RETHINKING SPATIAL REASONING IN MLLMS: LEARNING TO FORM AND USE DEPTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning is essential for multimodal models to understand scenes and support embodied interaction, yet representing spatial information does not ensure that it guides their answers. We investigate how models can learn to form useful depth representations and turn them into evidence for spatial judgments. Our finding that depth information becomes readable before it has its strongest influence on answers motivates a training strategy that couples early depth formation with its subsequent use. Using existing question-answering annotations, we supervise object depth in early layers and train later layers to answer consistently with controlled changes in represented depth relations, requiring no additional geometry data, encoder, or teacher. Across VSI-Bench, CV-Bench, BLINK spatial tasks, MMSI-Bench, and ViewSpatial-Bench, our method raises mean scores by 3.8, 1.7, and 2.5 percentage points over supervised fine-tuning on identical data for Qwen2.5-VL-3B, Qwen3-VL-4B, and Qwen3.5-4B, respectively. These findings support early depth formation and downstream use as complementary training targets for improving spatial reasoning while preserving the original inference architecture.

1 INTRODUCTION

Spatial reasoning enables multimodal large language models (MLLMs) to judge object relations, distances, and viewpoints, supporting scene understanding and embodied interaction. It requires recovering spatial structure from visual observations and using it in relational judgments. This remains challenging even for models with strong visual recognition and language capabilities (Yang et al., 2024).

Recent work addresses this challenge through richer supervision and geometric representations. SPAR provides diverse spatial question-answering (QA) tasks grounded in 3D scenes (Zhang et al., 2025a). VG-LLM strengthens visual inputs with pretrained geometric features (Zheng et al., 2025). GeoVR further internalizes geometry through teacher-guided training without an additional geometry encoder at inference (Wang & Huang, 2026). Together, these advances establish the value of spatial supervision and geometric knowledge.

Yet representing spatial information does not ensure that the model uses it to answer. For example, its internal states may distinguish which object is closer while its final judgment remains insensitive to that relation. Prior analysis links such failures to suppressed positional sensitivity and imbalanced multimodal attention (Qi et al., 2025). What remains unclear is *where depth becomes available and where it influences spatial judgments*. Without distinguishing these processes, improved geometric representations or aggregate QA scores alone offer limited guidance on what training should change.

We therefore ask: *how can training connect the formation of spatial representations to their use in answering?* We focus on depth, which requires reasoning beyond image-plane positions. Our analysis finds that peak depth readability precedes peak answer response; VG-LLM exhibits earlier readable depth and stronger early answer effects than its same-data fine-tuned backbone, Qwen-SFT (Section 3). Fig. 1 illustrates the earlier depth availability. These findings motivate *early depth formation followed by downstream processing and use* as a training principle.

Our approach jointly develops depth availability and downstream responsiveness (Fig. 5). Alongside natural QA, we ground depth in early states using existing QA annotations and train subsequent lay-

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.

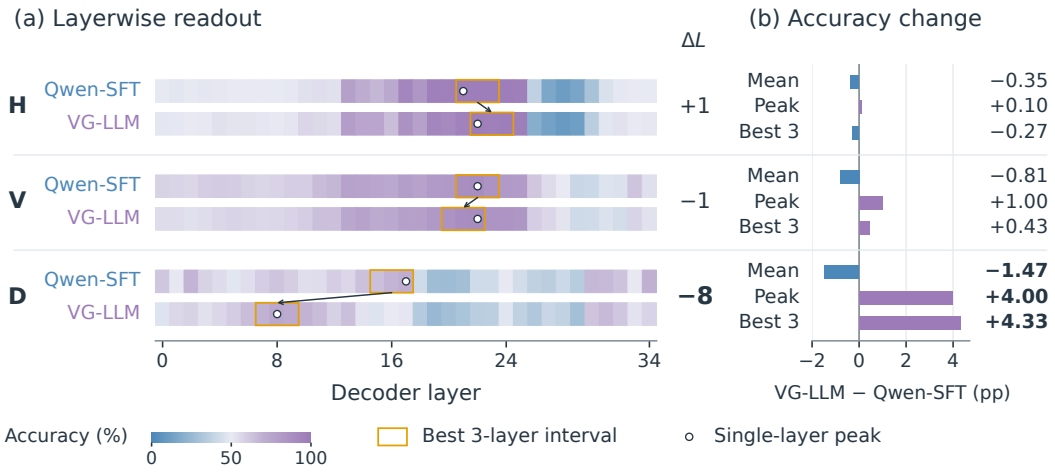

Figure 1: Spatial relation readout. (a) Horizontal (H), vertical (V), and depth (D) accuracies share a 0–100% scale. Boxes mark best three-layer intervals and circles mark peaks; ΔL gives the interval-center shift (VG-LLM minus Qwen-SFT). (b) Accuracy differences for the layer mean, peak, and best three-layer mean (Best 3), in percentage points.

ers to answer consistently with controlled changes in represented depth relations. Separate learning signals guide formation and use without additional geometry data, encoders, or teachers; inference retains the original architecture.

Our contributions are threefold:

- **We reveal a separation between depth readability and answer influence.** Their layer-wise ordering motivates early formation followed by downstream use.
- **We propose learning to form and use depth together.** Independent post-training diagnostics show improved early readability and stronger early depth-directed answer effects.
- **Our method improves spatial reasoning across backbones.** All five benchmark scores improve over same-data SFT, with mean gains of 3.8, 1.7, and 2.5 points on Qwen2.5-VL-3B, Qwen3-VL-4B, and Qwen3.5-4B, respectively; the method also exceeds VG-LLM's mean on each (Table 2).

2 Related Work

Geometry-enhanced spatial reasoning. VG-LLM (Zheng et al., 2025) provides a compelling demonstration of the value of geometry priors for spatial reasoning, fusing pretrained VGGT features (Wang et al., 2025) with visual tokens. Related architectures combine geometric and semantic encoders (Wu et al., 2025) or align and fuse geometry with language (Fan et al., 2025; Zhang et al., 2026; Zhao et al., 2025). We use VG-LLM to trace depth formation and use within spatial reasoning, then derive training objectives for the original backbone.

Learning spatial representations. SPAR constructs spatial QA from 3D scenes (Zhang et al., 2025a); subsequent work develops spatial training data and curricula (Li et al., 2025b; Yang et al., 2025b;a). GeoVR uses a geometry teacher to supervise depth, pose, scale, and multi-layer features (Wang & Huang, 2026); SpatioLM supervises a spatial module with depth and camera targets (Wu et al., 2026). Our analysis motivates coupling early object-depth supervision with training the decoder's subsequent use of these coordinates.

Mechanistic analysis of spatial reasoning. Mechanistic studies connect spatial representations to their effects on predictions. Qi et al. (2025) analyze visual-token scale and positional sensitivity through probes and targeted interventions. We examine object depth across decoder layers:

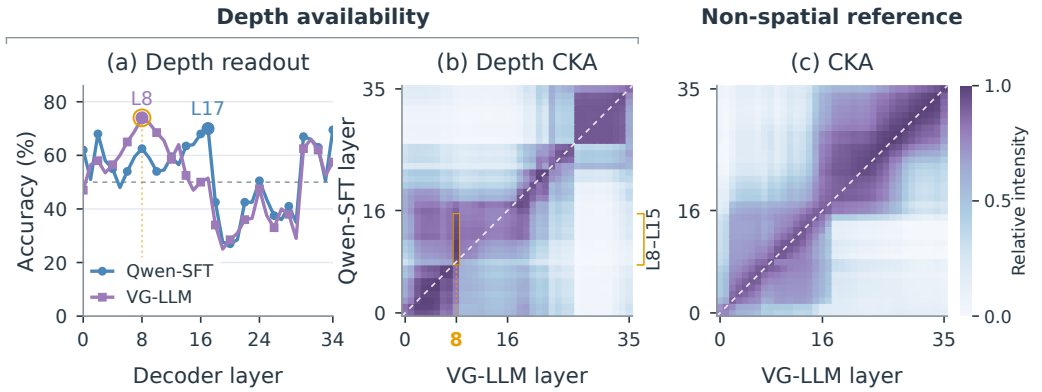

Figure 2: Depth readout and cross-layer similarity. Orange markers link VG-LLM L8 across (a,b); the box spans Qwen-SFT L8–L15. Panel (c) provides a non-spatial reference. CKA uses per-panel min–max normalization and a shared relative-intensity scale. The gray horizontal line marks 50% accuracy; white diagonals mark equal layers.

relation readout and cross-layer CKA characterize depth availability and representation correspondence, while directional edits and attribution-guided channel ablations test effects on answer scores and generation. These measurements jointly motivate early formation and subsequent use.

Semantic supervision and intervention training. Concept bottleneck models supervise semantic variables (Koh et al., 2020). Interchange intervention training (IIT) supervises counterfactual outputs after activation swaps, can include variable supervision, and differentiates through source and base computations (Geiger et al., 2022). We ground depth at multiple early layers and supervise answers after paired directional edits. Detachment assigns edited-answer supervision to downstream computation, while depth supervision learns the coordinates.

3 Depth Formation and Use in Spatial Reasoning

Using frozen VG-LLM and Qwen-SFT models trained on the same spatial QA data, we trace how readable depth influences answers. Readout locates accessible relations; CKA compares representation structure across models and layers. We then test downstream answer responses through relation edits and generated judgments through spatial-channel ablations. Layers are zero-indexed.

3.1 Depth becomes readable earlier

Locating readable depth. We construct H/V/D object pairs from SPAR's ScanNet and ScanNet++ scenes (Zhang et al., 2025a; Dai et al., 2017; Yeshwanth et al., 2023), with scene-disjoint estimation and evaluation splits. H/V labels use bounding-box centers; D uses object-center depth in the same camera frame. At block ℓ's output, we extract each object's last name-subtoken residual before final RMSNorm. Write the pair mean as $c_{i,\ell} = (h_{i,\ell,a}+h_{i,\ell,b})/2$ and relation state as $r_{i,\ell} = h_{i,\ell,a}-h_{i,\ell,b}$.

Let $F_{i,\ell,g}(r)$ be the final positive-minus-negative answer-logit margin for relation $g \in \{H, V, D\}$, when the two states are $c_{i,\ell} \pm r/2$ and other token states stay fixed. Averaging its gradient over N_g direction-estimation examples gives

$$\bar{v}_{\ell,g} = \frac{1}{N_g}\sum_i \nabla_r F_{i,\ell,g}(r_{i,\ell}), \qquad u_{\ell,g} = \frac{\bar{v}_{\ell,g}}{\|\bar{v}_{\ell,g}\|_2}, \qquad \hat{y}_{i,\ell,g} = \mathrm{sign}(u_{\ell,g}^\top r_{i,\ell}). \tag{1}$$

This readout measures how reliably each fixed answer-gradient axis recovers ground-truth relations on independent confirmation examples; positive depth means that entity a is closer than b. Appendix A gives split sizes, the estimator derivation, and full H/V/D results.

The H/V/D profiles in Fig. 1a show high H/V readout accuracy in both models, with best three-layer intervals differing by only one layer. VG-LLM's depth readout peaks at 74.00% at L8, with its best

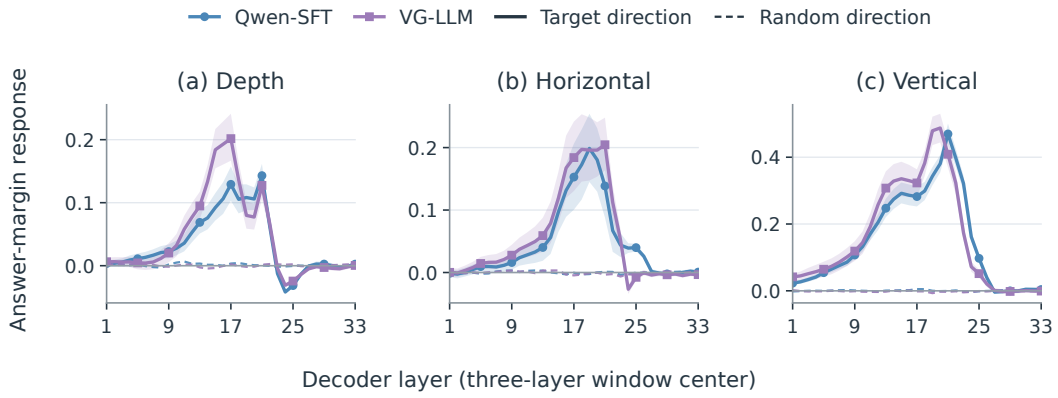

Figure 3: Directional answer responses at $\rho = 10\%$. Solid curves show target directions, dashed curves equal-norm random controls, and shading 95% bootstrap confidence intervals. All three directions use three-layer window centers L1–L33.

three-layer mean at L7–L9; Qwen-SFT peaks at 70.00% at L17, with its best window at L15–L17. The best depth interval thus moves eight layers earlier in VG-LLM.

VG-LLM's depth peak and best-three-layer mean exceed Qwen-SFT's by 4.00 and 4.33 percentage points, respectively; H/V differences in these two statistics are at most 1.00 point in magnitude (Fig. 1b). Layer-average depth accuracies are 51.01% for VG-LLM and 52.49% for Qwen-SFT (Table 4). The higher peak and best-window mean accompany an earlier concentration of readable depth.

Cross-layer correspondence. Debiased linear CKA (Kornblith et al., 2019; Song et al., 2012) compares the similarity structure of object-token representations across samples, complementing relation readout. We use 600 CV-Bench Depth questions (Tong et al., 2024) and 600 non-spatial reasoning samples from LLaVA-Hound (Zhang et al., 2024a;b). With corresponding samples stacked in matching order, $M_{\ell k} = \mathrm{CKA}_{\mathrm{deb}}(X_\ell^{\mathrm{SFT}}, X_k^{\mathrm{VG}})$ compares Qwen-SFT layer ℓ (rows) with VG-LLM layer k (columns) over L0–L35. Higher values indicate greater similarity; Appendix A gives the estimator.

On depth questions, $M_{\ell,8}$ is high for $\ell = 8, \ldots, 15$ (Fig. 2b), with off-diagonal values exceeding the corresponding same-layer similarities $M_{\ell\ell}$. VG-LLM's early states therefore organize depth-question samples similarly to later Qwen-SFT states. The non-spatial reference has more diagonal correspondence. Together, readout and CKA support earlier depth availability. We next test whether editing these depth relations changes answer scores and locate the strongest response.

3.2 Peak answer influence follows peak readability

We edit relation states along the readout axes while holding the image, question, model parameters, and pair mean fixed. With $B_\ell = [u_{\ell,H}, u_{\ell,V}, u_{\ell,D}]^\top$, choose

$$\begin{aligned} w_{\ell,D} &= \arg\min_w \|w\|_2^2 \quad \text{s.t.} \quad B_\ell w = (0,0,1)^\top, \\ h_a^{(s)} &= h_a + \frac{s\eta}{2} w_{\ell,D}, \qquad h_b^{(s)} = h_b - \frac{s\eta}{2} w_{\ell,D}. \end{aligned} \tag{2}$$

Here $s \in \{-1, +1\}$; sample and layer indices on h are omitted. The edit changes the depth reading by $s\eta$ while preserving edited-layer H/V readings. Subsequent layers run normally. Strength ρ is the median per-object relative state change, calibrated per model and layer. Let $M_{i,\ell,\rho}^{s,\pi}$ be the edited logit margin for the fixed relation "a closer than b," with entity order $\pi \in \{(a,b),(b,a)\}$. The Directional Utilization Score averages downstream effects over both signs, orders, and N examples:

$$U_{i,\ell,\rho} = \frac{1}{4} \sum_\pi \left(M_{i,\ell,\rho}^{+,\pi} - M_{i,\ell,\rho}^{-,\pi} \right), \qquad \bar{U}_{\ell,\rho} = \frac{1}{N} \sum_i U_{i,\ell,\rho}. \tag{3}$$

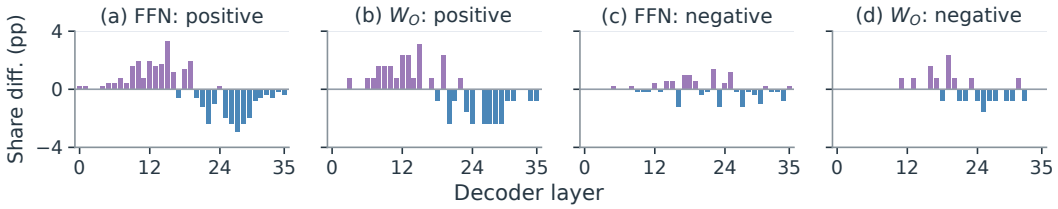

Figure 4: Layer-share differences (VG-LLM minus Qwen-SFT, pp) for positive (a,b) and negative (c,d) sets.

Positive values follow the intended relation change. Equal-norm random controls are orthogonal to the H/V/D span; Appendix A gives calibration and all six strengths.

Depth responses peak at L17 in VG-LLM and L21 in Qwen-SFT at every strength. The H/V high-response regions in Fig. 3 overlap more closely, with V peaks near L20–L21. Both depth curves turn negative near L24 and return toward zero; random responses remain near zero. The controlled edits establish a causal effect of depth-direction states on answer scores, concentrated earlier and more strongly in VG-LLM.

Peak readability precedes the strongest answer response in both models: L8 versus L17 in VG-LLM, and L17 versus L21 in Qwen-SFT. VG-LLM thus exhibits both earlier depth formation and earlier answer influence, with the larger advance in readability making their layerwise separation more pronounced. These profiles and the early-to-late CKA correspondence support early depth formation followed by subsequent relational processing and use. H/V's more closely aligned profiles identify depth as the prominent difference between models.

3.3 Spatial-channel ablations change generated answers

We complement depth-direction effects on answer scores with spatial-channel ablations, testing generated answers and examining channel distributions. On SPAR questions, we rank FFN channels entering `down_proj` and attention channels entering `o_proj` (W_O). Let $m(x)$ be the correct-answer logit minus the log-mean-exp of incorrect candidate logits. For channel j, with activations $z_{t,j}$ at visual and text prefill positions $\mathcal{P}_x$, define

$$s_j(x) = \sum_{t \in \mathcal{P}_x} z_{t,j} \frac{\partial m(x)}{\partial z_{t,j}}. \tag{4}$$

We average contributions over semantic question variants, penalize absolute contributions on non-spatial controls, and account for scene-grouped standard error. Positive candidates support the correct-answer margin; negative ones oppose it. Ranking across layers selects fixed sets of 512 FFN or 128 W_O channels per model and sign. Appendix A.4 specifies the ranking and controls.

We zero these fixed sets during visual and text prefill, then decode normally on CV-Bench. Positive-set removal lowers accuracy, negative-set removal raises it, and equal-size random sets have smaller effects (Table 1). These ablations establish the sets' behavioral relevance.

The layer-share differences in Fig. 4 show an earlier concentration of positive channels in VG-LLM, complementing the depth-specific readout and intervention evidence. Negative sets are relatively stable; the complete distributions retain many positive channels in later layers (Fig. 11).

Table 1: CV-Bench channel ablations (accuracy, %). Random sets: mean ± standard deviation.

Channels	Model	Unedited	Positive	Negative	Random
FFN (512)	Qwen-SFT	77.748	65.011	83.397	77.74 ± 0.108
	VG-LLM	80.250	71.531	84.799	80.50 ± 0.085
W_O (128)	Qwen-SFT	77.748	69.712	81.691	77.80 ± 0.284
	VG-LLM	80.250	69.560	83.586	80.88 ± 0.089

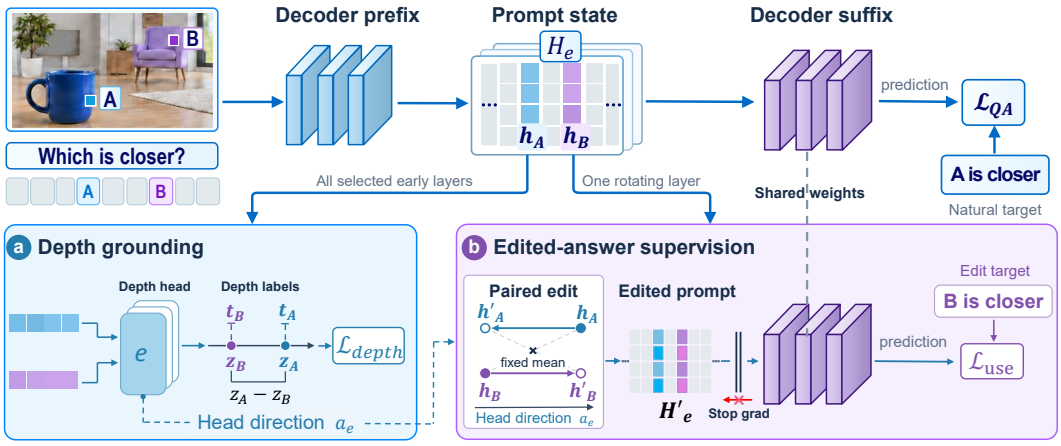

Figure 5: Depth and edited-answer supervision. (a) The natural pass grounds depth at all selected early layers. (b) A paired edit at a rotating early layer changes the relation; detachment directs its answer loss to the shared decoder suffix and output head. Here, the target changes from A to B.

3.4 From early depth formation to downstream use

Readout and CKA support earlier depth formation in VG-LLM; interventions show earlier answer influence whose peak still follows peak readability. Channel ablations link the selected spatial computations to generated answers. These findings motivate *forming useful depth representations early and explicitly training their downstream use*. Depth supervision grounds early coordinates in object geometry; edited-answer supervision trains the decoder to answer according to changes in those coordinates. Section 5.2 tests both training goals through independent readout and directional interventions.

4 Spatial reasoning through early depth formation and use

We implement the two auxiliary objectives alongside natural QA supervision. The natural pass grounds depth at all selected early layers; a continuation from one feasible paired edit learns to answer consistently with the represented relation (Fig. 5).

4.1 Grounding depth at multiple early layers

Select early blocks $\mathcal{E} = \{\ell_1, \ldots, \ell_M\}$. For natural-QA batch $\mathcal{B}$, let $\mathcal{B}_{\text{aux}}$ contain examples with camera-axis depths recovered from existing QA annotations. Targets t_o are standardized negative log depths, so larger values mean closer objects (Appendix B.1). A/B follow first mention in a same-view question q, with label difference $r^\star = t_A - t_B$ and natural answer $y = G(r^\star, q)$ determined by the wording and options.

Each layer has an independent affine head shared across entities, applied to the object-token states of Section 3:

$$z_o^{(\ell)} = a_\ell^\top h_o^{(\ell)} + b_\ell, \qquad \hat{d}^{(\ell)} = z_A^{(\ell)} - z_B^{(\ell)}, \tag{5}$$

$$\mathcal{L}_D^{(\ell)} = \mathbb{E}_{\text{obj}}\left[\text{Huber}_1(z_o^{(\ell)} - t_o)\right] + \beta \mathbb{E}_{\text{pair}}\left[\text{Huber}_1(\hat{d}^{(\ell)} - r^\star)\right]. \tag{6}$$

The first loss term grounds individual depths; the second trains their comparison. Expectations average labeled entities and same-view pairs in $\mathcal{B}_{\text{aux}}$. All heads share targets and the natural pass; $\mathcal{L}_{\text{depth}} = M^{-1} \sum_{\ell \in \mathcal{E}} \mathcal{L}_D^{(\ell)}$ averages their losses.

4.2 TRAINING DOWNSTREAM RESPONSES TO DEPTH EDITS

To train downstream use, we supervise answers consistent with edited depth relations. Once editing is enabled, the selected layer $e = e_k$ cycles through $\mathcal{E}$ across steps. We edit single-image, same-view questions whose readouts $\hat{d} = z_A^{(e)} - z_B^{(e)}$ pass relation-sign, margin, and error checks.

A per-entity relative-change cap $\rho_{\max} = 0.10$ bounds feasible targets. We choose preservation or reversal with equal probability, then sample a nonidentity difference d' from the corresponding feasible empirical label differences. Each example with a sampled target receives one paired edit:

$$h_A^{(e)\prime} = h_A^{(e)} + \frac{d' - \hat{d}}{2\|a_e\|_2^2} a_e, \qquad h_B^{(e)\prime} = h_B^{(e)} - \frac{d' - \hat{d}}{2\|a_e\|_2^2} a_e. \tag{7}$$

This sets the readout difference to d' while preserving the pair mean and components orthogonal to a_e. Training edits use learned depth coordinates; diagnostic edits constrain answer-gradient H/V/D readings. The target $y' = G(d', q)$ changes only for relation reversal; the image and natural QA target stay unchanged. Appendix B.3 specifies reliability checks and target sampling.

4.3 ASSIGNING FORMATION AND USE TO SEPARATE GRADIENT PATHS

The edited-answer objective trains the downstream interpretation of depth while depth supervision learns the coordinates. The shared decoder suffix and output head $f_{\theta,>e}$ continue from the edited prompt state H'_e:

$$p_\theta^e = f_{\theta,>e}(\mathrm{sg}(H'_e)) \,, \tag{8}$$

Here sg stops gradients through all prompt states and the edit, including the depth head. For edited examples $\mathcal{V}_e$, natural distribution p_θ, and answer-token cross-entropy ℓ_{ans}, optimize

$$\begin{gathered} \mathcal{L}_{\mathrm{QA}} = \mathbb{E}_{\mathcal{B}}[\ell_{\mathrm{ans}}(p_\theta, y)], \qquad \mathcal{L}_{\mathrm{use}}^{(e)} = \mathbb{E}_{\mathcal{V}_e}[\ell_{\mathrm{ans}}(p_\theta^e, G(d', q))], \\ \mathcal{L} = \mathcal{L}_{\mathrm{QA}} + \lambda_D(s)\mathcal{L}_{\mathrm{depth}} + \lambda_U(s)\mathcal{L}_{\mathrm{use}}^{(e_k)}. \end{gathered} \tag{9}$$

Each loss averages its participating terms; s is training progress. QA updates the full trainable language path; depth loss updates each head and trainable parameters through its layer; edited-answer loss updates only the downstream path. The schedule in Appendix B.4 warms up depth coordinates before ramping up edited-answer supervision. Section 5.4 tests this gradient separation. Inference uses the original model without auxiliary heads, edits, or a geometry encoder.

5 INTERNAL CHANGES AND SPATIAL QA

We test the two training goals through post-training diagnostics and spatial QA benchmarks, then examine the semantic edits, gradient separation, and early layer placement in controlled comparisons.

5.1 COMPARISON SETUP

We compare SFT, VG-LLM, and Ours on identical natural QA data using Qwen2.5-VL-3B (Bai et al., 2025b), Qwen3-VL-4B (Bai et al., 2025a), and Qwen3.5-4B (Qwen Team, 2026), separately from the initial analysis pair. Evaluation covers VSI-Bench (Yang et al., 2024), CV-Bench (Tong et al., 2024), BLINK spatial tasks (Fu et al., 2024), MMSI-Bench (Yang et al., 2025c), and ViewSpatial-Bench (Li et al., 2025a). Appendix Tables 5–7 and Section B.5 give data, configurations, evaluation, and model selection.

5.2 TRAINING CHANGES THE FORMATION AND USE OF DEPTH

We repeat Section 3's readout and six-strength interventions on frozen Qwen2.5-VL-3B checkpoints: R0 (QA), R1 (+ depth), and R2 (+ edited answers). Each uses its own layerwise answer-gradient axes, estimated separately from the training heads on the same estimation split and evaluated on the independent confirmation set.

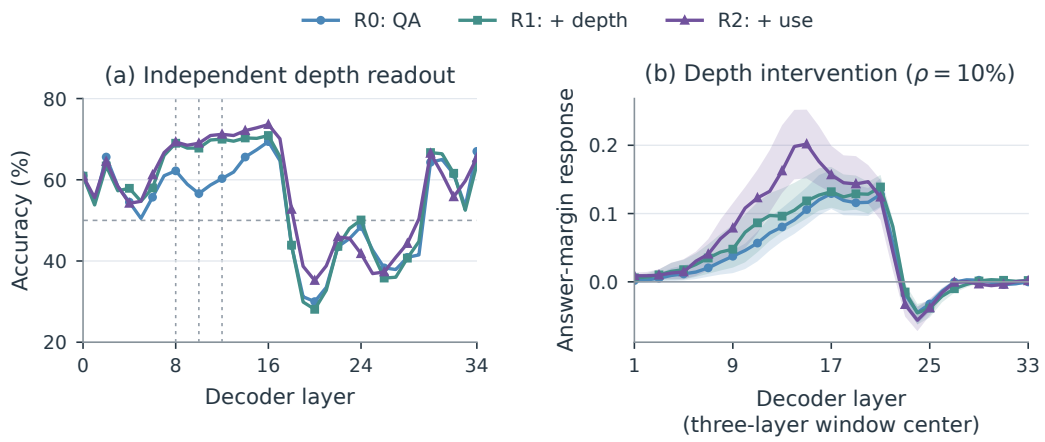

Figure 6: Post-training depth diagnostics. (a) Independent readout; guides mark supervised layers L8/L10/L12. (b) Depth-directed answer response at $\rho = 10\%$. Shading shows 95% bootstrap confidence intervals.

Depth supervision improves early availability. R1 improves readout throughout L7–L16, extending beyond the directly depth-supervised layers L8/L10/L12 (Fig. 6a). Mean accuracy at those three layers rises from 59.7% (R0) to 68.9% (R1) and 69.8% (R2).

Edited-answer supervision strengthens early answer responses. Fig. 6b shows that R1 also increases directional responses; R2 further strengthens the early-to-middle region. Across all six strengths, R2 peaks at L14–L15 versus L17–L21 for R0/R1 (Fig. 12). All groups turn negative near L24 and return toward zero later; R2's enhancement lies in the preceding positive-response region.

R1 supplies most of the early readout gain; R2 adds 0.9 points at the supervised layers and moves peak answer influence earlier, still downstream of those layers. Together, these profiles connect the training objectives to their intended internal changes: stronger early depth representations and more responsive subsequent computation.

5.3 Spatial QA improves across three backbones

Table 2: Same-data spatial QA across three backbones (point estimates in %; higher is better). Avg. is the five-benchmark mean before rounding; bold marks within-backbone column maxima, including ties.

Backbone	Training	Avg.	VSI	CV	BLINK	MMSI	View
Qwen2.5-VL-3B	SFT	50.7	45.6	77.8	63.7	26.9	39.4
	VG-LLM	53.0	46.6	81.4	68.3	**29.0**	40.0
	Ours	**54.5**	**48.1**	**84.6**	**69.6**	28.7	**41.6**
Qwen3-VL-4B	SFT	55.5	52.5	84.6	67.6	26.2	46.5
	VG-LLM	54.7	53.8	**85.7**	59.6	27.7	46.6
	Ours	**57.2**	**54.2**	**85.7**	**69.9**	**28.5**	**47.8**
Qwen3.5-4B	SFT	57.6	59.7	85.0	63.9	29.2	50.1
	VG-LLM	59.6	62.1	**86.9**	66.2	30.7	52.3
	Ours	**60.1**	**62.6**	86.4	**67.4**	**30.9**	**53.2**

In Table 2, Ours improves all five benchmark scores over same-data SFT on each backbone, raising their means by 3.8, 1.7, and 2.5 points for Qwen2.5-VL-3B, Qwen3-VL-4B, and Qwen3.5-4B, respectively. The gains hold under both full fine-tuning and LoRA.

Ours exceeds VG-LLM's mean by 1.5, 2.5, and 0.5 points, respectively, while retaining the original inference architecture. VG-LLM scores higher on Qwen2.5 MMSI and Qwen3.5 CV; Qwen3 CV is tied. Appendix C reports all eight VSI tasks for each backbone and external comparisons under their published settings.

5.4 Objectives, gradients, and layer configuration

The Qwen2.5-VL-3B ablations in Table 3 share natural QA data and ordering, initialization, fine-tuning mode, preprocessing, training steps, and model selection. Appendix B.5 specifies each group.

Table 3: Qwen2.5-VL-3B ablations (point estimates in %; higher is better). The final pair matches use-loss updates to L23–L35. Avg. uses unrounded scores; bold marks column maxima.

Aspect	Variant	Avg.	VSI	CV	BLINK	MMSI	View
Objectives	R0 (SFT)	50.7	45.6	77.8	63.7	26.9	39.4
	R1 (+ depth)	51.9	46.2	80.7	65.4	27.4	39.7
	R2 (Ours)	**54.5**	**48.1**	**84.6**	**69.6**	**28.7**	**41.6**
Controls	R1-replay	52.0	46.2	81.1	65.5	27.8	39.6
	R2-random	51.6	45.9	80.3	65.2	27.4	39.4
	R2-E2E	52.2	47.7	80.8	66.7	27.9	38.1
Placement	R3-late	52.3	46.7	81.5	67.4	26.6	39.2
	R3-early-L23	53.8	47.5	84.2	69.1	28.2	40.0
	R3-late-L23	52.1	46.1	81.5	65.8	27.5	39.6

Formation and use contribute complementary gains. Depth supervision raises the mean from 50.7 (R0) to 51.9 (R1); edited-answer supervision raises it to 54.5 (R2), with gains on every benchmark at each step. Fig. 6 shows that these QA gains accompany better early readout and earlier directional answer responses.

The edited relation supplies the additional training signal. R1-replay matches R2's branch examples, count, layer schedule, and gradient boundary, using natural states and original answers. R2-random uses equal-norm paired perturbations orthogonal to the learned depth direction, preserving depth readouts and using original answer targets. Both remain near R1, while R2 exceeds them by 2.5 and 2.9 points, supporting depth-semantic edits and corresponding answer supervision.

Separate gradients direct learning toward downstream response. R2-E2E removes detachment, allowing the edited-answer objective to change the upstream states and depth head defining the edit. Its mean falls from 54.5 to 52.2, supporting the separation of depth-coordinate learning from downstream answer-response learning.

The early joint configuration is stronger. R3-late moves both depth supervision and editing from $\{8, 10, 12\}$ to $\{18, 20, 22\}$. R2 scores higher on all five benchmarks, with a 2.2-point mean advantage. Restricting both configurations' use-loss updates to L23–L35 still favors early placement: R3-early-L23 scores 53.8 versus R3-late-L23's 52.1. Early edits traverse the intervening blocks before the common trainable suffix, supporting earlier grounded relations as input to subsequent processing and use.

6 Conclusion

We connect depth formation to spatial judgments through joint analysis of representations and answer behavior. VG-LLM exhibits earlier depth readability and answer influence than Qwen-SFT, while peak readability precedes peak answer response in both models. These findings motivate grounding depth early and teaching later layers to respond to changes in represented relations. Internal diagnostics, gains across three backbones, and controlled ablations support these complementary training targets, showing how mechanistic evidence can guide spatial reasoning improvements without additional inference components. Future work can examine how depth, orientation, and viewpoint should interact across layers when questions require several spatial relations at once.

AI USE STATEMENT

Generative AI tools were used to assist with writing refinement, manuscript organization, literature retrieval and reference checking, data preprocessing, and selected aspects of code implementation. All research decisions, experiments, reported results, and conclusions were reviewed and verified by the authors.

ETHICS STATEMENT

This work analyzes spatial reasoning in vision–language models using existing image and video datasets. Errors in spatial judgments can create risks in navigation, robotics, and other safety-sensitive applications. Such applications require validation in the intended operating environment in addition to benchmark evaluation.

REPRODUCIBILITY STATEMENT

Sections 3–5 define the representation analyses, training objectives, gradient boundaries, and controlled comparisons. Section 5.2 describes independent readout and intervention evaluation after training. Appendix A provides the complete readout and intervention curves and channel-selection procedure. Appendix B specifies data construction, optimization, loss schedules, and ablation protocols. Appendix C gives per-task VSI-Bench results and comparison conditions. To facilitate reproducibility, we also provide the code used in this study in the Supplementary Material.

REFERENCES

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, 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, 2025b. URL https://arxiv.org/abs/2502.13923.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes, 2017. URL https://arxiv.org/abs/1702.04405.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction, 2025. URL https://arxiv.org/abs/2505.20279.

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, 2024. URL https://arxiv.org/abs/2404.12390.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Ting Huang, Zhenyu Zhang, Wenyuan Huang, Jian Yang, and Hao Tang. ConsiSpace: Learning Geometric Consistency Matters for Video Spatial Reasoning, 2026. URL https://arxiv.org/abs/2607.17599.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

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, 2024. URL https://arxiv.org/abs/2408.03326.

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, 2025a. URL https://arxiv.org/abs/2505.21500.

Haoyuan Li, Qihang Cao, Tao Tang, Kun Xiang, Zihan Guo, Jianhua Han, JiaWang Bian, Hang Xu, and Xiaodan Liang. Thinking with Geometry: Active Geometry Integration for Spatial Reasoning, 2026. URL https://arxiv.org/abs/2602.06037.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive Training for Spatial Reasoning in Vision-Language Models, 2025b. URL https://arxiv.org/abs/2510.08531.

OpenAI. GPT-5 System Card, 2025. URL https://openai.com/index/gpt-5-system-card/.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond Semantics: Rediscovering Spatial Awareness in Vision-Language Models, 2025. URL https://arxiv.org/abs/2503.17349v2.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature Selection via Dependence Maximization. *Journal of Machine Learning Research*, 13(47):1393–1434, 2012. URL https://jmlr.org/papers/v13/song12a.html.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs, 2024. URL https://arxiv.org/abs/2406.16860.

Haibo Wang and Lifu Huang. Learning Geometric Representations from Videos for Spatial Intelligent Multimodal Large Language Models, 2026. URL https://arxiv.org/abs/2606.05833.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer, 2025. URL https://arxiv.org/abs/2503.11651.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM Capabilities in Visual-based Spatial Intelligence, 2025. URL https://arxiv.org/abs/2505.23747.

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. SpatioLM: Towards General Physical Spatial Intelligence in Vision-Language Models, 2026. URL https://arxiv.org/abs/2608.01899.

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, 2024. URL https://arxiv.org/abs/2412.14171.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual Spatial Tuning, 2025a. URL https://arxiv.org/abs/2511.05491.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-S: Towards Spatial Supersensing in Video, 2025b. URL https://arxiv.org/abs/2511.04670.

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, 2025c. URL https://arxiv.org/abs/2505.23764.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes, 2023. URL https://arxiv.org/abs/2308.11417.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From Flatland to Space: Teaching Vision-Language Models to Perceive and Reason in 3D, 2025a. URL https://arxiv.org/abs/2503.22976.

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. SpatialStack: Layered Geometry-Language Fusion for 3D VLM Spatial Reasoning, 2026. URL https://arxiv.org/abs/2603.27437.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. Direct Preference Optimization of Video Large Multimodal Models from Language Model Reward, 2024a. URL https://arxiv.org/abs/2404.01258.

Xiaofeng Zhang, Fanshuo Zeng, and Chaochen Gu. Simignore: Exploring and enhancing multimodal large model complex reasoning via similarity computation. *Neural Networks*, 184:107059, 2025b. doi: 10.1016/j.neunet.2024.107059.

Xiaofeng Zhang, Fanshuo Zeng, Yihao Quan, Zheng Hui, and Jiawei Yao. Enhancing Multimodal Large Language Models Complex Reason via Similarity Computation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10):10203–10211, 2025c. doi: 10.1609/aaai.v39i10.33107.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning With Synthetic Data, 2024b. URL https://arxiv.org/abs/2410.02713.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. SpaceMind: Camera-Guided Modality Fusion for Spatial Reasoning in Vision-Language Models, 2025. URL https://arxiv.org/abs/2511.23075.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from Videos for 3D World: Enhancing MLLMs with 3D Vision Geometry Priors, 2025. URL https://arxiv.org/abs/2505.24625.

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, 2025. URL https://arxiv.org/abs/2504.10479.

A Additional mechanistic analysis

A.1 Relation readout and cross-layer correspondence

For each horizontal (H), vertical (V), or depth (D) relation, 2,400 direction-estimation, 400 development, and 1,000 confirmation examples are separated by physical scene group. H/V labels use bounding-box centers; D uses object-center depth in the camera frame. At decoder block ℓ's output, $h_{i,\ell,a}$ and $h_{i,\ell,b}$ are the residual states at the last subtokens of the two object names, before final RMSNorm. Decompose them into their mean and relation state:

$$c_{i,\ell} = \frac{h_{i,\ell,a} + h_{i,\ell,b}}{2}, \qquad r_{i,\ell} = h_{i,\ell,a} - h_{i,\ell,b}. \tag{10}$$

Holding c and all other token states fixed, reconstruct the pair as $c \pm r/2$. Let $F_{i,\ell,g}(r)$ be the final answer-logit difference between the positive and negative semantic endpoints of relation g. The chain rule gives

$$\nabla_r F_{i,\ell,g} = \tfrac{1}{2}\left(\nabla_{h_a} F_{i,\ell,g} - \nabla_{h_b} F_{i,\ell,g}\right). \tag{11}$$

The gradient measures the local sensitivity of the final relation score to changes in r. Averaging over the N_g direction-estimation examples and normalizing gives a fixed relation axis and its sign prediction:

$$\bar{v}_{\ell,g} = \frac{1}{N_g}\sum_i \nabla_r F_{i,\ell,g}(r_{i,\ell}), \qquad u_{\ell,g} = \frac{\bar{v}_{\ell,g}}{\|\bar{v}_{\ell,g}\|_2}, \qquad \hat{y}_{i,\ell,g} = \operatorname{sign}(u_{\ell,g}^\top r_{i,\ell}). \tag{12}$$

Directions are fitted separately for each model, layer, and relation; geometry-label signs match the answer semantics. Positive D means that a is closer than b. Development sets determine analysis settings, and independent confirmation examples provide accuracy and balanced accuracy. Under this fixed polarity, below-50% accuracy indicates an inverse correspondence with ground-truth ordering along the measured axis. The plotted readout measurements cover L0–L34.

For CKA, we use 600 CV-Bench Depth questions (Tong et al., 2024) and 600 non-spatial reasoning samples from LLaVA-Hound (Zhang et al., 2024a). Object-token states from corresponding samples in matching order form representation matrices X, Y. Center them across samples and form $K = \bar{X}\bar{X}^\top$ and $G = \bar{Y}\bar{Y}^\top$. Let $\widetilde{K}, \widetilde{G}$ be these Gram matrices with zero diagonals. Following Song et al. (2012), for N samples,

$$\begin{aligned}\mathrm{HSIC}_u(K,G) = \frac{1}{N(N-3)}\Bigg[\operatorname{tr}(\widetilde{K}\widetilde{G}) + \frac{(\mathbf{1}^\top\widetilde{K}\mathbf{1})(\mathbf{1}^\top\widetilde{G}\mathbf{1})}{(N-1)(N-2)} \\ - \frac{2}{N-2}\mathbf{1}^\top\widetilde{K}\widetilde{G}\mathbf{1}\Bigg].\end{aligned} \tag{13}$$

The normalized cross-layer measure is

$$M_{\ell k} = \mathrm{CKA}_{\mathrm{deb}}(X_\ell^{\mathrm{SFT}}, X_k^{\mathrm{VG}}) = \frac{\mathrm{HSIC}_u(K_\ell, G_k)}{\sqrt{\mathrm{HSIC}_u(K_\ell, K_\ell)\,\mathrm{HSIC}_u(G_k, G_k)}}. \tag{14}$$

Here K_ℓ and G_k are the centered linear Gram matrices for Qwen-SFT layer ℓ and VG-LLM layer k. Each 36×36 matrix covers L0–L35, with Qwen-SFT rows and VG-LLM columns. The box in Fig. 2b marks VG-LLM L8 against Qwen-SFT L8–L15. Higher values indicate more similar relations among the same input samples.

A.2 Complete spatial-readout results

Fig. 7 and Table 4 give the full H/V/D profiles and their summary statistics.

Table 4: Spatial readout accuracy (%). Mean averages L0–L34; Peak gives the best measured layer; Window mean is the best consecutive three-layer mean. Layer indices are zero-based.

Axis	Model	Mean	Peak	Layer	Window mean	Window
H	Qwen-SFT	60.35	98.20	21	98.00	21–23
H	VG-LLM	59.99	98.30	22	97.73	22–24
V	Qwen-SFT	65.21	89.20	22	87.93	21–23
V	VG-LLM	64.40	90.20	22	88.37	20–22
D	Qwen-SFT	52.49	70.00	17	67.33	15–17
D	VG-LLM	51.01	74.00	8	71.67	7–9

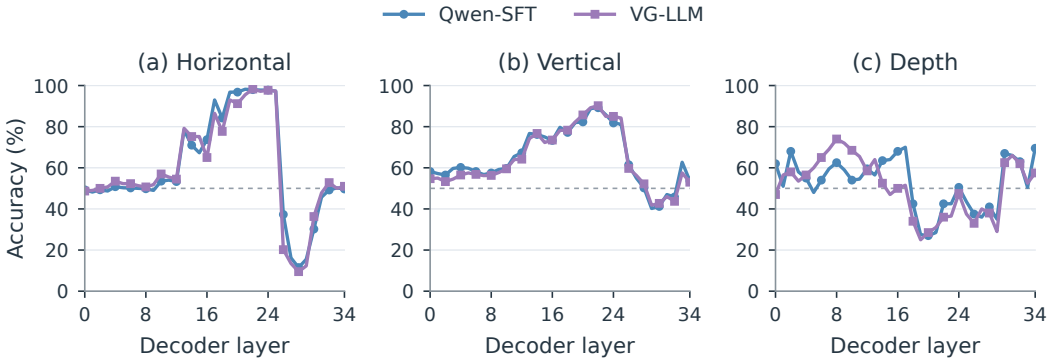

Figure 7: H/V/D readout profiles. The dashed line marks 50% accuracy.

A.3 Intervention calibration and complete directional responses

The matrix $B_\ell = [u_{\ell,H}, u_{\ell,V}, u_{\ell,D}]^\top$ stacks the three unit relation directions as rows. We seek the minimum-norm $w_{\ell,D}$ satisfying $B_\ell w_{\ell,D} = e_D$, where $e_D = (0,0,1)^\top$. For linearly independent directions,

$$w_{\ell,D} = B_\ell^\top (B_\ell B_\ell^\top)^{-1} e_D. \tag{15}$$

For $s \in \{-1, +1\}$, the relation-state edit $r^{(s)} = r + s\eta w_{\ell,D}$ changes D by $s\eta$ and leaves edited-layer H/V readings fixed. Reconstructing $c \pm r^{(s)}/2$ gives

$$h_a^{(s)} = h_a + \frac{s\eta}{2} w_{\ell,D}, \qquad h_b^{(s)} = h_b - \frac{s\eta}{2} w_{\ell,D}, \tag{16}$$

where sample and layer indices on h are omitted. The image, question, model parameters, pair mean, and other token states stay fixed; subsequent layers run normally.

For entity presentation order $\pi \in \{(a,b),(b,a)\}$ and object $o \in \{a,b\}$, the relative state change is $\eta \|w_{\ell,D}\|_2 / (2\|h_{i,\ell,o}^{(\pi)}\|_2)$. The calibrated amplitude is

$$\eta_{\ell,\rho} = \frac{2\rho}{\mathrm{median}_{i,o,\pi}\left(\|w_{\ell,D}\|_2 / \|h_{i,\ell,o}^{(\pi)}\|_2\right)}. \tag{17}$$

The median spans calibration samples, both objects, and both orders, separately for each model and layer. We use $\rho \in \{2.5, 5, 7.5, 10, 15, 20\}\%$.

Let $M^{s,\pi}_{i,\ell,\rho}$ be the edited answer-logit margin for the fixed relation "a closer than b" under perturbation sign s and entity order $\pi \in \{(a,b),(b,a)\}$. Answer logits are remapped to these identities across question orders and near/far wording. Both signs start from the same unedited state, with margin $M^{0,\pi}_i$. The Directional Utilization Score is

$$\begin{aligned} U_{i,\ell,\rho} &= \frac{1}{4}\sum_{\pi}\left(M^{+,\pi}_{i,\ell,\rho} - M^{-,\pi}_{i,\ell,\rho}\right) \\ &= \frac{1}{4}\sum_{\pi}\sum_{s\in\{-1,+1\}} s\left(M^{s,\pi}_{i,\ell,\rho} - M^{0,\pi}_{i}\right). \end{aligned} \tag{18}$$

Curves show the mean over N examples, $\bar{U}_{\ell,\rho} = N^{-1}\sum_i U_{i,\ell,\rho}$. Positive and negative values indicate movement toward and against the intended relation change. Random controls use equal-norm directions in the orthogonal complement of the H/V/D span. Shading in the intervention figures denotes 95% bootstrap confidence intervals.

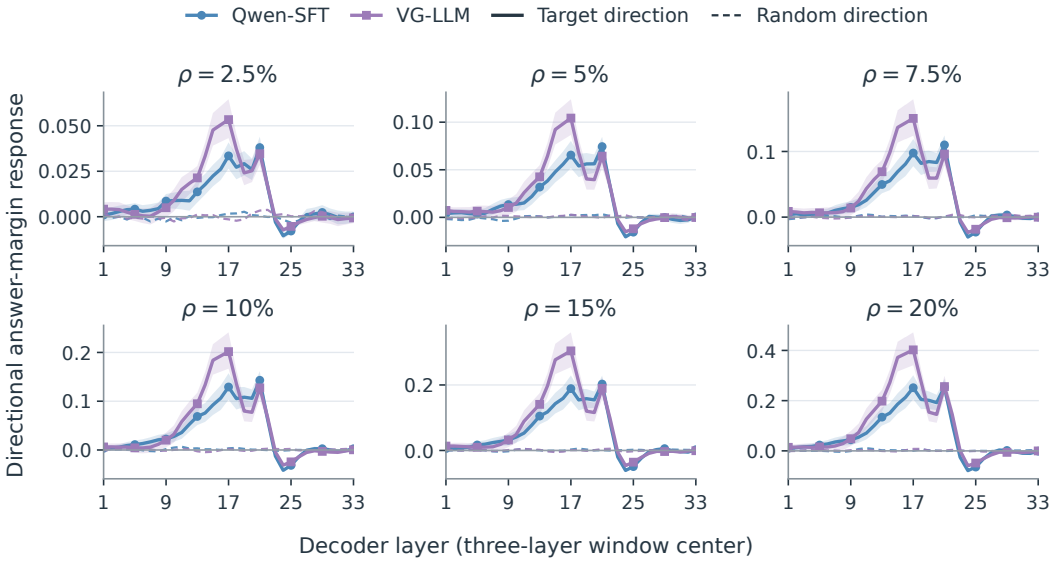

Figure 8: Depth-directed answer responses across six strengths, aggregated over 80 scenes.

The H/V comparisons target left/right and up/down relations using the corresponding basis vector in place of e_D. All three directions use the same relative-magnitude calibration, directional scoring, and three-layer window aggregation, with aligned entity orders and perturbation signs. Fig. 9 and Fig. 10 extend the main-text H/V comparison to all six strengths.

A.4 Attribution and Fixed-Channel Interventions

Prior analyses of image–text information flow motivate similarity-based visual-token filtering (Zhang et al., 2025b;c). We estimate channel contributions on SPAR spatial questions with frozen model parameters. For input x, candidate-answer letters $\mathcal{A}_x$, correct letter g_x, and candidate-token logit $L_a(x)$ at the last valid prefill position, the attribution target is

$$m(x) = L_{g_x}(x) - \left[\operatorname{logsumexp}_{a\in\mathcal{A}_x\setminus\{g_x\}} L_a(x) - \log(|\mathcal{A}_x| - 1)\right]. \tag{19}$$

Index j identifies module, layer, and channel. FFN activations follow gating and enter `down_proj`; attention activations enter `o_proj` before projection into the residual. Let $z_{t,j}(x)$ be the corresponding channel activation, and let $\mathcal{P}_x$ include all valid prefill positions for the visual input, ques-

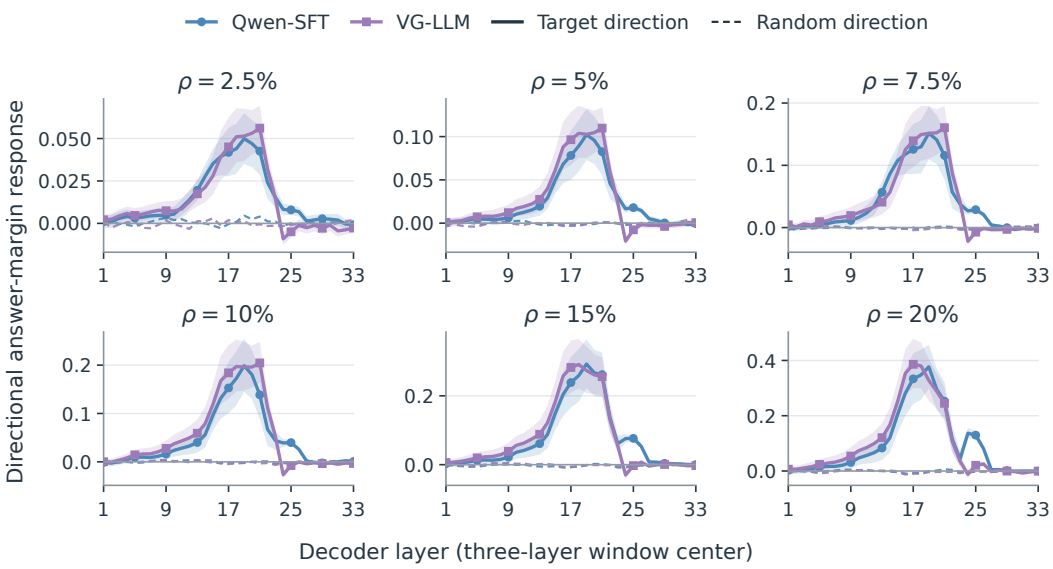

Figure 9: Horizontal-directed answer responses across six strengths, aggregated over 80 scenes.

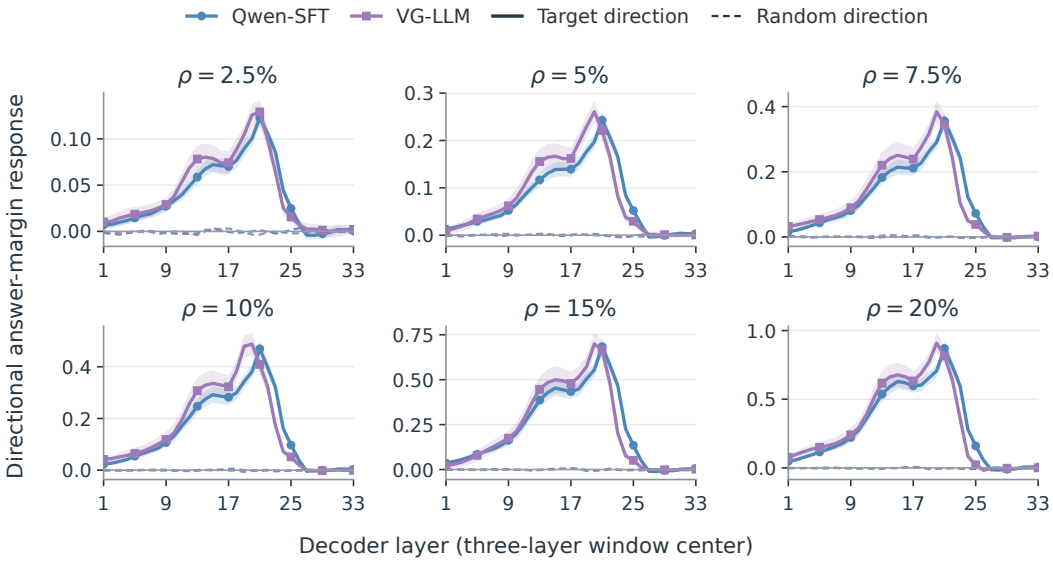

Figure 10: Vertical-directed answer responses across six strengths.

tion, and options. The signed activation–gradient contribution is

$$s_j(x) = \sum_{t \in \mathcal{P}_x} z_{t,j}(x) \frac{\partial m(x)}{\partial z_{t,j}(x)}. \tag{20}$$

If all these channel activations are scaled by γ, their margin $m_j(x;\gamma)$ satisfies

$$\left.\frac{\partial m_j(x;\gamma)}{\partial \gamma}\right|_{\gamma=1} = s_j(x), \qquad m_j(x;0) - m_j(x;1) \approx -s_j(x). \tag{21}$$

This first-order relation gives the attribution sign its interpretation; actual generation is evaluated after fixed-set removal.

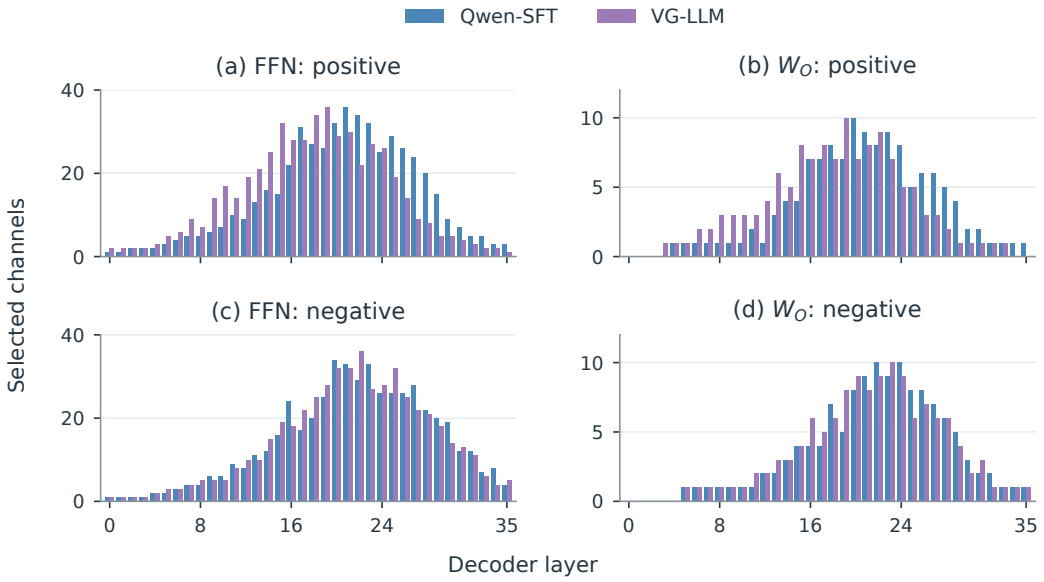

Figure 11: Spatial-channel counts by layer. Top/bottom: positive/negative sets. Left: FFN (512 channels per model and sign); right: W_O (128).

For original question i, target variants $\mathcal{V}_i$ combine option order, entity order, and spatial-relation wording with semantically correct answer remapping. Non-spatial controls $\mathcal{C}_i$ retain the visual input and QA format but ask for text information, attribute binding, or simple arithmetic specified by explicit textual facts. Aggregate within each original question:

$$T_{ij} = \frac{1}{|\mathcal{V}_i|} \sum_{x \in \mathcal{V}_i} s_j(x), \qquad C_{ij} = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} |s_j(x)|. \tag{22}$$

Controls take the absolute value after summing over positions for each input. Across original questions, $\bar{T}_j = \mathbb{E}_i[T_{ij}]$ and $\bar{C}_j = \mathbb{E}_i[C_{ij}]$. For candidate sign $p = \text{sign}(\bar{T}_j)$, define $\mu_j^{(p)} = p\bar{T}_j - 0.7\bar{C}_j$ and rank by

$$R_j^{(p)} = \mu_j^{(p)} - \text{SE}\left(\mu_j^{(p)}\right). \tag{23}$$

The standard error uses original-question paired contributions $pT_{ij} - 0.7C_{ij}$, grouped by physical scene to retain within-scene dependence. For each model, candidates are separated by the sign of $\bar{T}_j$ and ranked across layers within each module. We select 512 positive and 512 negative FFN channels, and 128 of each sign for W_O.

Fig. 11 reports all four raw-count comparisons. For each set of size K, Fig. 4 shows $100(n_\ell^{\text{VG}} - n_\ell^{\text{SFT}})/K$ in percentage points, where n_ℓ counts selected channels at layer ℓ and K is 512 for FFN or 128 for W_O.

For selected set $\mathcal{M}$, gate $z_{t,j}(x)$ to zero when $j \in \mathcal{M}$ and $t \in \mathcal{P}_x$. Both visual and text prefill positions are affected; subsequent decoding steps receive no direct gate. We evaluate generated answers on CV-Bench using seed 0, temperature 0, and greedy decoding. Equal-size random sets provide controls. Table 1 reports the analysis models' baselines and results; Table 2 reports the separately retrained comparisons.

A.5 Post-training depth interventions at all strengths

For Qwen2.5-VL-3B, R0 uses QA supervision, R1 adds early depth regression, and R2 also adds edited-answer supervision. With each checkpoint frozen, we estimate its own answer-gradient axes on the direction-estimation split and measure readout on the independent confirmation set. We then apply the paired interventions, calibration, and score in Equations 16, 17, and 18. Fig. 12 extends Fig. 6b to all six strengths. R2's earlier positive-response peak persists across the sweep.

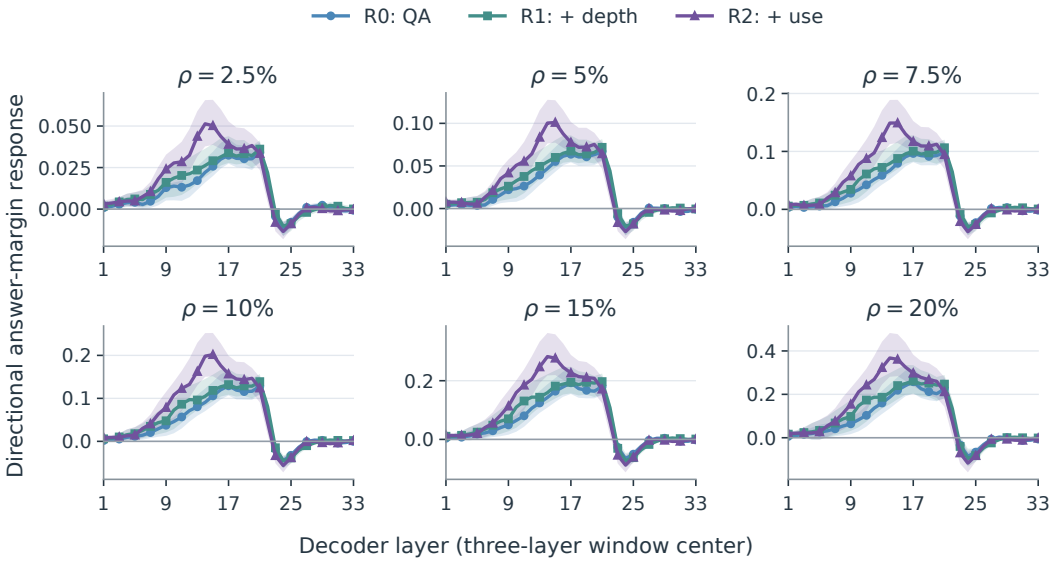

Figure 12: Post-training depth-directed answer responses across six strengths.

B Training protocol and feasible semantic edits

This appendix specifies the shared data, three backbone configurations, target sampling, optimization, and Qwen2.5-VL-3B controls.

B.1 Data construction and depth labels

Table 5 specifies the shared mixture. Its base follows VG-LLM (Zheng et al., 2025). SPAR provides spatial QA derived from 3D scenes (Zhang et al., 2025a), including ScanNet and ScanNet++ (Dai et al., 2017; Yeshwanth et al., 2023). The Hound split of LLaVA-Video supplies general video QA (Zhang et al., 2024a;b). Every comparison group receives the same natural QA examples.

Table 5: Shared training data. Base components count QA records; the additional set counts unique entity pairs. Question and option variants belong to their original entry.

Source	Examples	Supervision
SPAR-234K	234,277	Single-image, multi-view, and video spatial QA.
LLaVA-Hound-64K	63,750	General video QA.
Additional SPAR-7M depth set	30,000	Natural QA, depth regression, and edited-answer supervision.
Total	328,027	Same natural QA mixture across groups.

SPAR splits use physical scene groups: scans, images, and question variants from a physical scene share its identifier. Scenes used for model selection and independent evaluation are excluded from both base spatial QA and the additional set. Direction estimation, development calibration, and confirmation use disjoint scene groups. Video QA is grouped by source video.

Selecting depth pairs. We use single-image `depth_prediction_oc` records from ScanNet and ScanNet++. Reference depth in the original question and the original answer provide the two entities' absolute depths. Parsing covers numerical answers and 48 verified absolute-depth sentence formats, with unambiguous entity–marker correspondence and consistent entity labels. Both camera-axis object-center depths are positive and differ by at least $0.2\,\mathrm{m}$, checked at the original annotation precision.

We deduplicate by original image and unordered entity-marker pair, excluding pairs already jointly present in the base QA. Added examples may reuse an image with a different pair; there are at most five distinct object pairs per image. Grouped sampling and soft balancing across sources, scenes, categories, depth ranges, and depth gaps control concentration.

Vertical strata and question construction. In a common reference image view, let $\Delta_y = (y^{\mathrm{ref}}_{\mathrm{near}} - y^{\mathrm{ref}}_{\mathrm{far}})/H_{\mathrm{ref}}$, where y^{ref} is the marker-center coordinate increasing downward and H_{ref} is image height. Sampling assigns 25% to $\Delta_y < -0.1$, 25% to $\Delta_y > 0.1$, and 50% to $|\Delta_y| \le 0.1$. These strata put the nearer object above, below, or at a similar height to the farther object. All backbones share the auxiliary example list and apply their own preprocessing, with both markers fully visible.

Each example asks which of two marked objects' 3D centers is closer to or farther from the camera along its depth axis. Numerical depths are retained as auxiliary labels; the prompt excludes these values and any text revealing the relation. Marker colors, entity mention order, near/far wording, and option positions are balanced. Marker descriptions precede object names, such as `red-marked doors`. A/B follow first-mention order in the opening comparison sentence. Name character spans locate their last subtokens after tokenization, chat templating, and visual-token expansion.

Numerical targets. For positive camera-axis depth d_o, the depth coordinates and pair labels are

$$u_o = -\log \max(d_o, \varepsilon_d), \qquad t_o = \frac{u_o - \mu_{\mathrm{train}}}{\sigma_{\mathrm{train}}}, \qquad r^\star = t_A - t_B. \tag{24}$$

Deduplicate labels by image and entity marker. Set ε_d to the minimum valid depth among unique training entities; μ_{train} and σ_{train} are the mean and population standard deviation of their u_o. The same statistics serve every backbone, group, and supervised layer. Same-view label differences form the shared empirical target pool $\mathcal{D}_{\mathrm{emp}}$.

For a door A at $3.6\,\mathrm{m}$ and a table B at $2.9\,\mathrm{m}$, $r^\star = \log(2.9/3.6)/\sigma_{\mathrm{train}} < 0$: B is closer. The function $G(r^\star, q)$ maps this relation through the question wording and option identities. Exchanging entity roles changes the sign and the corresponding answer mapping.

B.2 Backbones and parameter updates

All conditions start from checkpoints before this spatial fine-tuning. For SFT, Ours, and the ablations, Qwen2.5-VL-3B updates all language parameters, including embeddings, decoder blocks, normalization, and output head; Qwen3-VL-4B and Qwen3.5-4B train LoRA adapters (Hu et al., 2021) with their original language weights frozen. These configurations freeze the visual encoder and visual-feature merger.

Table 6: Backbone configurations. Layer indices are zero-based; learning rates apply to the trainable language parameters.

Setting	Qwen2.5-VL-3B	Qwen3-VL-4B	Qwen3.5-4B
Decoder blocks	36	36	32
Language tuning	Full	LoRA	LoRA
Learning rate	10^{-5}	5×10^{-5}	5×10^{-5}
Early set $\mathcal{E}$	$\{8, 10, 12\}$	$\{8, 10, 12\}$	$\{6, 8, 10\}$
Late set $\mathcal{E}_{\mathrm{late}}$	$\{18, 20, 22\}$	$\{18, 20, 22\}$	$\{16, 18, 20\}$
LoRA rank / α	—	64 / 128	64 / 128
LoRA dropout	—	0.05	0.05

LoRA targets language attention projections `q_proj`, `k_proj`, `v_proj`, `o_proj` and MLP projections `gate_proj`, `up_proj`, `down_proj`. In Qwen3.5's full-attention blocks, these attention projections receive adapters; its linear-attention blocks additionally use `in_proj_qkv`, `in_proj_z`, `in_proj_b`, `in_proj_a`, and `out_proj`. MLP adapters span all decoder blocks. The LoRA scale is $\alpha/r = 2$.

VG-LLM patch alignment and fusion. VG-LLM follows its published geometry-branch and fine-tuning configuration (Zheng et al., 2025), adapted to each backbone with the shared training

mixture. Qwen2.5-VL and VGGT share a 14×14 patch size. For Qwen3-VL and Qwen3.5, we retain their 16×16 patches and VGGT's 14×14 patches. Both branches use the same image view after common resizing and any cropping. From Qwen's `image_grid_thw`, we obtain the spatial grid (h, w) and resize this view to $14h \times 14w$ for VGGT, matching Qwen's grid for its $16h \times 16w$ input. Both h and w are even for Qwen's 2×2 merger.

After removing VGGT's camera/register tokens, we concatenate each 2×2 patch block in row-major order. RMSNorm followed by an MLP ($4C_G \to 4096 \to D$, with C_G the VGGT feature width and D the language hidden size) yields a geometry grid of $(h/2) \times (w/2)$ tokens. We add these features positionwise to Qwen's merged visual features, preserving its visual-token count, ordering, and positional-encoding interface.

Depth coordinates and supervision. Let $\mathcal{E} = \{\ell_1, \dots, \ell_M\}$ be the ordered supervised layer set in Table 6. For entity o, $h_o^{(\ell)}$ is its last name-subtoken state at block ℓ's output, before final RMSNorm. Each layer has an independent affine head shared across entities:

$$z_o^{(\ell)} = a_\ell^\top h_o^{(\ell)} + b_\ell, \qquad \hat{d}^{(\ell)} = z_A^{(\ell)} - z_B^{(\ell)}. \tag{25}$$

For natural-QA batch $\mathcal{B}$, let $\mathcal{B}_{\text{aux}}$ contain the examples with depth labels. All selected layers use the same targets on the natural pass:

$$\mathcal{L}_D^{(\ell)} = \mathbb{E}_{\text{obj}}\Big[\text{Huber}_1(z_o^{(\ell)} - t_o)\Big] + \beta \mathbb{E}_{\text{pair}}\Big[\text{Huber}_1(\hat{d}^{(\ell)} - r^\star)\Big], \tag{26}$$

$$\mathcal{L}_{\text{depth}} = \frac{1}{M} \sum_{\ell \in \mathcal{E}} \mathcal{L}_D^{(\ell)}. \tag{27}$$

The expectations average entities and same-view pairs in $\mathcal{B}_{\text{aux}}$; Huber_1 has transition 1. Averaging across layers keeps the auxiliary weight comparable across choices of $\mathcal{E}$.

B.3 Reliable coordinates and feasible target sampling

Once editing is enabled, step $k = 0, 1, \dots$ selects $e_k = \ell_{1+(k \bmod M)}$. All extra branches at that step use $e = e_k$. For eligible single-image, same-view near/far questions, write $z_o = z_o^{(e)}$ and $\hat{d} = z_A - z_B$. Reliable readouts satisfy

$$\begin{aligned} &\text{sign}(\hat{d}) = \text{sign}(r^\star), \qquad \min\{|\hat{d}|, |r^\star|\} \geq \tau, \\ &|\hat{d} - r^\star| \leq \delta_{\text{pair}}, \qquad \max_{o \in \{A,B\}} |z_o - t_o| \leq \delta_{\text{obj}}. \end{aligned} \tag{28}$$

These conditions check relation direction, margin, and numerical error at the selected layer.

Feasibility of paired edits. For a candidate target difference d', the paired edit is

$$h_A^{(e)\prime} = h_A^{(e)} + \frac{d' - \hat{d}}{2\|a_e\|_2^2} a_e, \qquad h_B^{(e)\prime} = h_B^{(e)} - \frac{d' - \hat{d}}{2\|a_e\|_2^2} a_e. \tag{29}$$

The affine bias cancels in $\hat{d} = a_e^\top (h_A^{(e)} - h_B^{(e)})$. The edited difference is therefore

$$\tilde{d} = \hat{d} + \frac{d' - \hat{d}}{\|a_e\|_2^2} a_e^\top a_e = d'. \tag{30}$$

Equal-and-opposite shifts preserve the pair mean, and shifts parallel to a_e preserve each state's orthogonal component. Their common norm is

$$\|h_o^{(e)\prime} - h_o^{(e)}\|_2 = \frac{|d' - \hat{d}|}{2\|a_e\|_2}, \qquad o \in \{A, B\}. \tag{31}$$

Let $H_o = \max(\|h_o^{(e)}\|_2, \varepsilon_h)$, with numerical stabilizer $\varepsilon_h > 0$. Requiring both relative changes to be at most $\rho_{\max}$ is equivalent to

$$|d' - \hat{d}| \leq R_e, \qquad R_e = 2\rho_{\max} \|a_e\|_2 \min(H_A, H_B). \tag{32}$$

Thus the feasible interval $[\hat{d} - R_e, \hat{d} + R_e]$ is available before constructing the edited state.

Conditional empirical targets. Let $P_{\text{emp}}(v)$ be the probability mass of difference v in $\mathcal{D}_{\text{emp}}$, combining the frequencies of repeated values in training entries. For reliable example x, let $s_x = \text{sign}(\hat{d}) = \text{sign}(r^\star)$. Define

$$\begin{aligned}\mathcal{C}_{e,\text{keep}}(x) &= \{v \in \mathcal{D}_{\text{emp}} : \text{sign}(v) = s_x,\ v \neq \hat{d},\ |v - \hat{d}| \leq R_e\},\\ \mathcal{C}_{e,\text{flip}}(x) &= \{v \in \mathcal{D}_{\text{emp}} : \text{sign}(v) = -s_x,\ |v - \hat{d}| \leq R_e\}.\end{aligned} \tag{33}$$

Choose $c \in \{\text{keep}, \text{flip}\}$ with probability $1/2$ each. For a nonempty selected set, sample

$$P(d' \mid x, e, c) = \frac{P_{\text{emp}}(d')\mathbf{1}\{d' \in \mathcal{C}_{e,c}(x)\}}{\sum_{v \in \mathcal{C}_{e,c}(x)} P_{\text{emp}}(v)}. \tag{34}$$

Examples with constructed edits form $\mathcal{V}_e$. Each uses one target and one paired edit with answer $y' = G(d', q)$. Preservation changes the depth difference while retaining the answer; reversal changes both. The image stays unchanged. All examples retain their natural QA supervision, and auxiliary examples retain depth supervision.

B.4 Optimization and gradient paths

Let H'_e be the complete prompt state after the paired edit. The shared decoder suffix and output path $f_{\theta,>e}$ produce

$$p^e_\theta = f_{\theta,>e}(\text{sg}(H'_e))\,, \tag{35}$$

where sg stops gradients through all prompt states and the edit, including the depth head. For natural distribution p_θ, let ℓ_{ans} be mean target-answer-token cross-entropy over the full vocabulary. The natural and edited branches optimize

$$\begin{aligned}\mathcal{L}_{\text{QA}} &= \mathbb{E}_{\mathcal{B}}[\ell_{\text{ans}}(p_\theta, y)],\\ \mathcal{L}^{(e)}_{\text{use}} &= \mathbb{E}_{\mathcal{V}_e}[\ell_{\text{ans}}(p^e_\theta, G(d', q))],\\ \mathcal{L} &= \mathcal{L}_{\text{QA}} + \lambda_D(s)\mathcal{L}_{\text{depth}} + \lambda_U(s)\mathcal{L}^{(e_k)}_{\text{use}}.\end{aligned} \tag{36}$$

Each loss averages its participating terms; s denotes training progress. QA updates the full trainable language path. Depth loss updates each head and trainable parameters through its layer. The edited-answer loss updates only downstream trainable parameters: the decoder suffix and output head under full tuning, or downstream adapters under LoRA. Table 7 collects the shared settings.

Table 7: Shared training and evaluation settings. Numerical thresholds use standardized depth units.

Setting	Value
Depth objective	Huber transition 1; pair weight $\beta = 0.5$.
Auxiliary weights	$\lambda_D^{\max} = \lambda_U^{\max} = 0.1$.
Reliability thresholds	$\tau = 0.1$; $\delta_{\text{pair}} = \delta_{\text{obj}} = 0.5$.
Target-type probability	0.5 each for preservation and reversal; conditional empirical sampling.
Relative state-change cap	$\rho_{\max} = 0.10$ per entity.
Depth-head learning rate	10^{-3}.
Training budget	Global natural-QA batch 64; 1 epoch.
Hardware	All experiments were conducted on four NVIDIA A100 GPUs (40 GB each).
Training / evaluation seed	0.
Evaluation decoding	Greedy; temperature 0.
Optimizer	AdamW; 3% learning-rate warmup; cosine decay.
Weight decay	0.01 for trainable language parameters; 0 for depth heads.

For training progress $s \in [0, 1]$, the auxiliary schedule is

$$\lambda_D(s) = \begin{cases}0.1, & 0 \leq s \leq 0.9,\\ 1 - s, & 0.9 < s \leq 1,\end{cases} \qquad \lambda_U(s) = \begin{cases}0, & 0 \leq s \leq 0.1,\\ s - 0.1, & 0.1 < s < 0.2,\\ 0.1, & 0.2 \leq s \leq 0.9,\\ 1 - s, & 0.9 < s \leq 1.\end{cases} \tag{37}$$

This schedule warms up depth readouts, introduces edited-answer supervision, and tapers both auxiliary objectives at the end of training. Examples in $\mathcal{V}_{e_k}$ execute the extra branch once editing is enabled at 10% progress.

B.5 MATCHED TRAINING GROUPS AND EVALUATION

All groups receive the same natural QA mixture, including the auxiliary set's normal answers. Table 8 summarizes objective, layer, and gradient changes; shared settings follow Tables 6–7.

Table 8: Qwen2.5-VL-3B training groups. R0 is SFT and R2 is Ours. Early/late sets are $\{8, 10, 12\}$ and $\{18, 20, 22\}$.

Group	Configuration
R0	Natural QA on the complete mixture.
R1	R0 plus depth supervision at all early layers.
R2	R1 plus one edited-answer branch, rotating over early layers.
R3-late	Move both R2 depth supervision and editing to the late set.
R1-replay	R1 plus a matched unedited-state branch with the original answer.
R2-random	R1 plus equal-norm paired random perturbations orthogonal to the learned depth direction, with the original answer.
R2-E2E	R2 without stop-gradient; the use loss differentiates through the edit.
R3-early-L23	Depth supervision and editing at L8/L10/L12; use-loss updates only L23–L35.
R3-late-L23	Depth supervision and editing at L18/L20/L22; use-loss updates only L23–L35.

Natural-state replay. R1-replay uses R2's branch examples and count, layer schedule, and gradient boundary, but supplies natural states and original answers.

Equal-norm random perturbations. R2-random follows R2's readout filtering and feasible target sampling. For a sampled target d', let $\kappa = |d' - \hat{d}|/(2\|a_e\|_2)$ be the per-entity edit norm in Equation (29). At the selected layer e, sample $v_0 \sim \mathcal{N}(0, I)$ and project it orthogonally to the learned depth-head direction:

$$v = v_0 - \frac{v_0^\top a_e}{\|a_e\|_2^2} a_e, \qquad \hat{v} = \frac{v}{\|v\|_2}, \\ \widetilde{h}_A^{(e)} = h_A^{(e)} + \kappa \hat{v}, \qquad \widetilde{h}_B^{(e)} = h_B^{(e)} - \kappa \hat{v}. \tag{38}$$

These equal-and-opposite shifts preserve the pair mean and match the corresponding semantic edit's per-entity norm. Since $a_e^\top \hat{v} = 0$, both affine depth readouts and their difference remain unchanged. The extra branch uses answer-token cross-entropy with the original answer y. Depth supervision, layer scheduling, stop-gradient, and loss weights follow R2.

Gradient and layer controls. R2-E2E removes the stop-gradient boundary, letting the use loss differentiate through the paired edit, depth head, and upstream states. In both L23 groups, QA and depth loss retain their original gradient paths; only use-loss updates are restricted to decoder blocks L23–L35.

BLINK-Spatial includes relative-depth, spatial-reasoning, and multi-view-reasoning tasks (Fu et al., 2024; Zheng et al., 2025). Evaluation uses the decoding settings in Table 7, with numerical parsing for numerical-answer tasks. Independent calibration-scene QA determines model selection.

C VSI-BENCH RESULTS AND GEOMETRY RESOURCES

Numerical answers use mean relative accuracy: average the indicator of $|\hat{y} - y|/y < 1 - \theta$ over $\theta \in \{0.50, 0.55, \ldots, 0.95\}$. Multiple-choice tasks use accuracy. Our same-data Avg. averages all eight VSI-Bench task scores (Yang et al., 2024) (Table 9).

Table 9: VSI-Bench scores (%; higher is better) and geometry resources. External rows use published settings; same-data blocks share a backbone and QA mixture. Bold marks within-block maxima, including ties. 3D: additional raw 3D data; Tr.: training geometry encoder/teacher; Inf.: inference geometry encoder. Superscript s marks scores from Wang & Huang (2026); Human and LLaVA-OV scores are from Yang et al. (2024).

	Geometry				Numerical answer				Multiple choice			
Model	3D	Tr.	Inf.	Avg.	OC	AD	OS	RS	RD	Dir	RP	AO
Human and proprietary models												
Human	N/A	N/A	N/A	79.2	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0
Gemini-2.5-Pro^{s} (Comanici et al., 2025)	NR	NR	NR	53.5	46.0	37.3	68.7	54.3	61.9	43.9	47.4	68.7
GPT-5^{s} (OpenAI, 2025)	NR	NR	NR	55.0	53.3	34.4	73.3	47.5	63.7	48.6	50.2	68.9
Open general models												
LLaVA-OV-72B (Li et al., 2024)	N	N	N	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
InternVL3-8B^{s} (Zhu et al., 2025)	N	N	N	42.1	66.0	34.8	43.6	47.5	48.0	39.3	26.2	31.3
Qwen2.5-VL-7B-I^{s} (Bai et al., 2025b)	N	N	N	31.4	40.9	14.8	43.4	10.7	38.6	40.1	33.0	29.8
Qwen3-VL-2B-I^{s} (Bai et al., 2025a)	N	N	N	50.3	62.1	40.2	71.4	49.7	52.2	42.0	30.4	54.5
Qwen3-VL-8B-I^{s} (Bai et al., 2025a)	N	N	N	57.9	67.5	47.0	76.3	61.9	58.0	50.9	35.0	66.3
Spatial models												
SpatialLadder-3B (Li et al., 2025b)	Y†	N	N	45.7	63.5	34.3	61.7	43.9	45.4	44.8	35.6	36.4
Spatial-MLLM-4B (Wu et al., 2025)	Y†	Y	Y	47.0	65.3	34.8	63.1	45.1	41.3	46.9	33.5	46.3
SpatialStack-4B (Zhang et al., 2026)	N	Y	Y	60.9	69.2	45.4	63.0	63.2	57.9	68.4	40.2	79.6
VLM-3R-7B (Fan et al., 2025)	Y†	Y	Y	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
SpaceMind-8B (Zhao et al., 2025)	N	Y	Y	69.6	73.3	61.4	77.3	74.2	67.2	88.4	44.3	70.6
Cambrian-S-3B^{s} (Yang et al., 2025b)	Y†	Y†	N	57.3	70.7	40.6	68.0	46.3	64.8	61.9	27.3	78.8
VST-3B-SFT (Yang et al., 2025a)	Y†	Y†	N	57.9	69.3	45.4	71.8	62.4	59.0	46.0	38.7	70.2
GeoVR-4B (Wang & Huang, 2026)	N	Y	N	69.6	65.6	58.1	75.5	71.6	70.0	89.3	41.2	85.6
GeoThinker-8B (Li et al., 2026)	N	Y	Y	72.6	NR	NR	NR	NR	NR	NR	NR	NR
ConsiSpace-8B (Huang et al., 2026)	Y†	Y	Y	76.6	77.9	78.9	82.2	76.3	72.1	90.1	54.7	79.9
SpatioLM-8B (Wu et al., 2026)	N	Y	N	71.6	81.9	55.7	75.2	71.5	78.2	74.4	47.9	87.8
Same-data Qwen2.5-VL-3B (Bai et al., 2025b)												
SFT	N	N	N	45.6	67.2	34.8	55.3	52.6	42.8	45.3	**32.5**	34.3
VG-LLM (Zheng et al., 2025)	N	Y	Y	46.6	67.5	**37.9**	55.0	52.8	**49.3**	43.8	31.4	34.6
Ours	N	N	N	**48.1**	**68.5**	37.8	**58.3**	**55.2**	48.7	**47.1**	**32.5**	**36.9**
Same-data Qwen3-VL-4B (Bai et al., 2025a)												
SFT	N	N	N	52.5	68.4	41.6	**73.1**	54.6	48.0	45.5	**32.0**	**56.8**
VG-LLM (Zheng et al., 2025)	N	Y	Y	53.8	68.2	43.3	71.9	60.1	55.4	46.7	**32.0**	52.4
Ours	N	N	N	**54.2**	**68.7**	**44.1**	72.5	**60.8**	**56.3**	**47.4**	**32.0**	51.6
Same-data Qwen3.5-4B (Qwen Team, 2026)												
SFT	N	N	N	59.7	70.7	40.4	66.5	65.0	63.8	65.0	35.6	**70.4**
VG-LLM (Zheng et al., 2025)	N	Y	Y	62.1	**71.0**	45.1	**69.1**	70.2	66.3	**68.8**	**36.1**	70.2
Ours	N	N	N	**62.6**	70.7	**49.4**	68.2	**71.3**	**70.0**	66.5	35.1	69.6

Notation. OC: object count; AD: absolute distance; OS: object size; RS: room size; RD: relative distance; Dir: relative direction; RP: route planning; AO: appearance order. LLaVA-OV: LLaVA-OneVision; I: Instruct. Y/N: yes/no; NR: not reported; N/A: not applicable; †: construction-only use. Geometry flags exclude ordinary 2D encoders, backbone pretraining, and upstream geometry in reused RGB/QA datasets. SpatioLM uses a training teacher and an internal spatial module at inference (Inf.=N).

Evaluation settings. Human results use the tiny split. Spatial-MLLM uses macro averaging; VST is the SFT model. GeoThinker uses the scaled Qwen3-VL-8B, 32-frame setting; ConsiSpace uses the final UC-SSRL model; SpatioLM uses SenseNovaSI-8B Understanding.