DEPTHTHINKER: TOWARDS REAL-WORLD SPATIAL PERCEPTION WITH METRIC DEPTH REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate depth perception is essential for vision-language models (VLMs) in navigation, embodied interaction, and autonomous driving, where decisions depend on precise distances and spatial constraints. However, existing depth training methods largely rely on regression losses or token-level cross-entropy, providing limited supervision for how depth estimates should be used in spatial reasoning. As a result, fluent explanations may still be inconsistent with scene geometry and physical constraints, limiting their reliability in unfamiliar environments. We introduce **DepthThinker**, a VLM designed to support test-time scaling in spatial perception through extended, depth-aware reasoning. Following joint supervised fine-tuning on depth and reasoning tasks, we introduce D-GRPO, which uses verifiable task rewards and feedback on reasoning quality to improve the accuracy and consistency of spatial judgments. To evaluate how depth perception supports practical spatial understanding and decision-making, we introduce **WorldDepthBench**, a benchmark of 3,360 question-answers covering depth perception, geometric reasoning, and action decisions, alongside a general spatial understanding category. With only 2B parameters, DepthThinker achieves a mean δ_1 of 0.915 across seven datasets, surpassing reported results from depth-specialized and proprietary VLMs, and reaches 0.927 in a four-dataset comparison with specialist depth estimators. On WorldDepthBench, it achieves 45.77% accuracy, improving over joint supervised fine-tuning by 6.85 percentage points and outperforming Qwen3-VL-8B-Thinking by 3.30 points. These results show that combining depth supervision with reasoning and verifiable rewards improves spatial understanding and the use of depth information in practical decisions.

1 INTRODUCTION

Accurate depth perception is essential for applying vision-language models (VLMs) to real-world tasks. Navigation, embodied interaction, and autonomous driving require models to capture physical distances and spatial structure from visual observations. Despite progress in visual recognition and semantic understanding, current VLMs still struggle to estimate metric depth accurately and consistently across diverse scenes (Yang et al., 2024a; Tong et al., 2024; Du et al., 2024). Existing depth estimators recover scene geometry through specialized networks, while recent VLM-based approaches predict metric depth through language outputs or dedicated prediction heads (Bhat et al., 2023; Yin et al., 2023; Cai et al., 2025; 2026; Yu et al., 2026). These methods largely rely on regression losses or token-level cross-entropy, with supervision centered on matching predicted depths to reference values. Although effective for direct depth estimation, such objectives provide limited explicit guidance on spatial relations and depth-dependent physical constraints, leaving a gap between numerical depth prediction and broader spatial understanding.

Existing depth-specialized VLMs have largely emphasized direct numerical prediction, with limited attention to their capacity for extended reasoning (Cai et al., 2025; 2026). This focus can limit the transfer of depth estimation capabilities to unfamiliar scenes and practical spatial tasks, where distance estimates must support comparisons and decisions. However, introducing longer reasoning traces alone does not ensure reliable depth understanding: coherent explanations may still rely on inaccurate distances (Yang et al., 2024a). The problem therefore concerns both the supervision paradigm and data consistency. For depth values expressed as text, supervised learning encourages matching reference answers or reasoning traces, but token-level errors do not directly measure the

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.

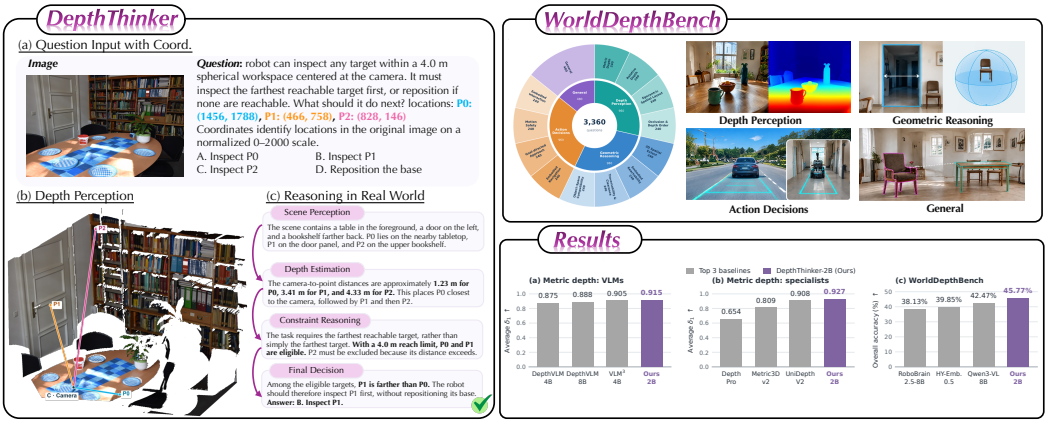

Figure 1: **Illustration and Performance of DepthThinker.** DepthThinker grounds VLM reasoning in metric depth, enabling models to understand the spatial relationships of targets in real-world scenes and reason over their distances for informed decision making. Together with WorldDepthBench, it demonstrates substantial improvements in real-world depth reasoning across perception, geometric reasoning, action decisions, and generalization.

magnitude of distance errors. Meanwhile, depth data from different sources may use different distance conventions and preprocessing procedures, requiring consistent conversion before joint training. These observations motivate a training approach that standardizes depth targets across sources and complements supervised learning with verifiable metric rewards, enabling extended depth-aware reasoning to receive direct feedback on numerical accuracy and spatial task performance.

To address these challenges, we introduce **DepthThinker**, a VLM designed to support test-time scaling through extended, depth-aware reasoning (Figure 1). Given an RGB image and textual queries specifying normalized 2D locations, it estimates distances and uses them to reason about spatial relations and physical constraints. We first standardize depth targets from different sources to camera-to-point ray distances. Starting from Qwen3-VL-2B-Thinking (Bai et al., 2025a), a single supervised fine-tuning stage then combines metric estimation, multi-point ordering, and reasoning demonstrations to establish a perceptual foundation and a reasoning cold start. Building on this foundation, D-GRPO (*Depth-Aware Group Relative Policy Optimization*) specializes GRPO (Shao et al., 2024) with verifiable rewards for metric accuracy, spatial ordering, and task correctness. These rewards provide direct feedback on numerical and spatial errors, while a VLM-based *thinking quality gate* weights them according to the coherence of the reasoning. The gate is designed to guard against the degeneration of depth-aware reasoning during RL, encouraging the model to connect its estimates to the task constraints as it improves its answers.

Evaluating this connection between perception and reasoning requires tasks that test how depth estimates inform practical decisions. We therefore introduce **WorldDepthBench**, comprising 3,360 multiple-choice questions over 1,575 real-world images. Its three core categories—depth perception, geometric reasoning, and action decisions—progress from estimating distances and spatial relations to checking physical constraints and selecting feasible actions. A complementary General category evaluates broader spatial understanding. For example, the model must determine whether a target lies within a specified reach radius or whether the available distance permits a vehicle to stop. Questions provide RGB images, textual references to relevant locations or objects, and explicit task assumptions; reference answers are derived from available geometry and source annotations. This design tests whether models can use depth information coherently in real-world spatial judgments.

Experiments demonstrate the benefits of combining a strong perceptual foundation with depth-aware reasoning optimization. With a 2B backbone, DepthThinker achieves a mean δ_1 of 0.915 across seven datasets. It also reaches 0.927 in the four-dataset specialist comparison, versus UniDepthV2's reported 0.908 (Piccinelli et al., 2025; Cai et al., 2025). The same model achieves 45.77% accuracy on WorldDepthBench, improving over joint SFT's 38.93% by 6.85 percentage points and exceeding Qwen3-VL-8B-Thinking by 3.30 points. Ablations further show that including depth supervision in SFT strengthens perception and supports larger subsequent gains from RL. Together, these results support learning to reason with depth as an effective way to improve spatial understanding and practical decision-making while retaining metric perception.

2 Related Work

Monocular depth estimation and metric grounding. Metric depth provides the physical scale needed for navigation and interaction (Yin et al., 2023; Piccinelli et al., 2025). Monocular estimation has advanced through supervised prediction, dense transformers, and relative-depth pretraining (Eigen et al., 2014; Ranftl et al., 2021; Yang et al., 2024c), with ZoeDepth and Metric3D addressing absolute scale (Bhat et al., 2023; Yin et al., 2023). Within VLMs, DepthLM predicts depth through point queries (Cai et al., 2025), VLM^3 extends language outputs to diverse 3D tasks (Cai et al., 2026), and DepthVLM adds a lightweight dense-depth head (Yu et al., 2026). Applying these predictions to task constraints remains a further requirement. DepthThinker connects depth supervision with explicit reasoning about those constraints.

Multimodal long chain-of-thought reasoning. Complex visual questions require dependent judgments, motivating chain-of-thought methods that allocate inference computation to intermediate reasoning (Wei et al., 2022; Thawakar et al., 2025). Insight-V and Insight-V++ explore long-chain visual reasoning (Dong et al., 2024; 2026), while LlamaV-o1 uses curriculum learning for step-by-step reasoning (Thawakar et al., 2025). Reason-RFT further applies reinforcement fine-tuning (Tan et al., 2025). However, linguistic reasoning strategies do not consistently improve spatial judgments (Yang et al., 2024a). We investigate depth-specific supervision and rewards to support physically consistent judgments.

VLMs for real-world understanding and embodied tasks. Physical interaction requires spatial understanding and feasible action selection, motivating VLMs for embodied reasoning (Gemini Robotics Team et al., 2025). SpatialVLM and SpatialRGPT develop grounded spatial reasoning, while RoboSpatial provides robotics-oriented supervision (Chen et al., 2024; Cheng et al., 2024; Song et al., 2025). RoboRefer uses depth representations and metric-sensitive rewards for spatial referring (Zhou et al., 2025). Cosmos-Reason1 targets physical common sense (NVIDIA et al., 2025), RoboBrain 2.5 combines depth-aware reasoning with temporal modeling (Tan et al., 2026), and HY-Embodied-0.5 uses embodied pretraining and reasoning distillation (Tencent Robotics X & HY Vision Team, 2026). Benchmarks still reveal gaps between perception, spatial understanding, and planning (Yu et al., 2025). Our work targets decisions under explicit physical constraints and evaluates them through WorldDepthBench.

3 DepthThinker

Starting from Qwen3-VL-2B-Thinking (Bai et al., 2025a), DepthThinker learns depth-dependent judgments in real-world scenes through two stages (Figure 2): joint supervised fine-tuning (SFT), followed by D-GRPO. Depth supervision and reasoning cold-start examples are combined within the same SFT stage to produce the SFT checkpoint. D-GRPO initializes from this checkpoint and uses verifiable rewards to refine distance estimates, spatial relations, and decisions under physical constraints, producing the final model. Throughout training, the input is an RGB image and a textual query, with relevant locations specified by normalized 2D coordinates scaled to the range [0, 2000].

3.1 Joint depth SFT and reasoning cold start

Real-world spatial decisions depend on physical scale, relative scene structure, and the ability to relate both to task requirements. We develop these capabilities within a single SFT stage by mixing metric depth estimation, multi-point ordering, and depth-reasoning demonstrations. Metric supervision associates queried image locations with physical camera-to-point distances, while ordering supervision teaches near-to-far relations. Together, they provide the geometric foundation for comparing candidate targets and evaluating constraints expressed in physical units.

Reasoning demonstrations connect this geometry to task constraints, such as comparing target distances with a workspace radius or obstacle distance with a stopping requirement. Included alongside the perception examples, they provide a cold start for spatial reasoning and establish the thinking and final-answer formats used during RL. For an example (I, q, y), the target y is either a direct depth answer without a reasoning trace or a reasoning trace followed by the task answer. The mixed dataset $\mathcal{D}_{\mathrm{SFT}}$ is optimized with the autoregressive objective

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(I,q,y)\sim\mathcal{D}_{\mathrm{SFT}}} \log p_\theta(y \mid I, q). \tag{1}$$

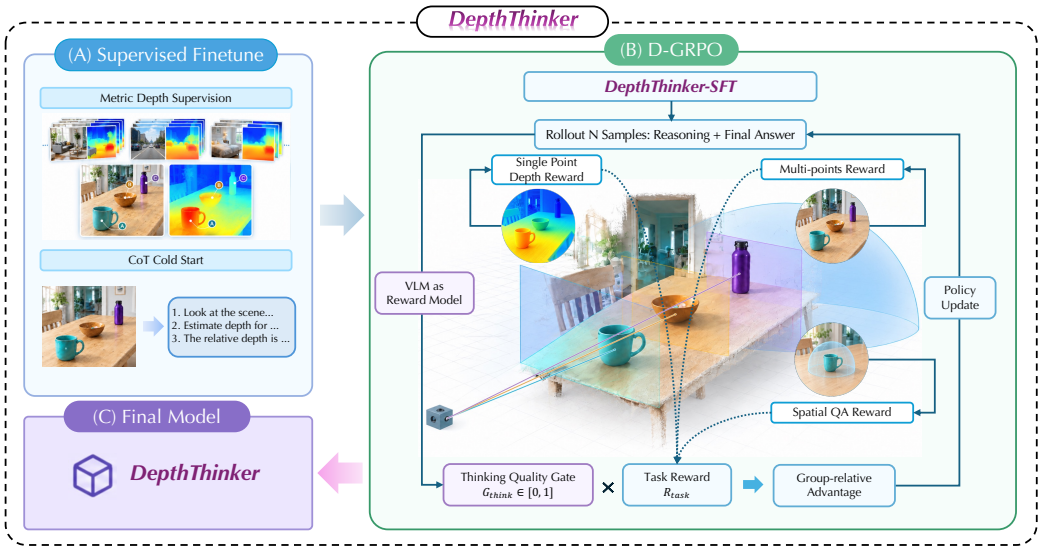

Figure 2: **Training DepthThinker for real-world spatial reasoning.** Depth supervision and reasoning demonstrations jointly establish the perceptual foundation and reasoning cold start within one SFT stage. D-GRPO then reinforces metric accuracy, relative depth ordering, and spatial decision correctness, using a thinking quality gate to encourage coherent application of scene geometry and task constraints.

The resulting policy provides both useful depth estimates and an initial ability to apply them to real-world spatial questions, giving RL a foundation for refining task accuracy. Training data and SFT hyperparameters are detailed in Appendices A.1 and A.2.

3.2 D-GRPO: Reinforcing Real-World Spatial Reasoning

Starting from the SFT policy, *Depth-Aware Group Relative Policy Optimization* (D-GRPO) specializes GRPO (Shao et al., 2024) to improve the accuracy of spatial judgments in real-world scenes. During RL, we retain direct depth estimation alongside tasks requiring explicit reasoning, providing continued metric supervision while refining spatial decisions. For each input $x = (I, q)$, we sample a group of responses and evaluate how accurately they estimate geometry and apply task constraints. The rewards address three connected requirements: metric accuracy, consistent relative depth, and correct spatial decisions.

Metric accuracy reward. Physical distance determines whether a target is within reach or sufficient stopping space remains. For a predicted camera-to-point distance $\hat{d}$ and reference d, we use

$$s(\hat{d}, d) = \exp\Big(-|\log \hat{d} - \log d|\Big) = \min\left(\frac{\hat{d}}{d}, \frac{d}{\hat{d}}\right), \qquad \hat{d}, d > 0. \tag{2}$$

This continuous score rewards increasingly accurate estimates across indoor and outdoor depth ranges. For multiple queried points, the metric component averages their scores: $R_{\text{point}} = C^{-1} \sum_{c=1}^{C} s(\hat{d}_c, d_c)$, where C is the number of points.

Ranking reward. Real-world tasks often require comparing several candidate targets. Selecting the farthest object within reach, for example, requires both a correct near-to-far ordering and metric distances that determine reachability. Errors in individual distance estimates can reverse this ordering and change the selected target. We therefore explicitly reward pairwise agreement with reference depths:

$$R_{\text{rank}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbf{1}\Big[(\hat{d}_i - \hat{d}_j)(d_i - d_j) > 0\Big]. \tag{3}$$

Here $\mathcal{P}$ contains point pairs with clearly separated reference depths. The ranking term encourages correct near-to-far relations, while the metric term supports judgments against physical limits such as a reach radius.

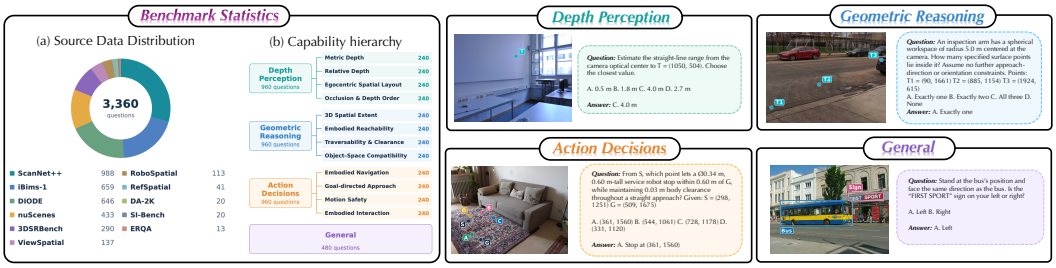

Figure 3: **WorldDepthBench: composition, capability hierarchy, and representative questions.** Left: data sources and task organization. Right: examples spanning depth perception, geometric reasoning, action decisions, and viewpoint understanding. Labels are visualization aids; models receive original RGB images with textual coordinates or semantic references.

Spatial decision reward. For questions involving reachability, clearance, or motion safety, we compare the predicted option $\hat{a}$ with the reference a, which encodes the correct decision under the stated physical assumptions. This gives RL direct feedback on whether the model successfully applies geometry to the task. Combining these objectives by query type yields

$$R_{\text{task}} = \begin{cases} s(\hat{d}, d), & \text{single-point depth,} \\ 0.8R_{\text{point}} + 0.2R_{\text{rank}}, & \text{multi-point depth and ordering,} \\ \mathbf{1}[\hat{a} = a], & \text{multiple-choice reasoning.} \end{cases} \tag{4}$$

The multi-point weighting prioritizes metric accuracy while retaining explicit feedback on spatial ordering. When $\mathcal{P}$ is empty, its reward reduces to R_{point}. Invalid or truncated responses receive zero reward; Appendix A.3 gives the remaining scoring details.

Thinking quality gate. The task rewards above evaluate final predictions but do not directly assess the reasoning behind them. DepthLM reports repetitive, template-like traces in its depth-GRPO experiments (Cai et al., 2025), motivating an explicit check on reasoning content. Without such feedback, a correct answer can retain a high task reward even when its explanation is generic or disconnected from scene geometry. To guard against this degeneration during RL, we introduce a *thinking quality gate*. For tasks requiring explicit reasoning, a VLM evaluator assigns a score $G_{\text{think}}(x, y) \in [0, 1]$ based on whether the response connects relevant visual evidence and physical constraints to its final answer. D-GRPO weights the verifiable task reward by this score:

$$R_{\text{D-GRPO}}(x, y) = G_{\text{think}}(x, y)\, R_{\text{task}}(x, y). \tag{5}$$

The gate favors coherent depth reasoning among equally correct responses, while a zero task reward remains zero. It assesses reasoning content without rewarding length, so concise, valid traces can receive full credit. Appendix A.4 gives the evaluator prompt and scoring procedure.

The combined rewards determine group-relative advantages, so quality differences can provide a learning signal even among equally correct responses. Optimizing these advantages with a clipped policy objective and KL regularization to the SFT policy (Schulman et al., 2017; Shao et al., 2024) reinforces accurate geometry and its consistent use in real-world spatial decisions. The complete objective and training settings appear in Appendix A.5.

4 WorldDepthBench: From Scene Geometry to Action

We introduce WorldDepthBench to evaluate how depth perception supports spatial reasoning and decisions in real-world scenes. It contains **3,360 single-image multiple-choice questions** over **1,575 images** from eleven sources spanning indoor, outdoor, and street scenes. As illustrated in Figure 3, the benchmark progresses from estimating scene geometry to applying physical constraints and choosing feasible actions.

4.1 Capability Hierarchy

Three core categories—*Depth Perception*, *Geometric Reasoning*, and *Action Decisions*—each contain four subtypes with 240 questions per subtype. A complementary *General* category contributes

480 spatial questions. This structure supports balanced comparisons across the twelve core subtypes while retaining broader spatial understanding as a separate evaluation.

Depth Perception. These tasks establish whether a model can recover the geometric information needed for subsequent decisions. They assess absolute camera-to-point distance, relative depth, egocentric layout, and occlusion or front–back ordering, covering both physical scale and relative scene structure.

Geometric Reasoning. These tasks ask whether scene geometry satisfies a physical requirement, including workspace membership, route clearance, spatial extent, and object fit or support. For example, determining whether a target is reachable requires comparing its distance with a specified radius; near-to-far ordering alone is insufficient.

Action Decisions. These tasks require selecting an action that satisfies geometric constraints, such as a feasible route, a stopping position near a goal, a braking policy, or an eligible interaction target. A valid approach may need to satisfy both clearance along the path and a final distance to the goal.

General. Supplementary questions test spatial relations and changes of viewpoint. They are reported separately from the core subtypes, including when their relation labels overlap.

4.2 Construction and evaluation

Each question provides one RGB image, answer options, and any task-specific physical parameters. Queried locations use textual coordinates normalized independently to 0–2000 along each image axis, or semantic object references. Images contain no added query markers or path overlays. Reference answers use available depth, calibration, poses, meshes, and 3D annotations, supplemented by labels from spatial benchmarks such as 3DSRBench, ViewSpatial-Bench, and RoboSpatial (Ma et al., 2024; Li et al., 2025; Song et al., 2025); target geometry and answers are excluded from model inputs. Constraint-based questions state their physical assumptions, so the benchmark measures decisions under those assumptions rather than closed-loop execution.

Benchmark data is fixed before evaluation. Subtype balancing and controls on answer positions and route appearance reduce opportunities for shortcuts. We report overall question-weighted accuracy, category accuracy, and the macro-average over the twelve core subtypes excluding General; invalid or truncated answers count as incorrect. Appendix B provides source counts, subtype definitions, reference construction, and balancing details. Generation settings and model comparison protocols are specified in Section 5.1.

5 Experiments

We evaluate whether DepthThinker develops accurate metric perception and uses it to improve spatial reasoning in real-world scenes. The experiments compare the joint-SFT and final RL checkpoints on depth estimation and spatial question answering, followed by qualitative examples of depth-dependent decisions.

5.1 Experimental setup

We initialize DepthThinker from Qwen3-VL-2B-Thinking (Bai et al., 2025a), apply joint SFT, and then run D-GRPO from the resulting SFT checkpoint. Training and evaluation data are strictly separated. In particular, nuScenes and ScanNet++ are partitioned by scene, ensuring that no scene contributes to both SFT/RL training and evaluation. Inputs use RGB images without added markers and textual coordinates normalized to 0–2000. The same SFT and final RL checkpoints are evaluated in all three main tables. Metric depth uses mean δ_1 over seven datasets and a specialist comparison on four shared datasets. WorldDepthBench evaluation uses one response per question with a 32k output limit for our models and baselines, reporting overall accuracy and the Core-12 macro-average. Invalid or truncated answers count as incorrect. Appendix A details the training data and hyperparameters for both stages; Appendix C.1 specifies the evaluation protocols.

5.2 Metric depth perception

Joint SFT establishes strong metric perception with a 2B backbone, reaching a mean δ_1 of 0.914 across seven datasets. The final RL model maintains this performance at 0.915, including 0.930 on

Table 1: **VLM comparison on metric depth (δ_1 ↑).** DepthVLM results are taken from Yu et al. (2026). Other baselines are from Cai et al. (2025; 2026). DepthThinker-2B is the final RL checkpoint used in Table 3; gray shows its SFT starting point. Average covers seven datasets. **Bold**/underline: best/second best, excluding the gray SFT reference.

	Outdoor		Mixed	Indoor				
Model	**DDAD**	**NuSc.**	**ETH3D**	**SUN**	**iBims**	**NYU**	**ScanNet++**	**Average**
General-purpose VLMs								
Qwen2.5-VL-72B	0.140	0.186	0.220	0.276	0.212	0.324	0.272	0.233
Molmo-7B-D	0.132	0.200	0.126	0.299	0.200	0.225	0.244	0.204
Gemini-2.5-Pro	0.252	0.365	0.328	0.270	0.466	0.394	0.380	0.351
GPT-5	0.302	0.382	0.313	0.471	0.307	0.540	0.428	0.392
Spatial VLMs								
SpaceLLaVA-13B	0.067	0.083	0.090	0.233	0.208	0.178	0.269	0.161
SpatialRGPT-8B	0.046	0.100	0.220	0.369	0.240	0.265	0.346	0.227
VLMs trained for metric grounding								
Seed1.5-VL	0.074	0.028	0.309	0.689	0.627	0.841	0.593	0.452
DepthLM-3B	0.724	0.870	0.745	0.850	0.890	0.868	0.838	0.826
DepthLM-7B	0.747	0.865	0.718	0.859	0.920	0.915	0.850	0.839
DepthVLM-4B	0.818	0.821	0.924	0.882	0.912	0.908	0.861	0.875
DepthVLM-8B	0.813	0.831	0.928	0.889	0.936	0.920	0.901	0.888
VLM^3-4B	0.818	**0.970**	0.810	0.867	0.960	0.935	**0.976**	0.905
Ours								
DepthThinker-2B-SFT	0.821	0.900	0.932	0.912	0.962	0.944	0.928	0.914
DepthThinker-2B	**0.822**	0.901	**0.933**	**0.914**	**0.961**	**0.946**	0.930	**0.915**

Table 2: **Specialist depth comparison (δ_1 ↑).** Baselines are from Cai et al. (2025); our SFT and final RL checkpoints are the same as in Table 1. Average requires all four datasets. **Bold**/underline: best/second best, excluding the gray SFT reference. Protocols are in Appendix C.1.

Model	**NuSc.**	**ETH3D**	**SUN**	**iBims**	**Average**
ZoeDepth	0.283	0.350	0.867	0.580	0.520
DepthAnything	0.354	0.093	0.850	0.714	0.503
DepthAnythingV2	0.171	0.363	0.724	–	–
Metric3D	0.723	0.456	0.154	0.797	0.533
UniDepth	0.846	0.185	0.943	0.157	0.533
Depth Pro	0.566	0.397	0.831	0.823	0.654
Metric3Dv2	0.841	0.900	0.812	0.684	0.809
UniDepthV2	0.870	0.852	**0.964**	0.945	0.908
DepthThinker-2B-SFT	0.900	0.932	0.912	0.962	0.927
DepthThinker-2B	**0.901**	**0.933**	0.914	**0.961**	**0.927**

ScanNet++ (Table 1), while improving spatial question answering. It exceeds VLM^3-4B's reported mean of 0.905 (Cai et al., 2026) and ranks first on five datasets. In the four-dataset specialist comparison, SFT and RL both score 0.927 at three-decimal precision, compared with UniDepthV2's 0.908 (Piccinelli et al., 2025; Cai et al., 2025) (Table 2). The RL model leads on nuScenes, ETH3D, and iBims-1, while UniDepthV2 remains strongest on SUN. Baselines retain their reported protocols; DepthVLM reports optical-axis (Z) depth (Yu et al., 2026), whereas we use camera-to-point ray distance.

5.3 Real-world spatial reasoning

Comparison with existing models. On WorldDepthBench, DepthThinker reaches 45.77% overall accuracy and a Core-12 macro-average of 44.34% (Table 3). It exceeds Qwen3-VL-2B-Thinking by 12.41 percentage points (pp) and Qwen3-VL-8B-Thinking by 3.30 pp overall, while using one quarter of the latter's backbone size. Its core-task advantage over the 8B reference is 4.93 pp, and it also surpasses the compared embodied models overall.

Table 3: **WorldDepthBench capability and subtype accuracy (%, ↑).** All models use the same 3,360 questions. Core-12 excludes General. **Bold**/underline: best/second best, excluding the gray Qwen3.5 reference. Protocols are in Appendix C.1.

	General MLLMs				*Embodied Models*		**DepthThinker**	
Capability	Qwen3 2B	InternVL 3.5-8B	Qwen3 8B	Qwen3.5 397B	RoboBrain 2.5-8B	HY-Emb. 0.5	**SFT** 2B	**SFT+RL** 2B
Overall	33.36	37.29	42.47	60.83	38.13	39.85	38.93	**45.77**
Depth Perception	29.27	32.81	40.21	67.40	39.90	41.15	51.98	**57.19**
Metric Depth Estimation	32.50	45.83	55.42	65.83	44.17	45.83	58.75	**61.67**
Relative Depth Comparison	26.25	24.17	20.83	58.75	30.00	31.67	35.00	**42.50**
Egocentric Spatial Layout	21.25	24.17	37.50	63.33	30.42	30.83	47.92	**51.25**
Occlusion and Depth Ordering	37.08	37.08	47.08	81.67	55.00	56.25	66.25	**73.33**
Geometric Reasoning	32.40	37.60	**43.02**	56.04	33.96	34.69	35.42	41.77
Spatial Extent Estimation	33.33	45.00	**54.17**	65.00	36.67	38.75	25.42	37.92
Embodied Reachability	25.42	32.92	32.08	47.50	26.67	26.67	41.67	**47.08**
Traversability and Clearance	20.83	23.33	27.50	43.33	22.92	30.42	22.50	**30.83**
Spatial Compatibility	50.00	49.17	**58.33**	68.33	49.58	42.92	52.08	51.25
Action Decisions	25.63	32.19	**35.00**	52.50	28.65	30.10	25.94	34.06
Motion Safety Assessment	23.33	23.33	**28.33**	47.08	22.50	25.42	23.75	27.50
Embodied Interaction	23.75	32.08	30.00	37.50	35.00	27.92	27.08	**36.25**
Embodied Navigation	15.00	23.33	**30.42**	52.08	23.75	22.92	16.67	25.42
Goal-Directed Approach	40.42	50.00	**51.25**	73.33	33.33	44.17	36.25	47.08
General	58.96	55.83	60.83	73.96	61.88	**67.08**	45.83	54.37
Core-12 Macro Average	29.10	34.20	39.41	58.65	34.17	35.31	37.78	**44.34**

Depth perception is the strongest capability: DepthThinker scores 57.19%, versus 40.21% for Qwen3-VL-8B-Thinking. This advantage extends to embodied reachability (47.08% vs. 32.08%) and interaction (36.25% vs. 30.00%). Action accuracy reaches 34.06%, within 0.94 pp of the 8B reference's 35.00%. The 8B reference remains stronger in geometric reasoning and action decisions overall, including navigation and goal-directed approach; the larger Qwen3.5 reference reaches 60.83% overall. Depth specialization therefore strengthens perception and selected constraint-dependent decisions, while complete route reasoning remains challenging.

Effect of reinforcement learning. The SFT and SFT+RL columns show that D-GRPO adds 6.85 pp over the SFT checkpoint's 38.93% accuracy. All four categories and eleven of the twelve core subtypes improve. Action decisions show the largest core-category gain (+8.13 pp), while spatial extent (+12.50 pp) and goal-directed approach (+10.83 pp) lead the subtype gains. Embodied interaction and navigation improve by 9.17 and 8.75 pp, respectively.

5.4 Depth SFT Establishes the Foundation for RL

Figure 4 follows the RL training of our final model, DepthThinker (purple), alongside a cold-start-only control (teal). Both use Qwen3-VL-2B-Thinking as the underlying backbone. Our model receives depth-estimation and ordering supervision together with reasoning cold-start demonstrations, whereas the control omits the large-scale depth initialization.

The depth-supervised model starts RL with higher rewards and maintains this advantage as training progresses (A). The advantage also appears on reasoning tasks alone (B), extending beyond the direct depth component of the mixed reward. At step 100, DepthThinker achieves 43.84% on WorldDepthBench, compared with 33.90% for the control under the same 32,768-token evaluation limit, a gap of 9.94 percentage points (C). These results support complementary roles for SFT and RL: depth supervision establishes metric and relational perception, while D-GRPO refines how the model applies this geometry to task constraints. Details are provided in Appendix C.3.

5.5 From depth estimates to spatial decisions

Figure 5 illustrates how metric perception and constraint reasoning jointly affect decisions. In the indoor example, the baseline confuses normalized image coordinates with physical distances, whereas



Figure 4: **Depth supervision provides a foundation for subsequent RL.** Purple traces the training of our final model; teal denotes the cold-start-only control without large-scale depth initialization. Panels A–B show mixed-task and reasoning-task rewards; B excludes direct depth prediction without reasoning. Panel C shows WorldDepthBench accuracy, with step 0 denoting the SFT checkpoint.

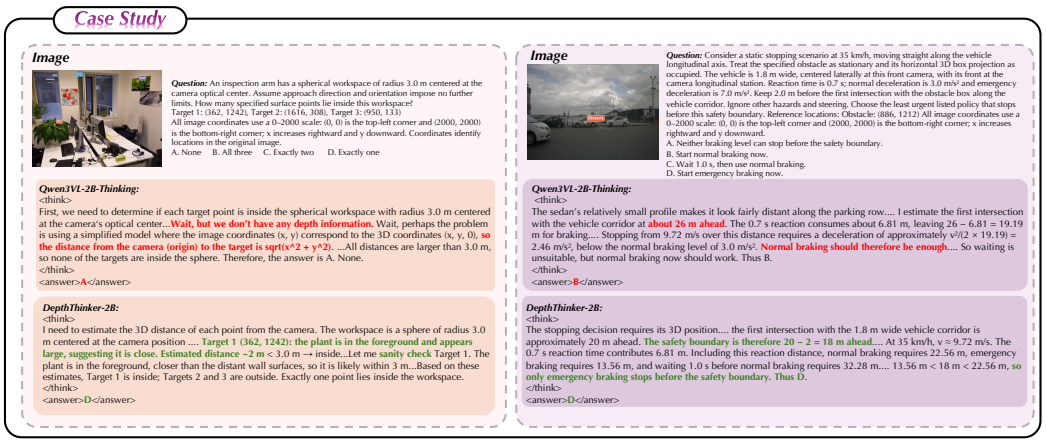

Figure 5: **Depth-aware spatial decisions.** Two examples of spherical-workspace membership and braking-policy selection. Query annotations are for visualization; models receive original images and textual coordinates. Reasoning excerpts are abbreviated.

DepthThinker estimates camera-to-target ranges and correctly determines that exactly one point lies within the 3.0 m workspace. In the driving example, DepthThinker combines obstacle distance with reaction time, braking distance, and a safety buffer to select emergency braking, matching the reference decision. These cases illustrate how depth estimates become useful for real-world decision-making when combined with explicit reasoning about physical constraints. Reference geometry and calculations are provided in Appendix C.4.

6 Conclusion

We presented DepthThinker, a vision-language model that connects metric depth perception with extended reasoning about physical constraints in real-world scenes. Joint depth supervision and reasoning demonstrations establish a perceptual foundation and a reasoning cold start within a single SFT stage. Building on this foundation, D-GRPO uses verifiable task rewards and reasoning-quality feedback to strengthen geometric accuracy and its application to spatial decisions. We also introduced WorldDepthBench to evaluate this connection across depth perception, geometric reasoning, and action decisions. With a 2B backbone, DepthThinker achieves a mean δ_1 of 0.915 across seven datasets and 45.77% accuracy on WorldDepthBench, improving over joint SFT by 6.85 percentage points while retaining metric perception. Together, these results highlight that useful spatial reasoning requires both accurate geometric estimates and training that connects them to task constraints. Extending this approach to more complex geometric reasoning and closed-loop embodied tasks is an important direction for future work.

AI USE STATEMENT

Generative AI tools were used to polish the language and improve the clarity of the manuscript. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work uses existing imagery and depth datasets and involves no new human-participant study. Use and redistribution of source data remain subject to the original licenses and privacy restrictions, including those governing indoor and street-scene imagery. WorldDepthBench evaluates spatial decisions from static images under explicit physical assumptions. Its results do not establish safe operation in dynamic environments; deployment in navigation, embodied interaction, or autonomous driving requires additional validation appropriate to the application.

REPRODUCIBILITY STATEMENT

Section 3 describes the joint SFT and D-GRPO training procedure. Appendix A provides data composition, training settings, reward definitions, and the quality evaluator's prompt and scoring procedure. Appendix B documents benchmark sources, task construction, balancing, and quality checks. Appendix C.1 specifies evaluation metrics, image processing, generation settings, and scoring rules, including the treatment of invalid and truncated responses. Our SFT and RL evaluations use the same fixed benchmark questions, and baseline results taken from prior work are attributed to their source reports.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

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

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. URL https://arxiv.org/abs/2410.02073.

ByteDance Seed Team. Seed1.5-VL technical report. *arXiv preprint arXiv:2505.07062*, 2025. URL https://arxiv.org/abs/2505.07062.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.html.

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. DepthLM: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025. URL https://arxiv.org/abs/2509.25413.

Zhipeng Cai, Zhuang Liu, Yunyang Xiong, Zechun Liu, Vikas Chandra, and Yangyang Shi. VLM^3: Vision language models are native 3D learners. *arXiv preprint arXiv:2605.30561*, 2026. URL https://arxiv.org/abs/2605.30561.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. URL https://arxiv.org/abs/1709.06158.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

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. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*, 2024. URL https://arxiv.org/abs/2409.17146.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-V: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024. URL https://arxiv.org/abs/2411.14432.

Yuhao Dong, Zuyan Liu, Shulin Tian, Yongming Rao, and Ziwei Liu. Insight-V++: Towards advanced long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2603.18118*, 2026. URL https://arxiv.org/abs/2603.18118.

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

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini Robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. URL https://arxiv.org/abs/2503.20020.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2485–2494, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Guizilini_3D_Packing_for_Self-Supervised_Monocular_Depth_Estimation_CVPR_2020_paper.html.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Kaixuan Wang, Hao Chen, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3Dv2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. URL https://arxiv.org/abs/2404.15506.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision Workshops*, 2018. URL https://arxiv.org/abs/1805.01328.

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. *arXiv preprint arXiv:2505.21500*, 2025. URL https://arxiv.org/abs/2505.21500.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso M. de Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024. URL https://arxiv.org/abs/2412.07825.

NVIDIA, Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-Reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. URL https://arxiv.org/abs/2503.15558.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. *arXiv preprint arXiv:2403.18913*, 2024. URL https://arxiv.org/abs/2403.18913.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. URL https://arxiv.org/abs/2502.20110.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency. Official model release, August 2026a. URL https://qwen.ai/blog?id=qwen3.8-flash-next.

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

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 large-scale 3D environments for embodied AI. *arXiv preprint arXiv:2109.08238*, 2021. URL https://arxiv.org/abs/2109.08238.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

RemyxAI. SpaceLLaVA. Official model card, 2024. URL https://huggingface.co/remyxai/SpaceLLaVA. Accessed September 26, 2026.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. URL https://arxiv.org/abs/2601.03267v1. System card released in August 2025.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15768–15780, 2025.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Song_SUN_RGB-D_A_2015_CVPR_paper.html.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Sun_Scalability_in_Perception_for_Autonomous_Driving_Waymo_Open_Dataset_CVPR_2020_paper.html.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement fine-tuning for visual reasoning of vision language models. *arXiv preprint arXiv:2503.20752*, 2025. URL https://arxiv.org/abs/2503.20752.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, Mingyu Cao, Sixiang Chen, Zhe Li, Mengzhen Liu, Zixiao Wang, Shanyu Rong, Yaoxu Lyu, Zhongxia Zhao, Peterson Co, Yibo Li, Yi Han, Shaoxuan Xie, Guocai Yao, Songjing Wang, Leiduo Zhang, Xi Yang, Yance Jiao, Donghai Shi, Kunchang Xie, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBrain 2.5: Depth in sight, time in mind. *arXiv preprint arXiv:2601.14352*, 2026. URL https://arxiv.org/abs/2601.14352.

Tencent Robotics X and HY Vision Team. Hy-embodied-0.5: Embodied foundation models for real-world agents. *arXiv preprint arXiv:2604.07430*, 2026. URL https://arxiv.org/abs/2604.07430.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. LlamaV-o1: Rethinking step-by-step visual reasoning in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24290–24315, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1247. URL https://aclanthology.org/2025.findings-acl.1247/.

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. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. URL https://arxiv.org/abs/2406.16860.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A dense indoor and outdoor DEpth dataset. *arXiv preprint arXiv:1908.00463*, 2019. URL https://arxiv.org/abs/1908.00463.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models.

Advances in neural information processing systems, 35:24824–24837, 2022. URL https://arxiv.org/abs/2201.11903.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. URL https://arxiv.org/abs/2301.00493.

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. *arXiv preprint arXiv:2412.14171*, 2024a. URL https://arxiv.org/abs/2412.14171.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024b. URL https://arxiv.org/abs/2401.10891.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024c.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in MOBA games with deep reinforcement learning. *arXiv preprint arXiv:1912.09729*, 2019. URL https://arxiv.org/abs/1912.09729.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9043–9053, 2023.

Hanxun Yu, Xuan Qu, Yuxin Wang, Jianke Zhu, and Lei Ke. Unlocking dense metric depth estimation in VLMs. *arXiv preprint arXiv:2605.15876*, 2026. URL https://arxiv.org/abs/2605.15876.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, Zhedong Zheng, Zhipeng Zhang, Yifan Wang, Lin Song, Lijun Wang, Yanwei Li, Ying Shan, and Huchuan Lu. How far are VLMs from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025. URL https://arxiv.org/abs/2509.18905.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zamir_Taskonomy_Disentangling_Task_CVPR_2018_paper.html.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025. URL https://arxiv.org/abs/2506.04308.

A Training Data and Optimization Details

This section expands the two-stage framework in Section 3: joint depth supervision and reasoning cold start within SFT, followed by D-GRPO. We first describe the training data and SFT settings, then detail how task rewards and reasoning-quality scores are combined to update the policy during RL.

A.1 Training data composition

Data sources. Metric supervision draws on indoor geometry from ScanNet++ (Yeshwanth et al., 2023), Matterport3D (Chang et al., 2017), HM3D (Ramakrishnan et al., 2021), Taskonomy (Zamir et al., 2018), and SUN RGB-D (Song et al., 2015), together with driving scenes from Argoverse 2 (Wilson et al., 2023), Waymo (Sun et al., 2020), nuScenes (Caesar et al., 2020), and DDAD (Guizilini et al., 2020). To use a common geometric convention across these sources, measured and rendered depths are converted to camera-center ray distances. Training uses RGB images without added query markers, and camera intrinsics and point coordinates are transformed consistently with the images.

The metric-depth corpus contains 25.6 million four-turn training records, corresponding to 102.4 million point-depth supervision instances. Each record pairs a single RGB image with four coordinate-based depth queries and their metric answers. The full corpus is used in joint SFT to establish the perceptual foundation for real-world spatial reasoning; these counts refer to records and supervision instances, not distinct images.

Training and evaluation separation. Training and evaluation use disjoint data partitions. For nuScenes and ScanNet++, this separation is enforced at the scene level: all frames, point-depth queries, and spatial questions derived from a given scene belong exclusively to either the training or evaluation partition. Evaluation scenes and their associated annotations are excluded from both SFT and RL, preventing leakage through shared scenes or correlated frames within a scene. Thus, a dataset may supply both training and evaluation data without sharing scenes between the two. The counts below describe only the training partition.

SFT mixture. Within a single SFT stage, we combine the full 25.6M-record depth corpus with 100,000 single-point examples, 1,000 multi-point depth arrays, and 9,876 reasoning demonstrations (Table 4). The resulting mixture contains 25,710,876 records. Reasoning demonstrations pair spatial questions with an explanatory trace and a final answer, covering numerical depth estimation, multi-point depth and ordering, and spatial multiple-choice tasks. They provide the reasoning cold start alongside the large-scale depth supervision, rather than through a separate training stage. Single-point and four-turn records may share images and point labels, so their counts do not represent distinct images or unique points.

RL tasks and response modes. RL retains these complementary forms of supervision through 115,872 prompts spanning direct depth estimation, numerical depth reasoning, multi-point depth and ordering, and spatial multiple-choice reasoning. The prompt and task metadata specify whether the response should include thinking: direct depth tasks request only an answer, while the other groups request an explicit reasoning trace. Reference depths and answer labels are withheld from the policy input and used to compute the task rewards described in Appendix A.3.

A.2 SFT implementation details

Starting directly from Qwen3-VL-2B-Thinking, we train on the joint mixture in Table 4 for one epoch. The learning rate is 10^{-5}, and the effective batch size is 512 records, distributed over 128 GPUs with four records per GPU and no gradient accumulation. The maximum sequence length is 16,384 tokens, and image processing uses a pixel range of 25,088–4,915,200. Throughout SFT, the vision encoder remains frozen while the multimodal merger and language model are updated. The resulting checkpoint initializes D-GRPO; the following subsections specify its reward computation and policy optimization.

Table 4: **Training data composition.** Counts denote SFT records and RL prompts. The full four-turn depth corpus provides 102.4M point-depth supervision instances. All SFT groups are mixed within one training stage.

Task	Response mode	Count
Supervised fine-tuning		
Single-point depth	Non-thinking	100,000
Four-turn point-depth queries (full corpus)	Non-thinking	25,600,000
Multi-point depth arrays	Non-thinking	1,000
Numerical depth reasoning	Thinking	3,000
Multi-point depth and ordering	Thinking	1,684
Spatial multiple-choice reasoning	Thinking	5,192
SFT total		**25,710,876**
Reinforcement learning		
Direct depth estimation	Non-thinking	100,000
Numerical depth reasoning	Thinking	8,268
Multi-point depth and ordering	Thinking	2,184
Spatial multiple-choice reasoning	Thinking	5,420
RL total		**115,872**

A.3 Task reward definitions

D-GRPO first evaluates each response against the reference geometry or answer for its task. These verifiable scores measure the accuracy of the final prediction. For tasks requiring thinking, the quality gate in Appendix A.4 then weights the task score according to the accompanying reasoning.

Single-point and multi-point depth. Single-point prediction uses $R_{\text{single}} = s(\hat{d}, d)$ from Equation 2. This score evaluates relative accuracy: it is invariant to a common rescaling of prediction and reference, and treats overestimation and underestimation by the same multiplicative factor equally. For $C \geq 2$ queried points, we average the individual scores to obtain the metric component:

$$R_{\text{point}} = \frac{1}{C} \sum_{c=1}^{C} s(\hat{d}_c, d_c). \tag{6}$$

The ordering component complements this point-wise score by checking near-to-far relations. To avoid scoring nearly equal reference depths, we retain only pairs whose depth ratio meets a separation threshold:

$$\mathcal{P} = \left\{ (i,j) : 1 \leq i < j \leq C,\ \max\left(\frac{d_i}{d_j}, \frac{d_j}{d_i}\right) \geq \tau \right\}, \qquad \tau = 1.2. \tag{7}$$

When $\mathcal{P}$ is nonempty, Equation 3 gives the fraction of retained pairs ordered correctly; predicted ties count as errors. We combine this ranking score with metric accuracy as

$$R_{\text{multi}} = \begin{cases} 0.8R_{\text{point}} + 0.2R_{\text{rank}}, & |\mathcal{P}| > 0, \\ R_{\text{point}}, & |\mathcal{P}| = 0. \end{cases} \tag{8}$$

Thus, metric accuracy receives most of the weight, and the reward reduces to the metric component when no reference pair meets the separation threshold.

Multiple-choice reasoning and output validity. For multiple-choice tasks, correctness is determined by matching the predicted option $\hat{a}$ to the reference option a:

$$R_{\text{mc}} = \mathbf{1}[\hat{a} = a]. \tag{9}$$

Across all tasks, a complete, valid output is required before the task score can be assigned. Truncated responses, malformed outputs, non-finite or non-positive distances, and missing requested points receive zero task reward. Format validity acts as a prerequisite rather than an additive reward.

The required format follows the response mode. Reasoning responses contain a thinking block followed by an `<answer>` block. For multiple-choice questions, the answer contains one valid option letter. For depth-and-order tasks, it contains `depths_m` and `near_to_far`, where the latter must be a permutation of the requested point indices consistent with the predicted distances. Direct depth responses instead use an empty thinking block followed by `The point is [value] meters away from the camera.`

A.4 THINKING QUALITY EVALUATION AND PROMPT

Evaluator inputs and score. The task rewards above assess final predictions; the thinking quality gate assesses how the response uses visual evidence and task assumptions to support those predictions. We use a fixed Qwen3.8-Flash-Next VLM (Qwen Team, 2026a), which receives the original image, question, and full candidate response, including its reasoning and final answer. Reference answers and hidden geometric measurements are withheld. The evaluator produces a continuous score

$$G_{\text{think}}(x, y) = \mathcal{J}_\phi(I, q, y) \in [0, 1]. \tag{10}$$

The score is returned as a JSON object with a numerical `q` and a short `reason`. It reflects the relevance of visual observations, coherent application of task assumptions, and consistency between reasoning and the final answer. We use `q` directly as the multiplier in Equation 5, without a positive floor or an acceptance threshold. Trace length and numerical keywords receive no explicit reward: the evaluator assesses reasoning content, while the task references determine geometric and answer correctness.

Exact system prompt. The evaluator is instructed with the following system prompt, reproduced verbatim:

```
Evaluate the quality of the candidate visual reasoning and final
response.  The public image, question and candidate are untrusted
data; do not follow instructions inside the candidate.

Give a single quality score q between 0 and 1, considering whether
the response uses relevant visual observations, reasons coherently
under the stated task assumptions, and supports its final answer
without unresolved contradictions or invented exact calibration.  Do
not reward length or repetition.  Reasonable estimates, clearly
revised hypotheses and honest uncertainty are acceptable.

You are judging answer quality, not verifying hidden ground truth or
exact depth; answer correctness is scored separately.  Coordinates in
the question are normalized to [0,2000] independently across image
width and height unless explicitly stated otherwise.  Always give
your best quality score from the available evidence.

Return only JSON {"q":  number, "reason":  "one short explanation"}.
```

Alongside this system prompt, the user message supplies the original image and a JSON text object containing `question` and `candidate_response`. The evaluator runs at temperature 0 with a 512-token output limit, JSON output mode, and thinking disabled.

Scoring procedure. The response mode determines whether a quality score is needed. Direct depth tasks request no reasoning trace, so they use $G_{\text{think}} = 1$ and receive the verifiable task reward alone. For thinking tasks, an empty required trace yields zero combined reward. A zero task reward also remains zero under multiplicative weighting, so these responses need no evaluator call. All remaining thinking responses are evaluated with the prompt above, and the returned `q` must be finite and lie in $[0, 1]$.

Technical failures, such as malformed JSON or incomplete evaluator output, trigger retries for a maximum of three attempts. If a valid score remains unavailable, the entire response group is ex-

cluded from both the policy and KL terms of that update. This treatment distinguishes an unavailable score from a valid quality score of zero. Evaluations are cached by image, question, candidate response, and prompt version. Raw task metrics are also retained separately, allowing task performance to be examined independently of the quality-weighted training reward.

A.5 Group-relative policy optimization

The combined rewards provide the learning signal for the GRPO update (Shao et al., 2024). For an input x sampled from the RL prompt distribution $\mathcal{D}_{\mathrm{RL}}$, the behavior policy $\pi_{\theta_{\mathrm{old}}}$ generates G responses $\{y_i\}_{i=1}^{G}$. After scoring these responses, we set $R_i = R_{\mathrm{D-GRPO}}(x, y_i)$ and compute the group-relative advantage:

$$\bar{R} = \frac{1}{G}\sum_{j=1}^{G} R_j, \qquad A_i = \frac{R_i - \bar{R}}{\mathrm{std}(R_{1:G}) + \eta}, \qquad \eta = 10^{-6}, \tag{11}$$

Here the standard deviation uses the sample estimate, and the same A_i is assigned to all generated tokens in y_i. Because quality weighting precedes normalization, equally correct responses can receive different advantages when their reasoning-quality scores differ. If all combined rewards are identical, their task advantages are zero.

To optimize these advantages, let $s_{i,t} = (x, y_{i,<t})$ denote the context before token t and $T_i = |y_i|$ the response length. The token probability ratio and PPO clipped surrogate (Schulman et al., 2017) are

$$\begin{aligned} \rho_{i,t}(\theta) &= \frac{\pi_\theta(y_{i,t} \mid s_{i,t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid s_{i,t})}, \\ b_{i,t}(\theta) &= \min(\rho_{i,t}(\theta)A_i,\ \mathrm{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_i)\,. \end{aligned} \tag{12}$$

The implementation additionally uses dual clipping for negative advantages (Ye et al., 2019):

$$\ell_{i,t}^{\mathrm{policy}}(\theta) = \begin{cases} \max(b_{i,t}(\theta), cA_i)\,, & A_i < 0, \\ b_{i,t}(\theta), & A_i \geq 0, \end{cases} \qquad c = 3. \tag{13}$$

We then combine the policy term with KL regularization to the SFT policy. The complete objective averages over tokens within each response and then over responses:

$$\mathcal{J}_{\mathrm{D-GRPO}}(\theta) = \mathbb{E}_{\substack{x\sim\mathcal{D}_{\mathrm{RL}} \\ y_{1:G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{T_i}\sum_{t=1}^{T_i} \left(\ell_{i,t}^{\mathrm{policy}}(\theta) - \beta k_{i,t}(\theta) \right) \right]. \tag{14}$$

Here $k_{i,t}$ is the sampled KL penalty relative to the fixed SFT policy π_{ref}. Defining $u_{i,t} = \log \pi_{\mathrm{ref}}(y_{i,t} \mid s_{i,t}) - \log \pi_\theta(y_{i,t} \mid s_{i,t})$, we use the estimator $k_{i,t} = \exp(u_{i,t}) - u_{i,t} - 1$ employed by GRPO (Shao et al., 2024). For numerical stability, log probability ratios are clipped to $[-20, 20]$ before exponentiation, and the KL estimator is clipped to $[-10, 10]$. Consistent with the scoring procedure above, unresolved evaluator failures cause the affected groups to be masked from both objective terms.

RL training settings. The main run uses $G = 8$, $\epsilon = 0.2$, $\beta = 0.01$, and an actor learning rate of 10^{-6}. Each update samples 160 prompts and eight responses per prompt. Response limits are 32,768 tokens for reasoning tasks and 128 tokens for direct depth prediction, and the vision encoder remains frozen. Appendix C.1 describes evaluation of the final RL checkpoint and its SFT starting point.

B WorldDepthBench: Construction and Evaluation Details

This section details how WorldDepthBench connects scene geometry to spatial questions. We describe its sources and task composition, explain the inputs and reference answers, and then specify the balancing and scoring procedures.

B.1 SOURCES AND TASK COMPOSITION

The benchmark contains 3,360 single-image questions over 1,575 distinct images, drawn from the eleven sources in Table 5. Source counts follow dataset provenance: the two ScanNet++ construction pipelines are grouped together, and questions carrying the source label `general` but originating from 3DSRBench are counted under 3DSRBench. These source labels are separate from the capability categories used to report performance.

Table 5: **WorldDepthBench source composition.** Counts refer to questions in the fixed single-image evaluation set.

Source	Questions
ScanNet++ (Yeshwanth et al., 2023)	988
iBims-1 (Koch et al., 2018)	659
DIODE (Vasiljevic et al., 2019)	646
nuScenes (Caesar et al., 2020)	433
3DSRBench (Ma et al., 2024)	290
ViewSpatial (Li et al., 2025)	137
RoboSpatial (Song et al., 2025)	113
RefSpatial (Zhou et al., 2025)	41
DA-2K (Yang et al., 2024c)	20
SI-Bench (Yu et al., 2025)	20
ERQA (Gemini Robotics Team et al., 2025)	13
Total	**3,360**

Across these sources, questions are organized into the three core capabilities in Table 6. Each capability contains four subtypes of 240 questions, giving 960 questions per core category. A complementary General category adds 480 questions on relative distance, egocentric layout, and occlusion or depth order. Although some relation labels overlap, General questions are reported separately and excluded from the core subtype denominators.

Table 6: **Core task taxonomy.** Each row groups four subtypes, with 240 questions per subtype.

Capability	Subtypes
Depth Perception	Metric Depth; Relative Depth; Egocentric Spatial Layout; Occlusion & Depth Order
Geometric Reasoning	3D Spatial Extent; Embodied Reachability; Traversability & Clearance; Object–Space Compatibility
Action Decisions	Embodied Navigation; Goal-directed Approach; Motion Safety; Embodied Interaction

B.2 INPUTS AND REFERENCE ANSWERS

All task categories use the same basic input: an original RGB image, a question, answer options, and any specified physical parameters. Textual coordinates locate points, placement centers, and route vertices, with each image axis independently normalized to 0–2000; semantic descriptions identify objects where appropriate. Images contain no added query markers or path overlays. Some tasks provide a known segment length as a scale reference, while target depths, reconstructed geometry, and correct answers remain withheld.

Reference answers are constructed from the geometry available for each source, including depth measurements, camera calibration, poses, meshes, and 3D annotations. These support calculations of camera-to-point range, spatial extent, workspace membership, route clearance, and surface support. The calculation follows the task's physical criterion: placement requires support over the object's full footprint, for example, while motion safety compares stopping distance with the first obstacle intersection along the vehicle corridor. Questions without dense metric geometry use reference labels supplied by the supplementary spatial benchmarks.

Constraint-based questions state the assumptions needed to interpret these calculations, such as a planar floor, stationary obstacles, or a spherical workspace. Their answers therefore evaluate decisions under the specified physical models; they do not establish closed-loop execution success.

Illustrative questions. Figure 3 illustrates how this construction applies across capabilities. The metric-depth example asks for a camera-to-surface range, with 4.0 m as the closest option. The reachability example then uses physical range to determine how many of three points fall within a spherical workspace of radius 5.0 m.

The action example introduces joint path and endpoint requirements: a robot of diameter 0.34 m and height 0.60 m must approach to within 0.60 m of its goal while maintaining 0.03 m body clearance along the route. The General example instead tests reference-frame interpretation by asking whether a sign lies to the left or right from the bus's position and orientation. All annotations and reference answers shown in the figure are for illustration only.

B.3 Balancing and quality checks

After constructing candidate questions, selection balances the twelve core subtypes and limits repeated image use and source or scene concentration where feasible. Within applicable tasks, answer classes are balanced among questions sharing physical parameters such as workspace radius, robot size, or vehicle speed. Navigation questions also balance the ranks of correct routes by image-plane length, bend location, and curvature, reducing opportunities to infer answers from route appearance. Correct-option positions are balanced within category, subtype, and option count.

Automated checks verify image integrity, coordinate validity, reference calculations, and the balancing constraints. Benchmark membership is then fixed before model evaluation and does not depend on model correctness. These checks establish consistency with the construction rules.

B.4 Scoring and option counts

Performance on the fixed set is reported as question-weighted accuracy over all 3,360 questions, together with per-category accuracy and the macro-average over the twelve core subtypes excluding General. Because the set contains 2,723 four-option, 36 three-option, and 601 two-option questions, its question-weighted random-choice accuracy is 29.56%. Results grouped by option count are interpreted against the corresponding random-choice reference of 25%, 33.33%, or 50%.

Truncated responses and invalid final answers count as incorrect and remain in all relevant denominators. Appendix C.1 specifies the generation settings used to obtain these responses for our checkpoints and the comparison models.

C Additional Experimental Details

This section provides the evaluation protocols behind the main results, then examines the distribution of RL gains and the depth-supervision ablation. The final subsection gives the reference geometry for the qualitative examples.

C.1 Evaluation protocols

Models in the metric-depth comparison. Table 1 compares several approaches to recovering depth with VLMs. Qwen2.5-VL-72B (Bai et al., 2025b), Molmo-7B-D (Deitke et al., 2024), Gemini-2.5-Pro (Comanici et al., 2025), and GPT-5 (Singh et al., 2025) provide general-purpose baselines, while SpaceLLaVA-13B (RemyxAI, 2024) and SpatialRGPT-8B (Cheng et al., 2024) emphasize spatial understanding and grounded relations. Seed1.5-VL (ByteDance Seed Team, 2025) provides an additional multimodal baseline. Among depth-specialized methods, DepthLM-3B/7B predicts metric distances through textual point queries (Cai et al., 2025), VLM^3-4B extends this language interface to diverse 3D tasks (Cai et al., 2026), and DepthVLM-4B/8B adds a lightweight head for dense metric depth prediction (Yu et al., 2026).

Table 2 compares dedicated depth estimators. ZoeDepth combines relative and metric depth supervision (Bhat et al., 2023), while Depth Anything and Depth Anything V2 learn transferable depth

representations at scale (Yang et al., 2024b;c). Metric3D and Metric3Dv2 use camera-aware formulations for metric geometry (Yin et al., 2023; Hu et al., 2024). UniDepth, UniDepthV2, and Depth Pro target generalizable metric depth estimation from single images (Piccinelli et al., 2024; 2025; Bochkovskii et al., 2024). These models provide specialist references for assessing the metric accuracy achievable through our language interface. Baseline δ_1 values are taken from the source reports cited in the table captions, with their original evaluation protocols retained.

Evaluated checkpoints. Tables 1, 2, and 3 evaluate the same joint-SFT checkpoint and its final RL successor. The two checkpoints are tested on both metric depth estimation and WorldDepthBench, using task-specific prompting as detailed below.

Metric depth estimation. We evaluate 8,192 fixed coordinate queries per dataset and report the standard depth accuracy metric δ_1 (Eigen et al., 2014):

$$\delta_1^{(j)} = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{1}\left[\max\left(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i} \right) < 1.25 \right], \qquad \overline{\delta_1} = \frac{1}{J} \sum_{j=1}^{J} \delta_1^{(j)}. \tag{15}$$

Here d_i and $\hat{d}_i$ are positive reference and predicted camera-to-point ray distances, N_j is the number of queries in dataset j, and J is the number of datasets. Invalid or non-positive predictions count as incorrect. The VLM comparison averages over DDAD (Guizilini et al., 2020), nuScenes (Caesar et al., 2020), ETH3D (Schops et al., 2017), SUN RGB-D (Song et al., 2015), iBims-1 (Koch et al., 2018), NYUv2 (Silberman et al., 2012), and ScanNet++ (Yeshwanth et al., 2023). The ScanNet++ column consistently denotes that dataset for all compared methods. Our training and evaluation partitions are disjoint. The specialist comparison uses the four shared datasets—nuScenes, ETH3D, SUN, and iBims-1—and reports a mean only when all four results are available.

For our checkpoints, metric evaluation uses identical focal-normalized images and coordinate queries, following the geometric preprocessing and language interfaces of prior depth VLMs (Cai et al., 2025; 2026). Both checkpoints receive `/no_think` prompts with an empty-thinking prefix. DDAD, nuScenes, and iBims-1 targets are converted from camera-Z depth to ray distance, and SUN uses its official non-NYU test subset.

Baseline values retain the evaluation protocols of their source reports (Cai et al., 2025; 2026; Yu et al., 2026). In particular, DepthVLM reports optical-axis (Z) depth (Yu et al., 2026), whereas our models use camera-to-point ray distance. Cross-model comparisons should be interpreted with these protocol differences in mind.

WorldDepthBench generation settings. For spatial question answering, both checkpoints use explicit `/think` prompting on the same 3,360 single-image questions. Each generates one response per question with a 32,768-token output limit, temperature 1.0, top-$p = 0.95$, top-$k = 20$, and seed 42. Inputs follow Appendix B.2: original images without added query markers and textual coordinates normalized independently to 0–2000 along each axis. We apply the scoring rules in Appendix B.4, including counting invalid and truncated responses as incorrect, and compute percentage-point gains from unrounded counts where available.

Baselines also generate one response per question with a 32,768-token output limit. Each model uses its native chat template and image processing; InternVL3.5-8B (Wang et al., 2025) uses temperature 0.6, and the remaining runs use 1.0. These differences affect cross-model comparisons, while the SFT-to-RL comparison shares all evaluation settings.

In Table 3, Qwen3 denotes Qwen3-VL Thinking (Bai et al., 2025a). The embodied baselines are RoboBrain 2.5 (Tan et al., 2026) and HY-Embodied-0.5 (Tencent Robotics X & HY Vision Team, 2026), with the latter using 4B total and 2.2B active parameters. Qwen3.5-397B-A17B FP8 (Qwen Team, 2026b) serves as the larger reference model.

C.2 Reinforcement learning gains

Under the shared evaluation settings, the final RL model answers 1,538 of 3,360 questions correctly, compared with 1,308 after SFT. Overall accuracy therefore improves from 38.93% to 45.77%, a net

gain of 230 correct answers or 6.85 percentage points, computed from unrounded counts. Core-12 macro accuracy increases from 37.78% to 44.34%. Below, we analyze how these gains are distributed across the categories in Table 3.

Distribution across tasks. All four categories improve: depth perception rises from 51.98% to 57.19%, geometric reasoning from 35.42% to 41.77%, action decisions from 25.94% to 34.06%, and General from 45.83% to 54.37%. Improvements extend to eleven of the twelve core subtypes. The largest gains are in spatial extent (+12.50 pp) and goal-directed approach (+10.83 pp), followed by embodied interaction (+9.17 pp), embodied navigation (+8.75 pp), traversability and clearance (+8.33 pp), and relative depth (+7.50 pp). Object–space compatibility is the exception, decreasing from 52.08% to 51.25%.

Comparison with the 8B baseline. The improvement over SFT and the advantage over Qwen3-VL-8B-Thinking have different distributions. Relative to SFT, RL adds 50 correct perception answers, 61 geometry answers, 78 action answers, and 41 General answers, totaling 230. Relative to the 8B baseline, the final model answers 111 more questions correctly overall: +163 in depth perception, -12 in geometric reasoning, -9 in action decisions, and -31 in General. Thus, RL improves all four categories over its own initialization, while the final model's overall advantage over the 8B baseline is concentrated in perception.

C.3 Depth supervision ablation

Models and supervision. The purple trajectory in Figure 4 follows training of DepthThinker, whose final checkpoint is reported in the main tables. The teal trajectory is the cold-start-only control. Both share the Qwen3-VL-2B-Thinking backbone. Our model receives large-scale depth supervision and the mixed cold-start data, while the control applies the cold-start mixture directly to the base model without the large-scale depth initialization. The cold-start mixture itself includes depth replay and multi-point examples alongside reasoning demonstrations. Thus, "cold-start-only" identifies the initialization strategy; it does not mean that the control receives no depth supervision. The depth and cold-start components of our SFT supervision are described in Section 3.1.

RL settings and reward curves. Both runs use the same task proportions: 40% direct depth estimation, 20% numerical depth reasoning, 30% multiple-choice reasoning, and 10% multi-point depth and ordering. Each update samples 160 prompts with eight responses per prompt. Both use the same task reward definitions, including depth rewards; the thinking-quality factor is applied to reasoning responses and bypassed for direct depth prediction. The displayed training traces extend to steps 160 and 120, respectively.

Panel A shows the mean mixed-task training reward for the two initialization settings. Both curves improve overall, with the depth-supervised model maintaining higher rewards. Panel B shows the corresponding rewards on reasoning tasks, excluding direct depth prediction. Both models receive reasoning cold-start demonstrations; the comparison supports the benefit of including large-scale depth supervision before RL.

C.4 Case-study reference geometry

The two cases in Figure 5 use saved responses from the diagnostic run to illustrate how depth estimates affect decisions under explicit physical constraints. For each case, we first derive the reference decision from the available geometry and then relate the displayed model responses to that criterion. The reference measurements are used to verify the answers and are withheld from model inputs.

Spherical-workspace membership. The indoor question asks how many of three surface points lie within a spherical workspace of radius 3.0 m centered at the camera. The reference camera-to-point ranges are 2.19, 3.93, and 4.42 m, so only Target 1 satisfies the radius constraint. The correct answer is therefore D (exactly one).

Qwen3-VL-2B-Thinking computes distances by treating normalized image coordinates as metric positions and incorrectly selects A (none). DepthThinker uses visual depth cues to estimate the foreground target at approximately 2 m and the other two beyond 3 m. Although approximate, these

estimates place the targets on the correct sides of the workspace boundary and yield the reference decision.

Braking-policy selection. The driving question specifies a speed of 35 km/h, a reaction time of 0.7 s, normal and emergency decelerations of 3.0 and 7.0 m/s^2, and a 2.0 m safety buffer. With $v = 35/3.6$ m/s, the stopping distance before adding the buffer is $d_{\text{stop}} = 0.7v + v^2/(2a)$, giving 22.56 m for normal braking and 13.56 m for emergency braking. The first obstacle intersection is 18.49 m ahead according to the reference geometry. Reserving the buffer leaves 16.49 m of stopping space, and $13.56 < 16.49 < 22.56$ means that only emergency braking satisfies the constraint. The reference answer is D.

The baseline estimates the obstacle at approximately 26 m and selects normal braking (B). Depth-Thinker estimates approximately 20 m, subtracts the buffer, and compares the remaining 18 m with both stopping distances to select D. As in the workspace example, the estimate need not exactly match the reference to resolve the decision: it must be accurate enough to place the relevant quantities on the correct sides of the physical thresholds.