS). Each item is a spatial question (e.g. "which region/object is closer to the camera?") with a metric ground truth derived from the depth. For a *base* item we additionally synthesize a *counterfactual* partner that shares the *same RGB* but has the depth modified so that the correct answer *flips*: an image-level edit (CF_SYNTH, swapping/relocating depth so the near/far ordering reverses) for depth-image models, and a value-level edit (CF_TEXT) for depth-as-text models (Figure 1).

The RGB-only guarantee. Because base and counterfactual share an identical RGB but have opposite answers, any function of RGB alone must be wrong on exactly one of the two; balancing them bounds RGB-only accuracy at chance *by construction*. Empirically an RGB-only model scores below chance on CF-Depth (CF-combined 0.31), unlike on natural spatial benchmarks where monocular priors leak. Above-chance accuracy on CF-Depth therefore requires using depth.

Controls and metrics. We report *cf-sensitivity*, the gap between accuracy on real vs. spatially *shuffled* depth: a model that genuinely reads depth collapses under shuffling, whereas one keying on a depth-image *signature* does not. We also report *CF-combined* accuracy over base+counterfactual pairs (chance $= 0.5$ for binary ordering). Shuffling is applied both to base and to counterfactual depth, giving a clean artifact check. The audit (§4) uses a binary ordering instance (chance $\frac{1}{2}$); the probe/patching/remedy (§5–7) use a three-way absolute-distance instance (chance $\frac{1}{3}$). A clarification: the $\leq$ chance bound is a property of the *paired* (base+CF) metric (any RGB-only predictor is wrong on one member of each pair, so paired accuracy $\leq \frac{1}{2}$). On the *counterfactual members only* (e.g. Table 4), $\frac{1}{3}$ is the random-guess baseline, not a by-construction ceiling; there depth use is established by the depth-removed/RGB-only/shuffle/patching controls.

4 ORDINARY TRAINING SHOWS LITTLE IMAGE-DEPTH USE ON CF-DEPTH

Image depth is shuffle-invariant under ordinary training. In every ordinary-training row that feeds depth through the *image* pathway (including 4-channel fusion with the full vision tower fine-tuned), shuffling or removing the depth leaves answers essentially unchanged (cf-sensitivity ≈ 0; Table 1), so above-chance CF-Depth accuracy cannot be a shuffle artifact. Shuffle ablation, however, is suggestive rather than decisive: shuffled depth can even *improve* accuracy, and on natural questions RGB monocular cues leak.

CF-Depth removes the RGB shortcut. CF-Depth removes this ambiguity (§3): where above-chance accuracy *requires* reading depth, our dual-image and 4-channel models sit at chance (cf-sensitivity indistinguishable from zero, CF-combined at chance), and RGB-only is *below* chance. Table 1 consolidates the audit.

Text *is* used; the image is not. The same depth supplied as metric *text* yields a large cf-sensitivity (≈ 0.43) with general VQA preserved: identical information is usable through the language stream but ignored through the image stream. Counterfactual *training* is the only signal we tested that

Table 1: CF-Depth audit across input pathways and training regimes ($n = 1136$ unless noted). cf-sensitivity = accuracy with real − spatially shuffled depth on the same items (95% CI; off-the-shelf rows use the cleanest available intervention: real − removed for SpatialBot/RoboRefer, real − shuffled for SD-VLM, App. K); CF-combined = paired accuracy over base+counterfactual items (chance 0.5). Depth through the *image* pathway is ignored under ordinary training in every variant (including 4-channel early fusion with the full vision tower fine-tuned), while the same metric values as *text* are used immediately; only counterfactual training makes the image pathway depth-sensitive, at a general-VQA cost that a reduced learning rate (1×10^{-4}; this row's default is 2×10^{-4}) largely recovers. Below-chance CF-combined values (e.g. 0.21, 0.39) reflect the model following an RGB/object prior that is systematically wrong on the counterfactual partner, not depth use.

depth input	training	cf-sensitivity	CF-combined	general F1
dual image (9B)[†]	ordinary	-0.02 (.23 vs. .25)	0.21	—
4-channel, frozen ViT	ordinary	-0.002 $[-.026, .022]$	0.39	—
4-channel, full ViT FT	ordinary	-0.003 $[-.009, .004]$	0.50	—
depth as text	ordinary	$+0.430$ $[.401, .460]$	0.97	0.51
depth as text	counterfactual	$+0.265$ $[.232, .298]$	0.73	—
dual image[¶]	counterfactual	$\approx+0.7$	—	0.15 (0.51 low-lr)
SpatialBot-3B[‡]	published (no FT)	$+0.02$ (.405 vs. .383)	—	—
RoboRefer-2B-SFT[§]	published (no FT)	-0.04 (.470 vs. .511)	—	—
SD-VLM-7B[‖]	published (no FT)	$+0.01$ (.339 vs. .327)	—	—
RGB-only (validity)	—	—	0.31 (≤ 0.5)	—

[†] $n = 200$, free-form parse 0.47: a qualitative consistency check. [¶] The 2B dual-image counterfactual row is likewise qualitative (free-form parsing limits precision; general F1 at default and reduced lr). [‡] Off-the-shelf, no FT ($n = 5{,}592$): SpatialBot *reads* depth (Spearman 0.59) but is invariant ($.405$ vs. $.383$ removed). [§] RoboRefer-2B-SFT (released SFT; $n = 1{,}136$): invariant ($.470$ vs. $.511$). [‖] SD-VLM-7B (no FT, $n = 2000$): invariant (92% identical real/shuffled). All: App. K.

makes the image pathway use depth (cf-sensitivity ≈ 0.7), at a general-VQA cost a reduced learning rate largely recovers (Table 1; §7). The effect extends beyond our models: three off-the-shelf depth-VLMs (SpatialBot-3B, RoboRefer-2B-SFT, and SD-VLM-7B) show the same behavior without fine-tuning: SD-VLM, the model our normalization analysis most directly predicts to fail, is invariant to depth content (cf-sensitivity 0.01; released-checkpoint caveat, App. K), and SpatialBot *reads* depth (Spearman 0.59) yet is invariant to it (Table 1, App. K).

5 Perception vs. Use: Probing and Causal Patching

To separate *perceiving* depth from *using* it, we fit ridge regressions from the depth-image token representations (pooled to a spatial grid) to per-cell depth. Under leakage-free *scene-grouped* 5-fold CV against *absolute metric* depth (the cross-scene–comparable target), depth is decodable at every intermediate layer of the base model ($R^2 = 0.54$ at the input embedding, 0.58 by layer 18; fold SD ≤ 0.10), and comparably so in models that use depth (CF: 0.70 at layer 12) and models that ignore it (no-CF: 0.63): the information is present and retained regardless of whether the model acts on it. The modest CF advantage (≤ 0.07) is small relative to the *causal*-use gap (patching changes $\sim$0.76 vs. $\leq$0.01 of answers, below). The final, answer-specialized layer collapses (near zero in both fine-tuned models), confirming intermediate layers are the fair measure of encoded content. (A within-dataset probe against the *normalized* depth gives a higher nominal $R^2 \approx 0.84$, Table 2, but collapses under scene-grouped CV.)

The probe target is not merely the (normalized) input image: regressing instead to *absolute metric depth in meters* (from the raw sensor, within-dataset split) is also highly successful ($R^2 = 0.75$ at the input embedding, 0.96 at the last layer), so the representation genuinely carries metric depth, not just reconstructable input texture. (We attribute the absolute-scale effect of §7 to *usage* under counterfactual rescaling, not to representational decodability, which a within-dataset probe cannot cleanly separate.)

Two controls sharpen the reading. (i) Depth is also decodable from the *RGB* tokens ($R^2 = 0.86$ under scene-grouped CV), evidencing a monocular RGB shortcut (Geirhos et al., 2020) that makes the depth map redundant for relative structure; this redundancy is within-distribution, as the lin-

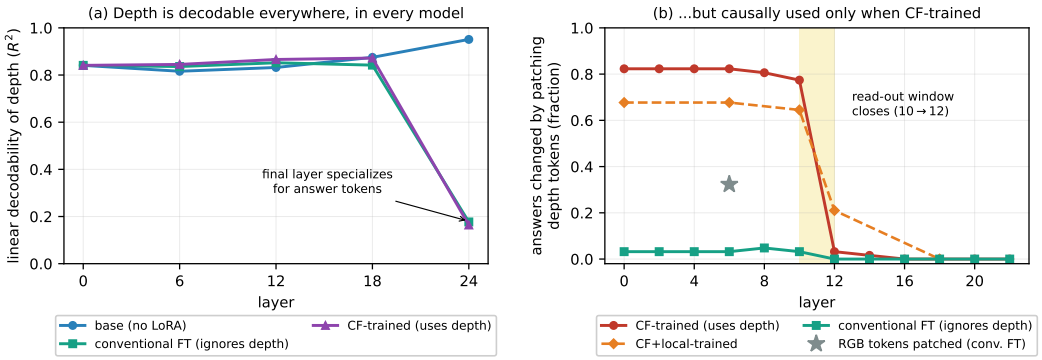


Figure 2: **Layer-resolved probing and patching.** Depth is decodable but not causally used. (a) Linear probes decode per-cell depth from the depth-image tokens at every intermediate layer, *identically* in models that use depth (CF-trained) and models that ignore it (conventional); only the answer-specialized final layer of fine-tuned models collapses (scene-grouped R^2, Table 2). (b) Activation patching of the same tokens ($n = 62$ pairs) shows they causally drive the answer *only* in counterfactually-trained models, and only through the early layers (flat $\approx$0.8, vanishing by layer 12, shaded; CF+local retains influence at layer 12). The conventional model is causally inert at every layer, yet patching its *RGB* tokens changes 32% of answers (star); the intervention works and RGB is used; depth, specifically, is perceived but not read out.

Table 2: Linear-probe R^2 for decoding per-cell *absolute metric* depth from depth-image tokens, by layer, under leakage-free *scene-grouped* 5-fold CV (the rigorous, cross-scene-comparable measure; fold SD ≤ 0.10). Intermediate layers are decodable *equally* whether the model uses depth (CF-trained) or ignores it (no-CF); only the final, answer-specialized layer of fine-tuned models collapses. Layer-0 equality across rows is structural (frozen encoder, LM-only LoRA); the informative comparison is at the intermediate LM layers. For reference, a within-dataset random-split probe against the *normalized* depth image gives inflated values ($R^2 \approx 0.84$–0.88 at intermediate layers, 0.95 for the base model, collapsing to 0.16–0.18 at the final layer of fine-tuned models).

layer	0 (embed)	6	12	18	24 (final)
base (no LoRA)	0.54	0.56	0.56	0.58	0.37
no-CF (ignores depth)	0.54	0.58	0.63	0.63	0.07
CF-trained (uses depth)	0.54	0.59	0.70	0.69	0.06

ear map does not transfer NYU→2D-3DS ($R^2 \approx 0.10$, App E). (ii) Spatially shuffling the depth roughly halves layer-0 decodability (normalized probe, $0.84 \rightarrow 0.53$), a control-task-style selectivity check (Hewitt & Liang, 2019) confirming the signal is content-dependent. Together these show the bottleneck is *read-out* rather than perception. The contrast is not circular: counterfactual training adds only marginal decodability (~ 0.06 at intermediate layers, far too small to explain the causal-use gap); it conditions the answer on an already-present signal that conventional training leaves causally inert, even with the vision tower fine-tuned (Table 1). As the encoder is frozen, the change is in read-out. This recruited read-out is not a memorized inversion of the training edit: it transfers to a held-out real sensor and task absent from training (2D-3DS; §8, App. E) and to rescaling magnitudes never seen in training (App. I), indicating a learned absolute-scale read-out rather than a counterfactual-specific artifact.

Causal evidence: activation patching. To make the read-out claim causal rather than inferential, we apply activation patching (causal tracing and path patching (Vig et al., 2020; Meng et al., 2022; Goldowsky-Dill et al., 2023; Wang et al., 2023; Zhang & Nanda, 2024; Geiger et al., 2021), ported to VLMs by Palit et al. (2023)) to the depth-image *token activations* at a single layer: on base/counterfactual pairs of the absolute-distance task (identical RGB, question, and object name; opposite answers), we run the counterfactual input but replace the depth-token hidden states at layer ℓ with those captured from the base input, and read the answer-bin logit (Table 3). In the

Table 3: Activation patching on base/counterfactual pairs ($n = 62$): fraction of answers changed by patching base depth-token activations into the counterfactual forward pass, by layer. In the counterfactually-trained model the depth tokens causally drive the answer with a sharp cutoff between layers 10 and 12: depth-token influence is confined to an early-layer window, and downstream of it the answer no longer depends on these positions. The conventionally trained model's answers are causally independent of its depth tokens at *every* layer, even though depth is equally decodable from them at every layer (Table 2): the decodability-use dissociation, layer-resolved. The profile reproduces on an independent converged seed (layers 0–10: 0.68–0.74; layers $\geq$12: 0.00). The per-layer values are the $n = 62$ profile; the bottom row gives the keystone layer-6 numbers over all 608 matched pairs (0.81 [0.78, 0.84] full-ViT). Protocol in Appendix C.

layer	0	2	4	6	8	10	12	14	16–22
CF-trained (uses depth)	.82	.82	.82	.82	.81	.77	.03	.02	.00
conventional (ignores)	.03	.03	.03	.03	.05	.03	.00	.00	.00

keystone L6, 608 pairs: CF 0.76 [0.72, 0.80] (6 converged; all-11 mean 0.42) vs. conventional 0.008 [0.002, 0.018]

counterfactually-trained model, a scene-clustered re-run over all 608 matched base/CF pairs shows layer-6 depth-token patching *changes the answer on* 0.76 *of pairs* (six of eleven trained checkpoints converge to depth use and five do not, so this is a conditional-on-convergence mean; unconditional 0.42; over converged seeds [0.72, 0.80], per-seed 0.72–0.83, App. B; 0.81 full-ViT) *and flips it on* 0.71: the depth tokens causally drive the answer. In the conventionally-trained model the very same intervention does nothing ($\leq 5\%$ changed at any layer): its answer is causally independent of the depth tokens, even though depth is equally decodable from them (Table 2). A positive control validates the intervention itself: patching the *same* conventionally-trained model's *RGB*-token activations with those of a different scene changes 32% of its answers; the hook is effective and RGB is causally used; depth specifically is perceived but not read out. The effect is content-specific: a *same-answer-bin* depth from another scene changes only 37% of answers, while an answer-flipping patch from another scene changes 87% [0.79, 0.92] ($n = 100$ each; both arms use foreign-scene donors), which with shuffle invariance and held-out-sensor transfer (App. C, E) argues against an edit-detection artifact. This is the decodability/use dissociation made causal; Figure 2 juxtaposes the two layer profiles, and the intervention is symmetric and sharply layer-localized (Table 3).

6 Absolute Scale: The Non-Redundant Depth Signal

If relative depth is already in RGB (§5), the principal non-redundant signal a depth map can add *on this task* is *absolute metric scale*. Yet the standard encoding (per-image percentile normalization to [0, 255]) rescales every map to the same range and thus *discards absolute scale*. To test the consequence we use CF-Depth's *absolute-distance* instance (three-way: near/medium/far against fixed metric thresholds), for which RGB lacks the decisive signal: monocular estimation recovers *relative* structure well (Ranftl et al., 2022), while metric scale requires injected priors (Bhat et al., 2023). In each counterfactual pair the RGB and question are identical and only the depth is rescaled so the correct answer flips (near $\leftrightarrow$ far); RGB-only is thus bounded at chance and an object-identity prior ("chairs are usually nearby") is actively misleading on the counterfactual member, so above-chance paired accuracy requires depth use.

Zero-shot, every model is at chance ($\frac{1}{3}$); after fine-tuning every encoding reaches ≈ 0.63–0.65 via a *language prior over object identity*, invariant to depth shuffling (0.637 vs. 0.638) and matched by the depth-free RGB-only model.

7 Scale Preservation and Counterfactual Training

Setup. Unless noted, mechanism experiments use Qwen3.5-2B with LoRA (rank 16, $\alpha = 32$, lr 10^{-4}, two epochs) on the language model only; the vision encoder is frozen. We replicate the headline result on a second architecture, SmolVLM2-2.2B, and at larger scale on Qwen3.5-9B, with identical LoRA hyperparameters and data (two epochs). The absolute-distance task asks for one of three bins (near < 1.9 m $<$ medium < 3.0 m $<$ far) about a named region; fixed-scale encoding

Table 4: Accuracy on the absolute-distance task (3-way, chance $= 0.33$). *Ordinary* training reaches ≈ 0.64 via an object-identity language prior, shuffle-invariant; under *counterfactual* training only *fixed-scale* depth reads absolute distance. Counterfactual numbers are on counterfactual examples, where the language prior is actively misleading ($n = 5{,}592$ ordinary; $4{,}406$ counterfactual). Within-run 95% Wilson CIs are within ± 0.015; the fixed+CF cell is seven seeds (bin-logit between-seed 95% CI $[0.809, 0.835]$, SD 0.014; free-gen $[0.54, 0.70]$, SD 0.08), normalized+CF three, other controls single runs. *Bin-logit* is the primary, parse-independent forced-choice readout; *free-gen* is a conservative, parse-limited lower bound (App. C). Only fixed-scale rises above the depth-blind floor (≈ 0.44–0.445); depth-removed's bin-logit (0.39) is the no-depth floor. The final row is a depth-image-free logistic regression over global depth statistics: it matches the model on this global-rescale split, so this table certifies a *global scale* read-out only; spatially resolved use is isolated by the histogram-fixed two-cell task (Table 5).

input	ordinary training	counterfactual training	
		free-gen	bin-logit
RGB-only	0.63	0.40	0.43
depth removed (gray)	—	0.36	0.39
normalized depth	0.65	0.39	0.44
fixed-scale depth	0.64	**0.62**	**0.822**
global-statistics logistic (no image)	—	0.69	—

uses $D_{\max} = 8000$ mm (8-bit linear, ≈ 31 mm/bin, which amply resolves the 1.1 m gap between the two decision thresholds though it coarsens near-range resolution). Counterfactual training/eval sets contain 22,380/5,592 examples (the 5,592-example eval split comprises 4,406 answer-flipping counterfactuals and 1,186 ordinary items); the histogram-matched local counterfactuals add 2,472/411, generated from image-disjoint train/val splits. Evaluation uses a parse-independent forced-choice bin-logit readout as the primary metric, with greedy free-generation accuracy as a conservative, parse-limited secondary; full protocol details are in Appendix C.

The diagnosis suggests two requirements for the image pathway to use depth: the input must *contain* absolute scale, and training must *force* the read-out to depend on it. We test both with a 2×2 over encoding (*normalized* vs. *fixed-scale*, pixel $= \mathrm{clip}(\mathrm{mm}/D_{\max}) \cdot 255$) and training (*ordinary* vs. *counterfactual*, where the depth map is rescaled so the answer changes while RGB and the object name are held fixed).

On counterfactual absolute-distance examples (chance $\frac{1}{3}$), only the fixed-scale + counterfactual model reads depth: Removing or normalizing the depth, or dropping to RGB-only, all fall to 0.36–0.44 under both readouts, confirming the gain is genuine depth use rather than a leak; fixed+CF reaches bin-logit 0.822 (seven-seed mean) against the 0.44–0.445 depth-blind floor (normalized+CF three seeds, 0.40–0.49; seed counts and per-seed values in App. B). The $2{\times}2$ isolates the mechanism: among the signals we test, counterfactuals are required to make depth *used*, and scale-preserving encoding yields *substantially stronger* use than normalized depth. The fix does not depend on a frozen encoder: unfreezing the full vision tower under counterfactual training also yields depth use (cf 0.67), while conventional training fails *even with the encoder fully fine-tuned* (Table 1). All seven training seeds exceed chance on counterfactual examples (free-gen cf 0.46–0.72, but bin-logit 0.805–0.844 with SD 0.014: the seed variance lies in generation behavior, not in the learned read-out; per-seed in App. B).

A second architecture, and scale. The fixed-scale + counterfactual result is not specific to Qwen3.5-2B. Across three seeds, SmolVLM2-2.2B (same fixed-scale encoding) reads counterfactual depth (cf 0.799, collapsing to 0.40 under shuffling); patching localizes it early ($n = 64$), while a conventional SmolVLM2 is near-inert (6%), mirroring Qwen. A 9B model reproduces the dissociation (three runs; App. D): highly decodable (probe R^2 0.92–0.99, within-dataset) yet conventionally inert (1.6%), while fixed+CF reads counterfactual depth (bin-logit cf 0.83, shuffling 0.70) through an early-layer causal window (changed-rate 0.87, $n = 64$).

Global and local counterfactuals recruit different read-outs. A histogram-matched local counterfactual (swapping two cells' depth across answer bins while flipping the queried cell) delimits

Table 5: Two-cell swap ($n = 1356$; chance $\frac{1}{3}$; the 256-bin global depth histogram is *exactly* fixed, $L1 = 0$ on all 678 pairs, so any global-statistics read-out answers base and swap identically, right on at most one per pair). Only local-CF tracks the swap (swapped rows 0.578; scene-clustered paired advantage vs. global-CF $+0.157$ $[0.132, 0.182]$; full breakdown in App. H).

	local-CF	global-CF	conventional	RGB-only	global-stats logistic
accuracy	**0.531**	0.374	0.386	0.378	0.406

what is read: the global-rescale model drops to chance on it (0.30–0.32, three seeds) while still solving global-rescale counterfactuals, so its read-out is a *global metric-scale anchor* plus RGB-derived layout, not spatially resolved. Yet the scope *tracks the training CF family*: training that adds local counterfactuals induces a spatially resolved read-out (local-cell accuracy rises from chance to 0.58 on three of four seeds; depth-shuffle control 0.37; App. D) while preserving the global read-out (0.63–0.71). Because a global-statistics-plus-region logistic baseline also reaches 0.57 here, a stricter *two-cell-swap* task removes the ambiguity (Table 5): with the global histogram exactly fixed, a global-only read-out answers each pair identically (right on at most one member), while the local-CF model tracks the swap; cell-restricted patching further supports a causal role for the edited cells ($76\%/53\%$ vs. 5%; App. C). The result is not architecture-specific: SmolVLM2 reads 0.78–0.88 (three seeds) vs. a 0.39 depth-withheld floor—stronger than Qwen on two-cell, weaker on local-cell (App. H). This local read-out is *non-synthetic*: on unedited NYU object-region depth (per-object absolute distance, $n = 2067$, chance $\frac{1}{3}$; App. F), the local-CF model shows the object-local signature—real $0.60 \gg$ shuffle $0.46 \gg$ removed 0.34, a $2.2\times$ larger shuffle drop than the global model—while the conventional model is depth-blind (0.64 with or without depth). Depth use here shows as *reliance* (collapse without depth), not raw accuracy: RGB-only 0.63 exceeds the model's 0.60. The global read-out also shifts monotonically under layer-6 linear steering (App. L).

8 Discussion

Implications for design. Because perception is not the bottleneck, a stronger depth encoder alone is unlikely to induce use; the binding constraints are a read-out-forcing training signal and a scale-preserving encoding. A text depth pathway is already effective for deployment (Table 1; Cai et al., 2025b); the image-pathway remedy shows what it takes for the depth *image* to be used.

Only counterfactual training recruited depth use, among tested signals. Three non-counterfactual routes failed (App. D): output-level KD, an auxiliary depth-readout head, and π-distill privileged-mode alignment (Penaloza et al., 2026), all at cf-sensitivity ≈ 0 (existence-level negatives at a matched budget, not tuned baselines; run counts in App. D). Explicit depth supervision improves *decodability*, not *use*. The recruited read-out also transfers: the CF model stays depth-driven on held-out real sensors (2D-3DS 0.56 vs. 0.34 removed; *human-authored* SpatialRGPT-Bench distance 0.53 vs. 0.33; App. E, J).

Low-cost mixing; limitations. Mixing counterfactuals into ordinary SpatialQA (RussRobin, 2024) training preserves general QA (Token-F1 0.55–0.56 vs. 0.55–0.62) while inducing depth use (cf $0.63/0.74$). *Limitations.* Our models are small (2B, LoRA) and the task is synthetic; the local-CF spatially-resolved read-out does not transfer across sensors, and CF_SYNTH edits can leave residual positional regularity, so the prior-controlled, transfer, held-out-magnitude, and two-cell results are load-bearing (App. J, E, I; Table 5). CF-Depth pits depth against RGB, so positive results are stress tests of *necessary* use under conflict, not cooperative-use or general-failure claims; real-sensor degradations are untested.

9 Conclusion

Adding a depth image does not imply using it: CF-Depth shows depth is perceived yet ignored under ordinary training, and restored by scale-preserving encoding with counterfactual training (globally, then locally). It remains an audit on small VLMs and a synthetic task, not a deployable recipe.

AI USE STATEMENT

In this work, we used generative AI tools to design and provide feedback on research methodology and experiments, to refine hypotheses, to implement methods (data-construction, training, and evaluation code, including the counterfactual-pair construction and statistical-analysis scripts), to clean and reformat datasets, and to assist in interpreting results. The counterfactual training and evaluation pairs themselves were produced by deterministic scripted transformations; generative AI was used to write those scripts but not to generate data content directly. We have not used generative AI tools to develop theoretical models or conceptual frameworks, to formulate mathematical claims or to write proofs, or to assist with translation, and the remaining tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, to draft and edit parts of the manuscript for readability, to prepare figures, to search for and identify relevant literature, and to suggest structure for the paper. We have reviewed all AI-assisted work: the code paths underlying every reported number were verified against per-sample logs (released as supplementary material, together with a script that recomputes the main tables from the raw logs), AI-drafted text was checked against the experimental records, and AI-suggested citations were manually verified against the cited papers. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experiments use public base models (Qwen3.5-2B, SmolVLM2-2.2B, Qwen3.5-9B) and public RGB-D sources (NYU Depth v2, 2D-3DS, SUN RGB-D). LoRA settings, optimizer, learning rates (including the 1×10^{-4} general-VQA-preserving variant of Table 1), epochs, the fixed-scale $D_{\max}$, bin thresholds, and counterfactual train/eval sizes are in §7; the patching, probe, steering, and readout-faithfulness protocols (with per-layer/per-seed details) are in Appendices C–L; per-seed numbers are in Appendix B. We will release, as anonymized supplementary material, the CF-Depth construction scripts, the evaluation/patching/probe/steering harnesses, per-sample prediction logs, and split identifiers. All data sources (NYU Depth v2, 2D-3DS, SUN RGB-D, SpatialQA, SpatialRGPT-Bench) are established research corpora used under their respective research licenses; we redistribute no images, only prediction logs and split identifiers, and the indoor-scene sources carry no personally identifying annotations beyond what the datasets themselves release.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proc. CVPR*, pp. 4971–4980, 2018.

Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2D-3D-Semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth, 2023. arXiv:2302.12288.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. In *Proc. ICRA*, 2025a. arXiv:2406.13642.

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, et al. DepthLM: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025b.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proc. CVPR*, 2024. doi: 10.1109/CVPR52733.2024.01370.

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. SD-VLM: Spatial measuring and understanding with depth-encoded vision language models. In *NeurIPS*, 2025. arXiv:2509.17664.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.01584.

Erik Daxberger et al. MM-Spatial: Exploring 3D spatial understanding in multimodal LLMs, 2025. arXiv:2503.13111.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *CVPR*, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proc. EMNLP*, 2021.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Computer Vision – ECCV 2024*, 2024. doi: 10.1007/978-3-031-73337-6_9.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Perceptual score: What data modalities does your model perceive? In *NeurIPS*, 2021.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching, 2023. arXiv:2304.05969.

Michal Golovanevsky, William Rudman, Michael Lepori, et al. Pixels versus priors: Controlling knowledge priors in VLMs through visual counterfacts. *EMNLP*, 2025. arXiv:2505.17127.

Minghao Guo, Meng Cao, Jiachen Tao, et al. GLaD: Geometric latent distillation for vision-language-action models. *arXiv preprint arXiv:2512.09619*, 2025.

Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *Proc. ECCV*, pp. 345–360, 2014. doi: 10.1007/978-3-319-10584-0_23.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proc. EMNLP-IJCNLP*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning Workshop*, 2015.

Xinmiao Huang, Qisong He, Zhenglin Huang, et al. Spatial-DISE: A unified benchmark for evaluating spatial reasoning in vision-language models. *ICLR*, 2026. arXiv:2510.13394.

Theerasit Issaranon, Chuhang Zou, and David Forsyth. Counterfactual depth from a single RGB image. In *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2129–2138, 2019. doi: 10.1109/ICCVW.2019.00268.

Mengdi Jia, Zekun Qi, Shaochen Zhang, et al. OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. *ICLR*, 2026. arXiv:2506.03135.

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

Fuhao Li, Wenxuan Song, Han Zhao, et al. Spatial forcing: Implicit spatial representation alignment for vision-language-action models. *arXiv preprint arXiv:2510.12276*, 2025a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, 2023.

Shawn Li, Jiashu Qu, Yuxiao Zhou, et al. Treble counterfactual vlms: A causal approach to hallucination. In *Findings of EMNLP*, 2025b. arXiv:2503.06169.

Xinzhuo Li, Adheesh Juvekar, Jiaxun Zhang, et al. Hallusegbench: Counterfactual visual reasoning for segmentation hallucination evaluation. *arXiv:2506.21546*, 2025c. v1 title; later retitled "Counterfactual Segmentation Reasoning".

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.

Yang Liu, Ming Ma, Xiaomin Yu, Pengxiang Ding, Han Zhao, Mingyang Sun, Siteng Huang, and Donglin Wang. SSR: Enhancing depth perception in vision-language models via rationale-guided spatial reasoning. In *NeurIPS*, 2025. arXiv:2505.12448.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *Proc. ICLR*, 2016.

Animesh Maheshwari, Divyansh Sahu, and Nishit Verma. Do vision-language models understand 3D scenes or just catalogue objects? *arXiv preprint arXiv:2605.20448*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, et al. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv:2604.09364*, 2026.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for BLIP. In *ICCV Workshops*, 2023.

Letitia Parcalabescu and Anette Frank. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models and tasks. In *ACL*, 2023.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, et al. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Xiaokang Peng et al. Balanced multimodal learning via on-the-fly gradient modulation (ogm-ge). *CVPR*, 2022. arXiv:2203.15332.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`. Model cards: `https://huggingface.co/Qwen/Qwen3.5-2B` (2B), `https://huggingface.co/Qwen/Qwen3.5-9B-Base` (9B).

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3), 2022.

Rajarshi Roy et al. ByDeWay: Boost your multimodal LLM with depth prompting in a training-free way. In *ICCV Workshops*, 2025. arXiv:2507.08679.

RussRobin. SpatialQA dataset. Hugging Face Datasets, 2024. License: CC-BY-4.0.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 21440–21455. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1195.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proc. ECCV*, pp. 746–760, 2012. doi: 10.1007/978-3-642-33715-4_54.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023. arXiv:2308.10248.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020.

Enrico Vompa, Tanel Tammet, and Mohit Vaishnav. Beyond the linear separability ceiling: Aligning representations in VLMs. *arXiv:2507.07574*, 2025.

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 *ICLR*, 2023.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020.

Hanxun Yu, Xuan Qu, Yuxin Wang, et al. Unlocking dense metric depth estimation in VLMs. *arXiv preprint arXiv:2605.15876*, 2026.

Songsong Yu, Yuxin Chen, Hao Ju, et al. How far are VLMs from visual spatial intelligence? A benchmark-driven perspective. *arXiv:2509.18905*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. SpatialStack: Layered geometry-language fusion for 3D VLM spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jusheng Zhang, Kaitong Cai, Yijia Fan, et al. CF-VLM: Counterfactual vision-language fine-tuning. *arXiv:2506.17267*, 2025a.

Yu Zhang, Jinlong Ma, Yongshuai Hou, et al. Evaluating and steering modality preferences in multimodal llms. *arXiv:2505.20977*, 2025b.

Enshen Zhou et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. In *NeurIPS*, 2025. arXiv:2506.04308.

Jinguo Zhu et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Orr Zohar, Miquel Farré, Andi Marafioti, Merve Noyan, Pedro Cuenca, Cyril Zakka, and Joshua Lochner. SmolVLM2: Bringing video understanding to every device. Hugging Face Blog, 2025. URL `https://huggingface.co/blog/smolvlm2`.

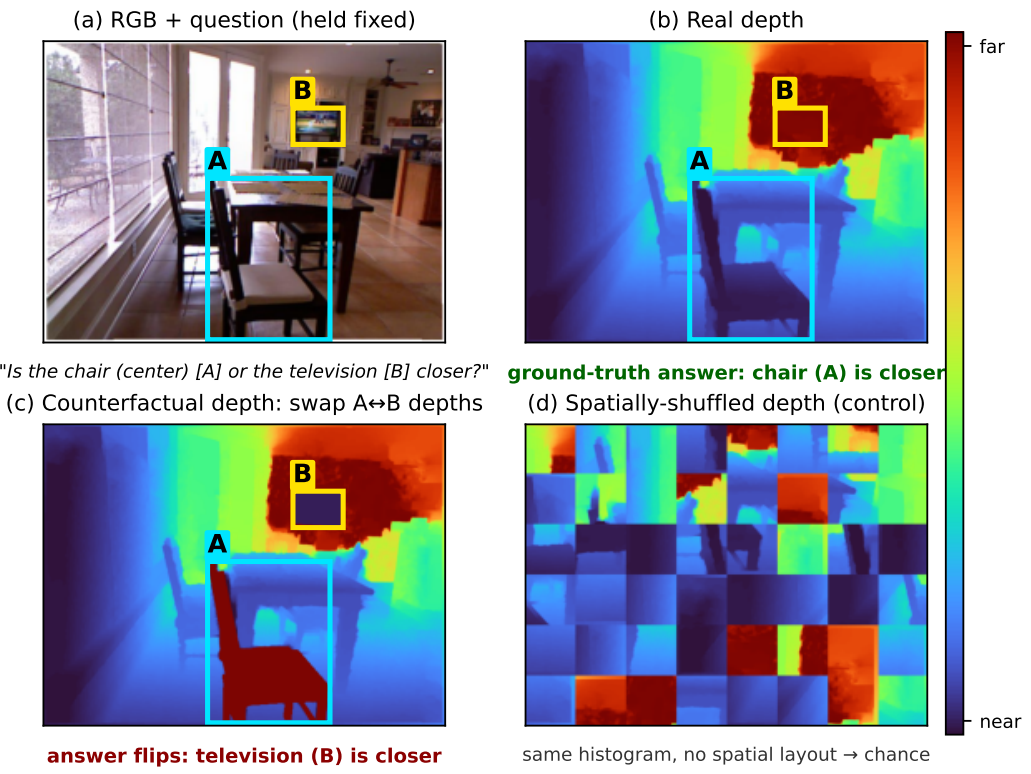


Figure 3: **CF-Depth edit, real-scene example.** A counterfactual depth edit on a real NYU Depth v2 scene. The RGB image and the question are held fixed; only the depth map is edited. **(a)** RGB with the two queried regions, A = chair (center) and B = television. **(b)** The real sensor depth places A nearer than B, so the ground-truth answer is "chair." **(c)** The image-level counterfactual (CF_SYNTH) swaps the depth of the two regions—a 7.6%-pixel edit localized to A and B (boxes)—inverting their near/far order and flipping the ground-truth answer to "television." Because RGB and the question are unchanged, any RGB-only predictor is wrong on exactly one member of each base/counterfactual pair (the construction behind the $\leq \frac{1}{2}$ RGB-only bound). **(d)** The spatially-shuffled control permutes depth blocks, preserving the global depth histogram while destroying spatial layout (the depth-content ablation). Depth is colorized with `turbo` (blue = near, red = far).

A A Counterfactual Depth Edit, Visualized

Figure 3 shows a single CF_SYNTH counterfactual from the object-region family: editing only the depth inverts the near/far relation of the two queried regions and flips the answer, while the RGB image and question are untouched. This is the image-level analogue of the schematic in Figure 1.

B Per-Seed Results

Table 6 reports both base and counterfactual accuracy for every training seed of the fixed-scale + counterfactual model (two-epoch budget unless noted). All seeds exceed chance (0.33) on counterfactual examples; convergence depth varies (seed 2 converges shallowly on both splits). Per-seed patching (§5) was run on the 11 fixed+CF checkpoints we trained at this budget (primary seeds plus shorter/longer-budget re-runs, a superset of the behavioral seeds above), with convergence judged by the same forced-choice readout the patching harness uses (counterfactual accuracy on the 608 matched pairs, > 0.5). Six converged (0.78–0.83; changed-rate 0.72–0.83, mean 0.76)—including behavioral seed 2, whose shallow free generation (0.456) reads 0.78 under the forced choice—and the other five re-runs did not (0.15–0.17, below chance; changed-rate ≤ 0.013, like a conventionally-trained model), giving an all-checkpoint mean of 0.42. We report the converged mean to localize

the mechanism *conditional on* successful behavioral depth use: activation patching can only move a read-out that training installed, so a non-converged run is uninformative about it rather than evidence against it. Both figures are stated, so the 0.76 headline is an explicit conditioning, not selective exclusion.

Table 6: Per-seed accuracy of fixed-scale + counterfactual training. Layer-6 depth-patching changed-rate at layer 6 over the full matched-pair set spans 0.72–0.83 across six converged frozen seeds (mean 0.76; 0.81 on the full-ViT variant), so the causal read-out (§5) is seed-robust, not a single-checkpoint artifact. The final column is the parse-independent bin-logit readout on counterfactual examples: between-seed mean 0.822, SD 0.014, 95% CI $[0.809, 0.835]$—an order of magnitude tighter than free generation (SD 0.08), locating the seed variance in generation behavior rather than in the learned read-out (seed 2's shallow free-gen 0.456 still reads 0.814 under bin-logit).

seed	base	counterfactual	bin-logit (cf)
42	0.659	0.617 (0.770 at three epochs)	0.835
1	0.672	0.682	0.805
2	0.367	0.456	0.814
3	0.581	0.634	0.814
4	0.614	0.722	0.817
5	0.606	0.678	0.844
6	0.587	0.623	0.826

C Patching Protocol

Pairs are formed from the counterfactual validation set by matching a base and a counterfactual item with the same scene and queried region whose answer bins differ; with at most one pair per (scene, region) and a cap of 100 candidate pairs this yields $n = 62$ valid pairs (no filtering on model correctness). For each pair we run the source input once, capture the hidden states at the depth-image token positions of a single layer, and re-run the target input with those positions overwritten at the same layer. The answer is read as the argmax over the three bin-word logits at the final position of the forced prefix "The {region} is"; on unpatched runs this agrees with penalty-free greedy generation (quantified below). Wilson 95% intervals for the headline rates at layer 6: changed-by-patch $0.82\,[0.71, 0.90]$, flip-to-source $0.73\,[0.61, 0.82]$. Patching is symmetric in both directions (§5). Patching establishes that the answer depends on these activations under substitution at the tested layer; it does not identify a complete causal circuit, and our claims are scoped accordingly. Two RGB-token controls bracket the intervention. Patching RGB positions between the members of a pair (which share the RGB image, so source and target activations coincide) leaves answers unchanged, confirming the hook is mechanically non-destructive; patching RGB positions *across* scenes (activations from a different image's forward pass) changes 32% of the conventionally-trained model's answers, confirming the same intervention does move answers through tokens the model actually uses. Cell-restricted patching uses the local-counterfactual pairs (base and cell-swapped depth, identical otherwise; $n = 100$): the queried cell of the 3×3 grid is mapped to its depth-token indices via the processor's patch grid, and only those positions are overwritten. The $0.76/0.53$ vs. 0.05 quoted in §7 are the layer-6 changed-by-patch rates on these pairs for two local-CF checkpoints (full-ViT and seed 1) vs. the global-CF model.

Content-specificity control (answer-preserving patch). To check that the effect tracks the metric *content* of depth rather than the mere presence of an edit, we patch the depth tokens of a target item with those of a *different scene whose answer bin is the same* (largest within-bin distance gap; same token positions and length, $n = 100$). This answer-preserving patch changes only $0.37\,[0.28, 0.47]$ of answers—far below both the $0.87\,[0.79, 0.92]$ of an answer-flipping *cross-scene* patch (provenance-matched: both arms use foreign-scene donors, differing only in the donor's answer bin) and the $0.83\,[0.75, 0.89]$ of the within-scene base→CF patch in the same harness, with non-overlapping CIs—and 80% of patched answers remain on the correct bin. The read-out is thus content-specific to the metric rather than a brittle "a depth edit occurred" signature; the residual 0.37 reflects boundary cases moved by the within-bin distance change.

Readout faithfulness and a conservative behavioral protocol. The mechanistic analyses (patching, steering, probe) read the answer-bin logit at the forced prefix "The {region} is"; the behavioral tables (§4, §7) instead use free greedy generation with a repetition penalty of 1.2 and last-bin parsing. On a matched 400-item sample the logit readout agrees with penalty-free greedy generation on 0.998 of answers ($399/400$; no generation emits more than one bin word), so the readout is faithful. The repetition penalty, however, penalizes the answer words *near*/*medium*/*far*—all three appear verbatim in the instruction—and compresses the behavioral numbers toward chance. Read out without the penalty on the full counterfactual set, the penalty-free readout reproduces the entire 2×2 of Table 4 with wider margins: the fixed+CF model scores cf 0.835 on the primary seed (versus 0.62 under the generation protocol) and collapses to 0.39 when depth is removed; a conventionally-trained fixed model scores 0.445 *with* depth and an identical 0.445 *without* it (demonstrably depth-blind under the faithful readout); and counterfactual training on *normalized* depth reaches only 0.44 on average (three seeds, 0.40–0.49), at the same floor. The reported behavioral gaps are therefore conservative—the faithful readout *widens* the depth-use gap ($0.835 \to 0.39$ on ablation, versus $0.62 \to 0.36$)—and every comparison in our tables uses one identical protocol across all conditions.

D Run Counts for Negative Results

Each alternative training signal (output-level KD; auxiliary depth-decoding head; π-distill-style privileged-mode alignment) used the two-epoch budget with the hyperparameters of the main recipe. KD and the depth-decoding head are each replicated across three seeds (cf-sensitivity $|\cdot| \leq 0.004$ throughout); π-distill is a single run. We report them as existence-level negatives (none produced cf-sensitivity above 0.01), not as tuned baselines. The local-counterfactual (CF+local) recipe shows the same convergence variability as the global one: three of four frozen-encoder seeds acquire the local read-out under free generation (local-cell accuracy 0.55–0.59; depth-shuffle control 0.37), while the fourth stays at its shuffle control (0.37) yet still reads the local counterfactual under the forced-choice readout (cell-pair accuracy 0.88); a full-ViT variant reads 0.50. The global read-out is preserved on all four seeds (0.63–0.71). Cell-restricted patching moves $0.76/0.53$ of answers (full-ViT and seed-1 local-CF checkpoints) vs. 0.05 for the global-CF model; counterfactual mixing is two seeds (general F1 $0.55/0.56$, cf $0.63/0.74$). The SmolVLM2 replication spans three two-epoch seeds (cf 0.799 ± 0.013; layer-0 patching changed-rate 0.833 ± 0.009, per-seed $0.828/0.828/0.844$; normal$-$shuffle gap 0.395 ± 0.050); its patching sweep reproduces the same early-layer window on both the one- and two-epoch checkpoints, and its conventional-training control mirrors the Qwen inert/active contrast (§7). The 9B replication likewise spans three runs (bin-logit cf $0.810/0.829/0.843$ vs. shuffled $0.653/0.736/0.698$; layer-6 changed-rate 0.859 on all three, conventional 0.016). The steering experiment is one run per condition (calibration on 25 pairs, evaluation on 120 disjoint items), with a norm-matched random-direction control (Appendix L).

E Cross-Dataset Real-Sensor Transfer

We test whether the absolute-scale read-out transfers to *real sensor depth from datasets absent from training* (training uses NYU Depth v2 only). The key comparison is CF vs. depth-removed/RGB-only, because the recruited read-out is a global metric-scale anchor (shuffle-invariant), so depth removal—not shuffling—is the clean ablation.

2D-3DS (Matterport sensor, different buildings; prior-free grid-cell camera-distance, $n = 420$). The CF model reads depth: 0.557 $[0.510, 0.605]$ with depth vs. 0.336 with depth removed (gray) and 0.362 for RGB-only—both at chance ($\frac{1}{3}$); the conventionally-trained model is also at chance (0.362). Removing the depth collapses the CF model to chance, so its above-chance accuracy on real cross-dataset sensor depth is *entirely* depth-driven. (Shuffling does not degrade it, 0.610, as expected for a global anchor whose histogram shuffling preserves.) Because the queried referent is a grid cell, no object-identity prior is available—this isolates depth use.

SUN RGB-D (Kinect/RealSense sensor; object-named camera-distance, $n = 450$). Here the object-identity prior alone is strong: RGB-only already reaches 0.529 and the CF model 0.538, so depth adds little *in absolute terms*—yet shuffling the depth still costs the CF model 0.10 ($0.538 \to 0.438$), showing depth is read but masked by the prior. This is precisely the masking that motivates CF-Depth, now observed on a real benchmark. The same prior-domination holds on the independent

SpatialBench (Cai et al., 2025a) (reach/size, $n = 80$): removing or shuffling the depth changes $\leq 2/78$ answers, so the model is prior-driven there too.

Takeaway. The read-out transfers to real cross-dataset sensor depth when the object-identity prior is controlled (2D-3DS); on prior-rich named benchmarks the prior dominates the headline number while depth use persists under the hood (SUN RGB-D shuffle drop). The *linear-probe* transfer to SpatialRGPT-Bench (estimated depth) is separately weak, as expected for a dataset-specific decoder; its *behavioral* fixed+CF result is nonetheless depth-driven (App. J).

Is the RGB→depth redundancy NYU-specific? The relative-depth decodability from RGB tokens (§5) is not an NYU layout artifact: a matched per-cell relative-depth probe from RGB tokens reaches $R^2 \approx 0.79$ *within* 2D-3DS (different buildings and sensor; scene-grouped 5-fold CV, base model, $n = 3726$ cells), comparable to NYU (0.86). The linear probe does not *transfer* across datasets, however—trained on NYU it predicts 2D-3DS relative depth at $R^2 \approx 0.10$ and absolute depth at $R^2 < 0$—so the redundancy holds *per dataset* through dataset-specific linear maps, consistent with absolute metric scale being a dataset-specific rather than a generalizing monocular cue.

F Real-Sensor Object-Level Depth Use

We test object-*local* use of *unedited* sensor depth (no counterfactual editing). From NYU Depth v2 instance masks we build a per-object absolute-distance benchmark: each object region's real median sensor depth is binned into near/medium/far tertiles ($e_1 = 1.9$ m, $e_2 = 3.0$ m), bin-balanced to $n = 2067$ (689 each; chance $\frac{1}{3}$), with the question "How far is the {object} from the camera?". Models read the fixed-scale *real* depth; conditions are full depth, pixel-*shuffled* depth (global histogram preserved, local structure destroyed), and depth-*removed* (gray); the RGB-only model uses no depth. Parse rate is 100%. Because the 2067 objects span only 264 NYU scenes, we report *scene-clustered* bootstrap 95% CIs (resampling scenes), $\sim 1.5\times$ wider than i.i.d. Wilson.

Table 7: Real-sensor per-object absolute distance (NYU Depth v2, $n = 2067$, chance $\frac{1}{3}$). Only the local-CF model shows the object-local signature (full $\gg$ shuffle $\gg$ removed); the conventional model is depth-blind.

model (training)	full depth	shuffled	removed
conventional (ordinary)	0.637	—	0.641
global CF	0.542	0.480	0.451
local CF (global+local)	**0.596**	0.459	0.341
RGB-only model	0.629	—	—

The conventional model is depth-blind (0.637 with depth $=$ 0.641 removed; CIs overlap), riding an RGB shortcut. Both CF models are genuinely depth-dependent, but only the *local*-CF model shows object-local read-out: a significant shuffle drop ($0.596 \to 0.459$; scene-clustered paired difference $+0.137$, 95% CI $[+0.106, +0.168]$) to a chance-level removed floor (0.341), versus the global model's smaller drop ($0.542 \to 0.480$) and above-chance floor (0.451)—the global/local distinction of §7 on *non-synthetic* data. The benchmark is partly RGB-solvable in-domain (RGB-only 0.629), so depth use is evidenced by the shuffle/removal contrasts, not raw accuracy; consistent with App. E, the global anchor is shuffle-invariant while local CFs add a shuffle-sensitive, spatially-resolved read-out that is in-domain and does not transfer across sensors. On the held-out 2D-3DS sensor (a region-metric construction distinct from the prior-free grid-cell transfer of App. E; $n = 420$), the local-CF model's shuffle-sensitivity vanishes (normal 0.419 vs. shuffled 0.433, overlapping CIs; depth-removed at chance, 0.333): only the coarser depth-*dependence* transfers across sensors, not the object-local spatial decoding.

G A Self-Trained InternVL3-2B

For cross-architecture breadth from our own models, we *train* a depth-augmented InternVL3-2B (LoRA, fixed-scale depth, the data of §7) and audit it on CF-Depth. Ordinary training leaves it depth-blind (CF-combined 0.486, near chance; cf-sensitivity 0.019), whereas counterfactual training

recovers depth use (CF-combined 0.797; cf-sensitivity 0.096, 95% CI $[0.084, 0.108]$, excluding 0). The effect is smaller than on Qwen3.5/SmolVLM2, but the pattern—ordinary training ignores depth, counterfactual training recruits it—holds on this fourth self-trained architecture (Qwen3.5-2B, Qwen3.5-9B, SmolVLM2, InternVL3-2B).

H Two-Cell Swap: Behavioral Evidence Beyond a Global Read-out

To test whether the local read-out exceeds *any* global read-out behaviorally (not only under cell-restricted patching), we construct a task on which any global-statistics read-out is provably capped (right on at most one member of each pair). From NYU 3×3 depth grids we form 678 base/swap pairs (239 scenes; chance $\frac{1}{3}$): each pair shares one RGB image and a "how far is the $\{$cell$\}$?" question, and the swap exchanges the depth of two equal-size cells in different distance bins, flipping the queried cell's answer. Because only two equal cells are exchanged, the 256-bin *global* depth histogram is identical on base and swap (verified $L1 = 0$ on all 678 pairs), so any global-statistics read-out emits the same answer on both and is right on at most one member of each pair (its base and swap accuracies sum to at most 1).

Table 8: Two-cell swap ($n = 1356$ rows $= 678$ base $+ 678$ swap; chance $\frac{1}{3}$; global depth histogram fixed, $L1 = 0$). Top block: Qwen3.5-2B models and baselines; only the local-CF model tracks the swapped cell. Bottom block: the same recipe on SmolVLM2-2.2B replicates the local read-out (seed 42 shown; all three seeds in the text below).

model / baseline	overall	base	swap
local-CF (global+local)	**0.531**	0.484	**0.578**
global-CF	0.374	0.435	0.313
conventional	0.386	0.478	0.295
RGB-only	0.378	0.496	0.260
global-stats logistic	0.406	—	—
SmolVLM2 local-CF (global+local)	**0.777**	0.684	**0.870**
SmolVLM2 local-CF, depth withheld	0.389	0.445	0.333

Only the local-CF model tracks the swap: $0.531\,[0.504, 0.557]$ overall and 0.578 on swapped rows, a base+swap sum of 1.06 (SmolVLM2: 1.55), above the ceiling of 1 for any read-out that answers a pair identically, whereas the global models collapse to 0.26–0.31 on swapped rows (they emit one answer per pair). This is direct behavioral evidence—independent of cell-restricted patching—that the recruited read-out is spatially resolved rather than a global scalar. A scene-clustered bootstrap (resampling the 239 scenes, $B = 4000$) confirms the separation: local-CF $0.531\,[0.496, 0.566]$, swapped rows $0.578\,[0.540, 0.615]$, and the *paired* local-CF$-$global-CF difference is $+0.157\,[0.132, 0.182]$, excluding 0 (likewise vs. conventional $+0.145$ and RGB-only $+0.153$); the spatial advantage over a global read-out thus survives scene-clustering.

The spatially resolved read-out also transfers across architectures, and on SmolVLM2 it replicates on all three training seeds. SmolVLM2-2.2B trained with the same global+local counterfactual recipe reaches $0.777/0.835/0.880$ overall (seeds $42/1/2$; $n = 1356$, parse rate 1.0; the weakest seed's CI is $[0.755, 0.799]$) and $0.870/0.935/0.944$ on swapped rows, base+swap sums 1.55–1.76; withholding depth at evaluation collapses every checkpoint to the same floor, $0.389\,[0.364, 0.415]$ overall and 0.333 on swapped rows (the three seeds' depth-withheld predictions are identical item-by-item), so the advantage is carried by the depth channel rather than by RGB or answer priors. On the local-cell counterfactual evaluation ($n = 435$, chance $\frac{1}{3}$) the three seeds read $0.457/0.503/0.480$. The local read-out is thus not a Qwen-specific artifact, though its strength varies with architecture and training run (SmolVLM2 exceeds Qwen3.5-2B on two-cell while trailing it on local-cell accuracy).

The two-cell signature shows the same seed variability as the other local read-out measures (App. B): of the eight local-CF checkpoints we evaluate (the Qwen primary, three Qwen seed re-runs, the Qwen full-ViT variant, and three SmolVLM2 seeds), six exceed the pair-constant ceiling clearly (base+swap sums 1.06, 1.07, 1.21, 1.55, 1.67, 1.76; overall 0.531–0.880; swapped rows 0.578–0.944), one marginally (1.01; swapped rows 0.608 against a weak base 0.397), and one Qwen seed

does not track the swap (0.81; overall 0.407). As with patching convergence, the claim is that local counterfactual training *can* install a spatially resolved read-out that no global account reproduces, not that every run converges to one.

I HELD-OUT RESCALING MAGNITUDES

To test whether the recruited read-out is a *general* absolute-scale function rather than a memorization of the training counterfactuals, we evaluate the fixed+CF model (§7) on counterfactuals built from rescaling magnitudes *disjoint* from training. Training counterfactuals use scale factors $\{0.5, 0.7, 1.4, 1.9\}$; here every counterfactual uses an unseen factor from $\{0.4, 0.6, 1.2, 1.6, 2.3\}$, with 0.4 and 2.3 *extrapolating* beyond the training range. On these held-out scales ($n = 4960$ counterfactual items, chance $\frac{1}{3}$, bin-logit readout) the model still reads depth at cf $0.80\,[0.79, 0.81]$—far above a conventionally-trained model evaluated on the same items ($0.44\,[0.43, 0.46]$) and above chance—and drops to $0.63\,[0.62, 0.64]$ under spatial shuffling (the residual reflects the read-out being a *global* scale anchor, §7, which shuffling leaves partly intact). The read-out therefore generalizes across rescaling magnitudes it never saw—interpolating and extrapolating a learned scale-to-answer mapping rather than memorizing the training counterfactuals; its cross-*dataset* generality is more limited (App. E).

J A NON-SYNTHETIC BENCHMARK: SPATIALRGPT-BENCH

Our absolute-distance task is constructed; to test the read-out on *human-authored* questions we evaluate on SpatialRGPT-Bench's distance split (Cheng et al., 2024) (three-way, chance $\frac{1}{3}$, real sensor depth), with object references *anonymized* so that an object-identity prior cannot substitute for depth ($n = 148$). The fixed+CF model's accuracy is depth-driven and content-specific: $0.534\,[0.453, 0.615]$ with real depth, falling to $0.405\,[0.324, 0.486]$ under spatial shuffling and to $0.331\,[0.257, 0.405]$ with depth removed—the with-vs-without-depth gap is significant (non-overlapping intervals) and the depth-removed score sits at chance. A conventionally-trained model is invariant to depth (0.385 normal vs. 0.399 shuffled) and an RGB-only model scores 0.365, both at the depth-removed floor. The recruited read-out thus carries from our synthetic task to a published benchmark on real questions; the modest n leaves the fixed+CF and conventional intervals overlapping at the boundary, so we read this as corroborating though not decisive. (This is the *behavioural* use result; the linear-probe transfer of §E is separately weak, as expected for a dataset-specific decoder.)

K OFF-THE-SHELF DEPTH-VLMS ON CF-DEPTH

SD-VLM. SD-VLM-7B (Chen et al., 2025)—the depth-VLM whose per-image depth normalization our mechanism most directly predicts to fail—is invariant to depth content on CF-Depth: real vs. spatially-shuffled depth give identical answers on 92% of items (cf-sensitivity 0.012; accuracy at chance, $n = 2000$). In its default mode it does not ingest an external depth image (it estimates depth from RGB); forced to read our depth, it applies per-image min-max normalization that discards absolute scale, consistent with the prediction. The released checkpoint has a depth-module weight mismatch with the released code, bypassed by the deterministic positional-encoding path we use—a reproducibility caveat. To test whether the read-out failure extends beyond models we trained, we run a *published* RGB+depth VLM—SpatialBot-3B (Cai et al., 2025a) (Phi-2+SigLIP, purpose-built for metric-depth spatial QA)—on the CF-Depth metric task without any fine-tuning. Depth is supplied in SpatialBot's native format (its lossless three-channel encoding of uint16-millimetre depth), reconstructed from our fixed-scale maps; answers are read by free generation, which SpatialBot emits as a clean near/medium/far word (parse rate 1.00). On the full validation set ($n = 5{,}592$, chance $\frac{1}{3}$), accuracy is *invariant* to the depth signal: 0.405 with real depth, 0.385 with spatially shuffled depth, and 0.383 with depth removed (constant gray)—a real-vs-removed gap of 0.02, all three barely above chance. On the $4{,}406$ counterfactual items, where depth is edited to flip the answer, it scores 0.408 with real depth versus 0.397 with depth removed: it does not follow the edit. **Fairness (positive control).** This null is not an artefact of an unreadable input format. Queried in SpatialBot's signature mode ("what is the depth value of point $\langle x, y \rangle$?"), its returned values rank-correlate with the true metric depth of our maps (Spearman $\rho = 0.59$, $n = 540$), confirming it *reads* the depth it does

not *use*. A model built and named for depth thus shows the same perceive-but-don't-use signature as our ordinary-trained models: the read-out failure spans both the models we train and off-the-shelf depth-VLMs rather than our particular training alone.

A second off-the-shelf depth-VLM: RoboRefer-2B-SFT. We repeat the audit on RoboRefer-2B-SFT (Zhou et al., 2025), the released supervised checkpoint of a depth-specialized referring model (SigLIP RGB and depth towers). On the binary region-ordering CF-Depth instance ($n = 1{,}136$, chance 0.5; identical RGB with the depth image swapped so the answer flips), its answer is *invariant to the depth image*: 0.470 with real depth, 0.505 with the counterfactually-swapped depth, and 0.511 with depth removed—all near chance, and removing depth does not lower accuracy. Across the $1{,}136$ pairs it returns the same parsed region on 97.4% and a byte-identical answer on 77.7% when the depth is swapped; the answer-flip rate is 0.001 $(1/1{,}136)$. Both arms use an identical single-channel→RGB depth conversion, so the only between-arm difference is depth content. RoboRefer's *reported* depth use arises from a subsequent reinforcement (forcing) stage (Zhou et al., 2025); that RFT-stage checkpoint is not publicly released, so the audit is necessarily scoped to the supervised stage. The supervised-checkpoint null is thus consistent with—indeed an instance of—our finding that ordinary supervision does not recruit image-depth read-out, while an explicit forcing signal does (§7); under our account, the unreleased RFT stage is exactly where read-out would be recruited. The null also holds at larger scale: the RoboRefer-*8B* SFT checkpoint is likewise invariant to the depth image (answer-flip rate 0.00, $n = 1{,}136$; 0.49 with real vs. 0.48 with depth removed, both at chance), so the read-out failure in published depth-VLMs is not a small-model artefact.

L Inference-Time Steering of the Depth Read-out

Following activation-steering / activation-addition methods (Turner et al., 2023; Li et al., 2023), we estimate a "farther" direction d as the mean, over counterfactual pairs, of the layer-6 depth-token activation difference (counterfactual minus base, sign-aligned so that increasing depth points toward "far"), calibrated on 25 pairs. On 120 disjoint held-out items (answer distribution near/medium/far $= 35/48/37$) we add $\pm\alpha d$ to the depth-token positions at layer 6 and read the answer bin (Table 9). In the counterfactually-trained model the answer moves dose-dependently and *every* induced change is in the steered direction (monotone in α; $+d$ and $-d$ move at matched rates because the held-out set is bin-balanced). Three controls show the effect is specific, not a generic perturbation: a *random* direction of the same norm changes $\leq 0.8\%$ of answers at every dose; the conventionally-trained model is unmoved at any dose; and the counterfactually-trained model is unmoved when d is added at layer 12 (outside the causal window of §5). The read-out is linearly manipulable exactly where—and only where—patching localizes it.

Table 9: Fraction of held-out answers changed by adding $\pm\alpha d$ to the layer-6 depth tokens ($n = 120$; Wilson 95% CI for the top row's $+d$: $\alpha = 1$ [.01, .08], 2 [.05, .15], 4 [.10, .23]). The depth direction moves answers dose-dependently; a norm-matched random direction, the conventional model, and layer 12 (outside the window) do not.

condition	$\alpha = 1$	2	4
CF-trained, L6, $+d$ (toward far)	.03	.08	.16
CF-trained, L6, $-d$ (toward near)	.03	.08	.17
CF-trained, L6, *random* dir.	.01	.01	.01
conventional, L6, d	.00	.00	.00
CF-trained, L12, d (outside window)	.01	.01	.01