# SPATIALSPEAK: QA-NATIVE RECONSTRUCTION WITH LOCAL AND GLOBAL CONTEXT FOR SPATIAL CHAIN-OF-THOUGHT REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Vision-language models (VLMs) can benefit from geometric priors for multi-view spatial reasoning, yet answer-only training does not directly supervise the intermediate geometric estimates and their use in deriving quantitative spatial answers. We hypothesize that spatial chain-of-thought (CoT) supervision becomes more effective when the VLM first jointly learns complementary local geometry and global scene context through multi-view reconstruction. We introduce **SpatialSpeak**, a two-stage framework that connects QA-native reconstruction pretraining with spatial CoT learning. In Stage I, QA-Native Reconstruction Pretraining (QA-RP) combines marked-point 3D queries for fine-grained local geometry with object-center queries for global scene context across views. Both tasks are formulated as text-based question answering, allowing geometric estimation and subsequent reasoning to share the same autoregressive output interface. In Stage II, spatial CoT with Visual Compensation (CoT-VC) trains the model to express question-relevant geometric estimates and use them to derive answers, with reliability assessment and visual compensation supporting answer refinement when needed. On ReVSI, QA-RP increases the gain from CoT-VC from 2.6 to 6.9 points, and ablations show that both local and global reconstruction supervision are beneficial. SpatialSpeak achieves state-of-the-art results on ReVSI, VSI-Bench, and SPAR-Bench, with a ReVSI score of 62.8 that exceeds the strongest compared baseline by 8.7 points. Code and pretrained models will be released.

## 1 INTRODUCTION

Multi-view vision-language models (VLMs) (Wu et al., 2025; Tong et al., 2024; Huang et al., 2025; Li et al., 2026a) are increasingly equipped with 3D geometric priors (Wang et al., 2025b;a; Wang & Xu, 2026) for spatial question answering. Existing approaches commonly inject features from pretrained reconstruction models into VLMs (Fan et al., 2026; Zheng et al., 2025a; Zhang et al., 2026a), a paradigm adopted by our baseline (Fig. 1a). Reconstruction- (Hu et al., 2026b) and generation-based (Hu et al., 2026a) objectives further improve scene representations. These advances strengthen spatial understanding (Yang et al., 2025b; Cao et al., 2023; 2026; Fan et al., 2024b; Yan & Xu, 2026), but geometric learning alone does not directly teach a VLM how to express and use question-relevant geometric estimates within explicit reasoning traces.

Answer-only supervision leaves this reasoning process implicit: it provides no direct supervision for the intermediate geometric estimates and their use to derive spatial answers. This limitation is particularly important for quantitative spatial questions, for which reliable answers depend on both accurate geometric estimation and appropriate geometric reasoning. We therefore introduce spatial chain-of-thought (CoT) supervision, which explicitly specifies question-relevant geometric estimates and the steps used to derive the final answer. We further hypothesize that such supervision is more effective when the VLM first jointly learns complementary local geometry and global scene context through multi-view reconstruction. Local reconstruction grounds image points in 3D, whereas global reconstruction captures the spatial arrangement of object instances across views.

Based on this insight, we introduce **SpatialSpeak**, a two-stage learning framework that combines QA-Native Reconstruction Pretraining (QA-RP) with spatial Chain-of-Thought with Visual Com-



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.

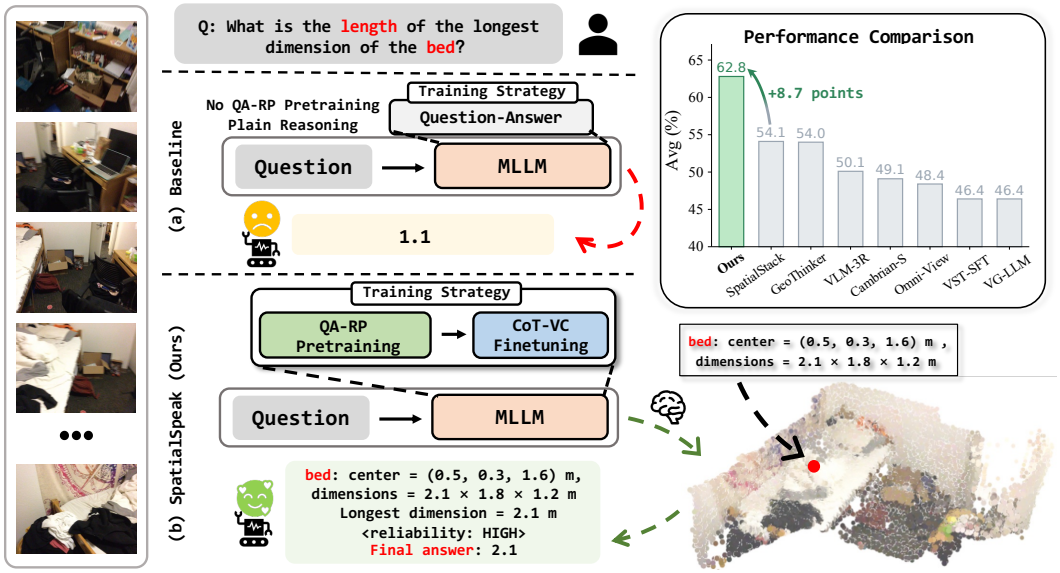


Figure 1: Panel (a) shows our baseline, which incorporates geometric priors through feature fusion and is trained on spatial-reasoning QA without QA-RP or CoT-VC. Panel (b) shows **SpatialSpeak**, which trains the VLM to conduct multi-view 3D reconstruction through QA-Native Reconstruction Pretraining (QA-RP) with complementary local geometry and global scene context, then to involve the learned geometry in explicit spatial reasoning through spatial Chain-of-Thought with Visual Compensation (CoT-VC), achieving leading results on ReVSI (Zhang et al., 2026c). The top-right chart compares average ReVSI scores with SpatialStack-4B (Zhang et al., 2026a), GeoThinker-8B (Li et al., 2026a), VLM-3R-7B (Fan et al., 2026), Cambrian-S-7B (Yang et al., 2026), Omni-View-7B (Hu et al., 2026a), VST-7B-SFT (Yang et al., 2025c), and VG-LLM-8B (Zheng et al., 2025a). SpatialSpeak-4B achieves 62.8, outperforming SpatialStack-4B (54.1) by 8.7 points.

pensation (CoT-VC) supervision (Fig. 1). In Stage I, QA-RP jointly supervises local point reconstruction and global object-center reconstruction. Local queries ask the VLM to predict the 3D location of a marked image point from multi-view inputs, providing fine-grained geometric supervision. Global queries ask it to enumerate visible object instances and predict their semantic categories and 3D centers in a shared coordinate system, providing scene-wide object-layout supervision. Crucially, both tasks are formulated as text-based question answering and trained with the standard next-token prediction objective. This QA-native design aligns geometric prediction with the autoregressive output interface later used for spatial reasoning.

In Stage II, CoT-VC trains the pretrained VLM to identify question-relevant entities, express their geometric estimates, and derive answers through explicit task-specific computations. Because geometric derivations can be approximate, CoT-VC also supervises reliability assessment and answer refinement using direct visual evidence when needed. On ReVSI, QA-RP increases the gain from CoT-VC from 2.6 to 6.9 points, and ablating either local or global supervision degrades performance. SpatialSpeak achieves state-of-the-art results on ReVSI (Zhang et al., 2026c), VSI-Bench (Yang et al., 2025b) and SPAR-Bench (Zhang et al., 2025).

Our contributions are as follows:

- We establish that multi-view reconstruction pretraining substantially increases the benefit of explicit spatial reasoning supervision: the gain from CoT-VC rises from 2.6 points without QA-RP to 6.9 points with QA-RP on ReVSI. SpatialSpeak achieves state-of-the-art results on ReVSI, VSI-Bench, and SPAR-Bench.
- We propose QA-Native Reconstruction Pretraining (QA-RP), which learns complementary local point geometry and global object-layout context through text-based QA targets in a shared 3D coordinate system.
- We introduce spatial Chain-of-Thought with Visual Compensation (CoT-VC), which explicitly supervises question-relevant geometric estimates, task-specific answer derivations, reliability assessment, and visually grounded answer refinement.

## 2 Related Work

**3D Reconstruction.** Given multi-view images without known poses, 3D reconstruction (Hartley & Zisserman, 2003; Leroy et al., 2024; Zhong et al., 2026; Mi et al., 2026) seeks to recover both the scene's geometry and the associated camera trajectories. Traditional pipelines split this objective into modules such as multi-view stereo (Wang et al., 2021; Zhang et al., 2020), feature matching (Lindenberger et al., 2023; Sun et al., 2021), and keypoint detection (Lowe, 2004; DeTone et al., 2018), among others. More recently, DUSt3R (Wang et al., 2024b) proposed a unified network that predicts scene structure directly, shifting the conventional paradigm. Then MASt3R (Leroy et al., 2024) adds an auxiliary head for correspondence estimation. However, both DUSt3R and MASt3R process only image pairs, which restricts global context, requires multiple forward passes, and entails an expensive global alignment stage. Fast3R (Yang et al., 2025a) mitigates these issues by consuming long sequences in a single pass and eliminating coordinate alignment. InstantSplat (Fan et al., 2024a) is a fast, self-supervised method that uses Gaussian Bundle Adjustment and co-visibility to jointly recover geometry from 2-3 unposed views. CUT3R (Wang et al., 2025b) introduces a recurrent transformer that incrementally produces a unified, metric-scale reconstruction from image streams. Beyond point maps, VGGT (Wang et al., 2025a) further estimates camera poses and other 3D attributes. DepthLM (Cai et al., 2026a) demonstrates expert-level single-image metric depth estimation with text-based SFT. It also extends to two-point distance estimation and camera displacement estimation from image pairs. Our focus is on learning multi-view geometry with local and global context to support spatial CoT.

**Spatial MLLMs.** Endowing multimodal large language models (MLLMs), also called vision-language models (VLMs) (Singh et al., 2025; Gemini Team, 2023; Bai et al., 2025a; Zheng et al., 2025b; Xu et al., 2025; Li et al., 2025c; Wang et al., 2025c; Zhu et al., 2026), with fine-grained spatial understanding has attracted growing attention. A dominant paradigm relies on pretrained feed-forward 3D reconstruction models (Wang et al., 2025a;b; 2024b) as external geometry providers and injects their reconstruction prior into the language model's representation space. Along this line, *token-level feature fusion* is a common mechanism. VG-LLM (Zheng et al., 2025a) directly adds geometric tokens to visual tokens. Spatial-MLLM (Wu et al., 2025) pairs a 2D semantic visual encoder with a geometry-prior spatial encoder and space-aware frame sampling to improve spatial reasoning. VLM-3R (Fan et al., 2026) applies cross-attention layers that allow visual representations to query geometric tokens. GeoThinker (Li et al., 2026a) further shifts from passive fusion to selective integration of geometric evidence across multiple levels. Beyond feature integration, other approaches incorporate geometric training objectives. GAP-MLLM (Zhang et al., 2026b) combines semantic labeling and sparse 3D point prediction with multi-level gated fusion to improve downstream 3D perception. $G^2$VLM (Hu et al., 2026b) integrates geometric and semantic experts through shared self-attention and learns scene geometry with dedicated prediction heads. In its released question-answering implementation, the semantic branch conditions answer generation on the geometric expert's hidden representations through attention. Omni-View (Hu et al., 2026a) jointly trains 3D scene understanding, novel-view synthesis, and depth and camera-pose estimation, using dedicated texture and geometry modules to strengthen scene understanding. Another strategy is *feature distillation* or alignment. 3DRS (Huang et al., 2025) transfers knowledge from a frozen reconstruction model into the visual encoder, while Spatial Forcing (Li et al., 2025b) enforces direct embedding alignment between visual and geometric streams during training. Beyond token fusion and distillation, SpatialStack (Zhang et al., 2026a) aligns and stacks multi-scale geometric features with the language backbone, and Map2Thought (Gao et al., 2026) leverages external metric-scale scene graphs to facilitate spatial reasoning. Our work investigates how jointly learning local point geometry and global scene context through QA-based reconstruction supervision supports subsequent spatial CoT learning. We explicitly supervise how question-relevant geometric estimates are expressed and used to derive answers, with visual compensation supporting refinement when needed.

## 3 Method

In this section, we present our approach in detail. The examples in Fig. 1 and Fig. 2 are constructed to show the workflow. We begin with *Preliminaries* (Sec. 3.1), introducing our baseline and the notation for its inputs and outputs. We then describe our two-stage reconstruction-to-reasoning learning framework: *Stage I: QA-Native Reconstruction Pretraining with Local and Global Context* (Sec. 3.2), which learns complementary local geometry and global scene context through QA-native

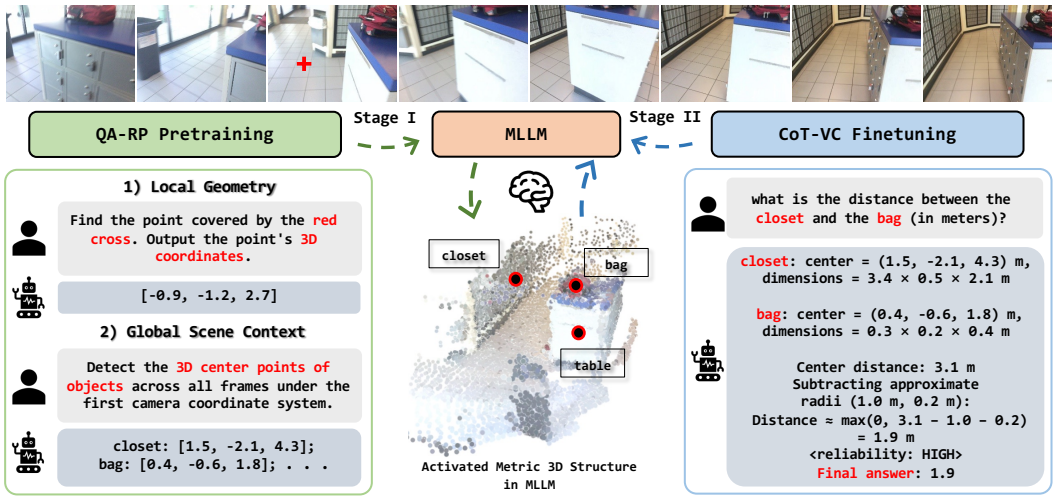


Figure 2: **Overview of the SpatialSpeak framework**. In Training Stage I, QA-Native Reconstruction Pretraining (QA-RP) trains the VLM to conduct metric-scale multi-view reconstruction within the standard QA interface. Local point queries and global object-center queries jointly supervise fine-grained local geometry and global scene context, with both targets expressed in the first-frame camera coordinate system. In Training Stage II, spatial Chain-of-Thought with Visual Compensation (CoT-VC) trains the model to reason explicitly with the learned geometry, assess the reliability of geometry-derived answers, and refine answers using direct visual evidence when needed.

reconstruction supervision, followed by *Stage II: Spatial Chain-of-Thought with Visual Compensation* (Sec. 3.3), which trains the VLM to express geometric estimates and use them to derive spatial answers, with visual compensation supporting answer refinement when needed.

### 3.1 Preliminaries

Our baseline ('MLLM' in Fig. 2) is built on the popular VG-LLM (Zheng et al., 2025a), which takes as input a sequence of $N$ RGB frames $\mathcal{I} = \{I_1, \ldots, I_N\}$ together with a natural-language query $q$ and produces a text response $y$ via autoregressive decoding.

**Feature Encoding.** Each frame $I_i$ is first processed by a vision encoder $\phi_{\text{vis}}$ to obtain a sequence of patch tokens:

$$\mathbf{v}_i = \phi_{\text{vis}}(I_i) \in \mathbb{R}^{L \times d_v}, \quad i = 1, \ldots, N, \tag{1}$$

where $L$ is the number of patch tokens per frame and $d_v$ is the visual feature dimension. In parallel, a frozen multi-view geometry encoder $\phi_{\text{geo}}$ of VGGT (Wang et al., 2025a) processes all $N$ frames jointly and extracts 3D-aware features, which are then fused with visual features, combining geometry priors for the reasoning. Let $\tilde{\mathbf{v}}_i$ denote the geometry-fused visual tokens.

**Language Modeling.** The text query is tokenized into $\mathbf{q} \in \mathbb{Z}^M$. The LLM $f_\theta$ receives the concatenation of all projected visual tokens and query tokens and produces the response via:

$$y = f_\theta\big([\phi_{\text{proj}}(\tilde{\mathbf{v}}_1), \ldots, \phi_{\text{proj}}(\tilde{\mathbf{v}}_N), \mathbf{q}]\big), \tag{2}$$

where $\phi_{\text{proj}}$ is a learned MLP projector aligning visual tokens to the LLM's hidden dimension. We optimize only the LLM parameters with a next-token prediction loss over the target response tokens, keeping the vision encoder, geometry encoder, and projector frozen.

### 3.2 Stage I: QA-Native Reconstruction Pretraining with Local and Global Context

To prepare the VLM for spatial CoT reasoning, Stage I jointly learns complementary local geometry and global scene context through multi-view reconstruction. Local point queries ground marked image points in 3D, while global object-center queries ask the model to enumerate object instances and predict their semantic categories and 3D centers across views. All predicted 3D coordinates are

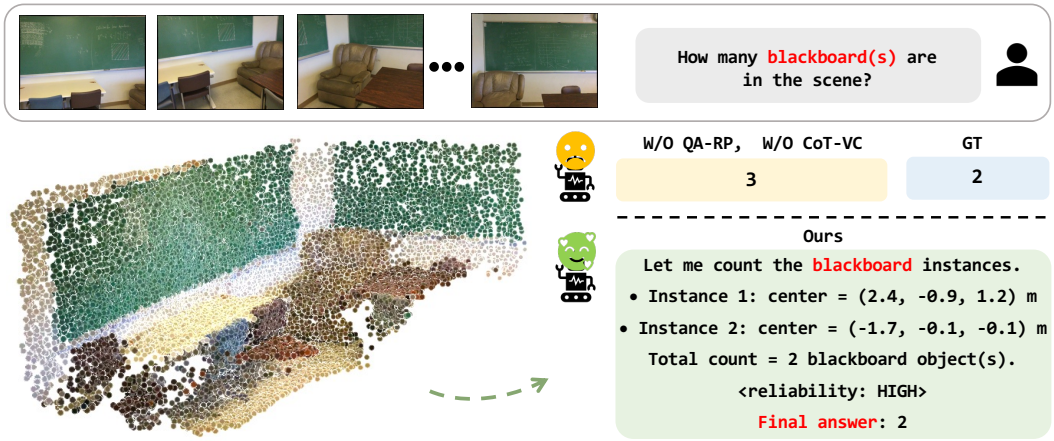


Figure 3: **Qualitative example**. The bottom-left panel shows the scene reconstruction predicted after Stage I (QA-RP), and the bottom-right panel shows the spatial CoT output after Stage II (CoT-VC). SpatialSpeak lists two blackboard instances with their estimated 3D centers and predicts a count of 2, matching the ground truth. The variant with neither QA-RP nor CoT-VC predicts 3.

expressed in the first-frame camera coordinate system, giving points and objects a shared spatial reference. Both tasks are formulated as text-based QA, allowing geometric prediction and subsequent spatial reasoning to share the same autoregressive output interface.

**Local Reconstruction Task.** As shown in 'QA-RP Pretraining' of Fig. 2, to ground fine-grained local geometry in explicit 3D coordinates, we supervise local point-wise 3D reconstruction inspired by prior work (Cai et al., 2026a; Zhang et al., 2026b). For each training sample we select a scene with $N$ video frames $\{I_1, \ldots, I_N\}$ and mark a 2D point $(u, v)$ in frame $I_i$ with a red cross. The local query $q_{\text{local}}$ asks for the point's 3D coordinates in the camera coordinate system of the first frame:

> "Find the point covered by the red cross. Output the point's 3D coordinates."

The ground-truth response is the lifted 3D point $\mathbf{p}^{(1)} = [x, y, z]^\top$ expressed in the first-frame camera coordinate system following VGGT (Wang et al., 2025a), computed from the ScanNet depth maps and known camera extrinsics:

$$\mathbf{p}^{(1)} = \mathbf{R}_{i\to 1}\,\mathbf{p}^{(i)} + \mathbf{t}_{i\to 1}, \tag{3}$$

where $\mathbf{R}_{i\to 1}$, $\mathbf{t}_{i\to 1}$ are the relative rotation and translation from frame $i$ to frame 1. The model outputs the prediction: $\hat{y}_{\text{local}} = \{\texttt{[}\hat{x}\texttt{,}\ \hat{y}\texttt{,}\ \hat{z}\texttt{]}\}$.

**Global Reconstruction Context.** While the local task supervises a single 3D point, it does not explicitly cover the distribution of objects across the scene. We therefore introduce a complementary *global* task that requires enumerating all visible object instances across all $N$ frames. Given the multi-frame input, the global query $q_{\text{global}}$ is:

> "Detect the 3D center points of objects across all frames under the first frame coordinate system."

The ground-truth response is a list of object detections, each consisting of a semantic label $c_k$ and a 3D center $\mathbf{c}_k^{(1)} \in \mathbb{R}^3$ transformed to the first-frame coordinate system. Note that objects are ordered by their first appearance (earlier frames first) and, within the same frame, by spatial location (left-to-right, top-to-bottom), to produce a deterministic, perceptually natural output sequence. This global task complements local point supervision with the semantic categories and spatial arrangement of scene objects in the same reference coordinate system.

**Joint Supervision.** We train the model on a mixture of local and global reconstruction samples using the standard next-token prediction objective:

$$\mathcal{L}_{\text{Stage I}} = \mathbb{E}_{\mathcal{D}_{\text{local}}}\big[-\log p_\theta(y_{\text{local}} \mid \mathcal{I}, q_{\text{local}})\big] + \mathbb{E}_{\mathcal{D}_{\text{global}}}\big[-\log p_\theta(y_{\text{global}} \mid \mathcal{I}, q_{\text{global}})\big]. \tag{4}$$

Both geometric targets are learned through text responses, without auxiliary 3D regression heads or additional geometric losses.

Table 1: **Training components**. QA-RP and CoT-VC each improve spatial reasoning, with larger gains when both designs are used.

| Method | Avg. |
|---|---|
| SpatialSpeak (full) | **62.8** |
| w/o CoT-VC | 55.9 |
| w/o QA-RP | 55.0 |
| w/o both | 52.4 |

Table 2: **QA-RP context**. Ablating global object-center queries, local point queries, or both from multi-view reconstruction pretraining.

| Method | Avg. |
|---|---|
| SpatialSpeak (full) | **62.8** |
| w/o global queries | 59.9 |
| w/o local queries | 59.3 |
| w/o both | 55.0 |

Table 3: **Reliability threshold**. Varying $\tau$ for reliability labels in CoT-VC (Sec. 3.3). All three outperform the variant w/o CoT-VC.

| Method | Avg. |
|---|---|
| SpatialSpeak ($\tau = 0.1$) | 59.6 |
| SpatialSpeak ($\tau = 0.3$) | **62.8** |
| SpatialSpeak ($\tau = 0.5$) | 59.5 |
| w/o CoT-VC | 55.9 |

### 3.3 Stage II: Spatial Chain-of-Thought with Visual Compensation

Stage I supervises scene geometry, but its reconstruction targets do not specify how to derive answers to spatial questions. Stage II therefore complements reconstruction pretraining with spatial CoT supervision that explicitly connects question-relevant geometric estimates to answer derivation. As shown in 'CoT-VC Finetuning' of Fig. 2, we construct structured responses that identify relevant objects, state their geometric estimates, and derive answers through explicit computations. We further include Visual Compensation (VC) to supervise reliability assessment and answer refinement using visual evidence when needed, yielding CoT-VC.

**Spatial CoT Data Construction.** Each CoT training sample in Stage II is a spatial QA pair $(q, a^\star)$ drawn from ScanNet (Dai et al., 2017) following VLM-3R (Fan et al., 2026) enriched with 3D object annotations (semantic labels, 3D bounding boxes). We implement deterministic CoT generators for four question types: *distance*, *size*, *count*, and *closest object*. For each type we derive a 3D estimate $\hat{a}$ from the annotations and compare it with the ground-truth answer $a^\star$ to assign a reliability label.

- **Distance.** Given two object categories $(c_1, c_2)$, we select the most-visible instance of each (the one appearing in the most frames) and compute an approximate closest-point distance as

$$d_{\text{cp}} = \max\big(0,\ \|\mathbf{c}_1^{(1)} - \mathbf{c}_2^{(1)}\|_2 - r_1 - r_2\big), \tag{5}$$

  where $r_k = \frac{1}{6}(s_k^x + s_k^y + s_k^z)$ is the mean half-side of object $k$'s axis-aligned bounding box, used as an approximate radius. The estimate is labeled HIGH if $|d_{\text{cp}} - a^\star|/a^\star \le \tau$; otherwise LOW.
- **Size.** We obtain the longest side length of the most-visible instance's bounding box, $\ell = \max(s^x, s^y, s^z)$, converted to centimeters. Reliability threshold: relative error $\le \tau$.
- **Count.** We count the number of detected instances of category $c$. The estimate is HIGH if and only if it equals the integer ground truth.
- **Closest Object.** For a multiple-choice question, we compute the approximate distance $d_{\text{cp}}$ from Eq. 5 for each option and predict the letter corresponding to the closest option. Reliability is HIGH if the predicted letter matches $a^\star$.

**Spatial CoT Template.** Because geometry-based computations can involve approximations, the derived answer need not always agree with the QA ground truth. We therefore include reliability-conditioned refinement in the CoT target, retaining the geometric derivation while allowing the final answer to be revised using visual evidence. Specifically, each response contains the 3D reasoning chain followed by a reliability token:

$$y_{\text{CoT}} = \underbrace{\mathcal{C}_{\text{3D}}}_{\text{3D reasoning chain}} \oplus \begin{cases} [\text{HIGH}] \oplus \hat{a} & \text{if } \hat{a} \text{ is reliable,} \\ [\text{LOW}] \oplus \mathcal{V} \oplus a^\star & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathcal{C}_{\text{3D}}$ is the 3D reasoning chain always present in the response, [HIGH] / [LOW] are the reliability tokens, $\hat{a}$ is the 3D-derived answer, $\mathcal{V}$ is the visual refinement note, and $a^\star$ is the ground-truth answer used as supervision. When the 3D estimate is reliable (HIGH), the model follows the geometric chain and outputs the 3D-derived answer. When it is not (LOW), the model is trained to acknowledge the limitation, invoke a visual-inspection fallback ("*refining based on visual observation of the scene*"), and output the ground-truth answer. This supervision trains the model to assess its geometric estimates and to refine the final answer with visual evidence when needed.

Table 4: **Spatial CoT**. All model variants retain QA-RP. Removing VC preserves spatial CoT. Removing CoT-VC retains only direct answer supervision in Stage II.

| Method | Avg. |
|---|---|
| SpatialSpeak (full) | **62.8** |
| w/o VC | 58.5 |
| w/o CoT-VC | 55.9 |

Table 5: **Pointmap reconstruction on ScanNet**. Acc. and Comp. are evaluated without alignment. Acc.* and Comp.* are evaluated after GT Sim(3) alignment. All errors are reported in cm. Lower values are better. Baselines are MapAnything (Keetha et al., 2026) and CUT3R (Wang et al., 2025b). Bold type marks the lowest error in each column.

| Method | Acc. ↓ | Comp. ↓ | Acc.* ↓ | Comp.* ↓ |
|---|---|---|---|---|
| MapAnything | 36.3 | 28.4 | 5.7 | 6.2 |
| CUT3R | 13.7 | 12.7 | **4.7** | **4.6** |
| SpatialSpeak (Ours) | **8.9** | **9.3** | 5.0 | 5.0 |




Ground Truth Ours CUT3R MapAnything

Figure 4: **Qualitative reconstruction comparison**. Red dashed arrows mark corresponding distances in meters. SpatialSpeak estimates the distance as 1.24 m, close to the ground truth of 1.25 m.

**Stage II Training.** The model is initialized from the Stage I checkpoint and fine-tuned on the spatial CoT dataset $\mathcal{D}_{\text{CoT}}$ using the same next-token prediction loss:

$$\mathcal{L}_{\text{Stage II}} = \mathbb{E}_{(q,y_{\text{CoT}})\sim\mathcal{D}_{\text{CoT}}}\big[-\log p_\theta(y_{\text{CoT}} \mid \mathcal{I}, q)\big]. \tag{7}$$

For question types not covered by these geometric CoT templates (*e.g.*, relative direction), we retain direct answer supervision to preserve coverage of the full VSI-Bench task distribution.

## 4 Experiments

**Datasets and Benchmarks.** Training Stage I uses the reconstruction pretraining data described in Sec. 3.2. For Training Stage II, following VG-LLM (Zheng et al., 2025a), we use the same training subsets from the LLaVA-Hound split of LLaVA-Video-178K (Zhang et al., 2024b) and SPAR-7M (Zhang et al., 2025), augmented with our spatial CoT data (Sec. 3.3). We evaluate spatial reasoning on ReVSI (Zhang et al., 2026c), VSI-Bench (Yang et al., 2025b), and SPAR-Bench (Zhang et al., 2025). ReVSI corrects annotation errors and reduces answer-distribution bias in VSI-Bench, and is used for all spatial-reasoning ablations. Metric-scale reconstruction is evaluated on ScanNet.

**Implementation Details.** We initialize from the popular Qwen3-VL-4B (Bai et al., 2025a) and fine-tune only the LLM parameters while keeping both the vision encoder and the MLP projector frozen. All stages share the same base configuration: a cosine learning rate schedule with a warm-up ratio of 0.03, weight decay of 0.01, and 32 frames sampled per video clip. Stage I trains for 1 epoch with a global batch size of 32 and a peak learning rate of $5 \times 10^{-6}$. Stage II trains for 1 epoch with a global batch size of 64 and a peak learning rate of $1 \times 10^{-5}$. More details are given in Appendix Sec. A.3.

### 4.1 Ablation Study

We conduct spatial-reasoning ablations on ReVSI and report the average score across its seven task categories. We also evaluate Stage I reconstruction on ScanNet. Throughout the ablations, 'w/o CoT-VC' retains Stage II training with answer-only supervision.

**Effectiveness of Key Components.** Tab. 1 reports the contribution of each component on ReVSI. The baseline without either component scores 52.4. QA-RP alone raises the score to 55.9, while CoT-VC alone raises it to 55.0. Combining both achieves 62.8, an improvement of 10.4 points over the baseline. Removing QA-RP or CoT-VC from the full method reduces the score by 7.8 or 6.9 points, respectively. The gain from CoT-VC increases from 2.6 without QA-RP to 6.9 points with QA-RP. The larger gain from CoT-VC after QA-RP supports our hypothesis that jointly learning local geometry and global scene context provides a stronger foundation for spatial CoT learning.

**Analysis of Reconstruction Pretraining.** Tab. 2 examines the local and global supervision in QA-RP. The full method scores 62.8, compared with 59.9 when global object-center queries are removed

Table 6: **Comparison with state-of-the-art methods on ReVSI**. We evaluate SpatialSpeak and official checkpoints of SpatialStack, GeoThinker, VG-LLM, and Omni-View under the same 32-frame setting. Other baseline results are sourced from the ReVSI paper (Zhang et al., 2026c).

| Method | Avg. | Obj. Count | Abs. Dist. | Obj. Size | Room Size | Rel. Dist. | Rel. Dir. | Route Plan |
|---|---|---|---|---|---|---|---|---|
| | | Numerical Answer | | | | Multiple-Choice Answer | | |
| SpaceR-7B (SG-RLVR) (Ouyang et al., 2025) | 30.5 | 30.7 | 34.5 | 52.0 | 18.6 | 22.8 | 34.5 | 20.2 |
| Spatial-MLLM-4B-135k (Wu et al., 2025) | 40.5 | 40.7 | 45.3 | 46.8 | – | 32.3 | 37.4 | – |
| Spatial-MLLM-4B-820k (Wu et al., 2025) | 40.9 | 41.5 | 40.0 | 53.1 | – | 30.7 | 39.2 | – |
| VST-7B-SFT (Yang et al., 2025c) | 46.4 | 35.4 | 52.6 | 67.9 | 47.2 | 49.2 | 36.9 | 35.4 |
| VG-LLM-8B (Zheng et al., 2025a) | 46.4 | 37.8 | 53.0 | 56.8 | 48.0 | 57.2 | 33.8 | 38.0 |
| Omni-View-7B (Hu et al., 2026a) | 48.4 | 43.1 | 52.1 | 64.2 | 52.4 | 45.9 | 38.5 | 42.2 |
| Cambrian-S-7B (Yang et al., 2026) | 49.1 | 48.4 | 60.5 | 65.5 | 46.7 | 37.1 | 48.5 | 37.0 |
| VLM-3R-7B (Fan et al., 2026) | 50.1 | 41.6 | 61.6 | 64.8 | 52.5 | 46.5 | 49.5 | 34.1 |
| GeoThinker-8B (Li et al., 2026a) | 54.0 | 37.1 | 71.5 | 66.0 | 53.0 | 61.9 | 47.5 | 41.5 |
| SpatialStack-4B (Zhang et al., 2026a) | 54.1 | 36.3 | 64.6 | 63.2 | 54.4 | 62.9 | 47.7 | 49.6 |
| SpatialSpeak-4B (Ours) | **62.8** | 64.9 | 70.2 | 62.5 | 66.0 | 71.9 | 51.1 | 52.7 |

and 59.3 when local point queries are removed. Removing both reduces the score to 55.0. These results support jointly learning local geometry and global scene context during reconstruction pretraining to better prepare the model for subsequent spatial reasoning.

**Analysis of the Reliability Threshold.** Tab. 3 evaluates the relative-error threshold $\tau$ used to construct reliability labels. Among the tested values, $\tau = 0.3$ achieves the highest score of 62.8, compared with 59.6 for $\tau = 0.1$ and 59.5 for $\tau = 0.5$. All three settings outperform the variant without CoT-VC (55.9).

**Analysis of Spatial CoT with Visual Compensation.** Tab. 4 examines the roles of spatial CoT and VC in Stage II. Starting from the model without CoT-VC, which scores 55.9, adding spatial CoT without VC improves the score to 58.5. Incorporating VC further raises it to 62.8, an additional gain of 4.3 points. These results support explicitly training the model both to reason with geometric estimates and to refine its answers when needed.

**3D Pointmap Reconstruction Accuracy.** In Tab. 5, we compare our Stage I model with CUT3R and MapAnything on the same 200 randomly sampled ScanNet validation scenes held out from training, under two protocols: no alignment and GT Sim(3) alignment. Without alignment, SpatialSpeak achieves the lowest Acc. and Comp. errors (8.9 and 9.3 cm), compared with CUT3R (13.7 and 12.7 cm) and MapAnything (36.3 and 28.4 cm). After GT Sim(3) alignment, SpatialSpeak obtains 5.0 cm on both metrics, lower than MapAnything (5.7 and 6.2 cm) but slightly higher than CUT3R (4.7 and 4.6 cm). The gaps narrow after GT Sim(3) alignment. These results demonstrate the Stage I VLM's ability to predict metric-scale geometry, while the ReVSI ablations show that reconstruction pretraining benefits spatial reasoning. These evaluations support our goal of learning geometry that serves downstream reasoning. Fig. 4 compares our Stage I reconstruction with CUT3R and MapAnything. SpatialSpeak's estimate of the marked distance is closer to ground truth. Enlarged visualizations and a comparison on an additional scene are provided in Fig. 7 in the appendix.

### 4.2 Main Results

We report results on ReVSI and SPAR-Bench, alongside VSI-Bench under normal and scaled training settings to cover a broader range of prior methods.

**ReVSI.** In Tab. 6, we compare against competitive methods on ReVSI (Zhang et al., 2026c), our primary spatial-reasoning benchmark. Other methods that do not report ReVSI results are skipped. Our method achieves an average score of 62.8, exceeding SpatialStack-4B (54.1), the highest-scoring competing method in the table, by 8.7 points. Our method obtains the best reported result on five of the seven task categories. Compared with SpatialStack-4B, the gains include object counting (64.9 vs. 36.3), room size (66.0 vs. 54.4), and relative distance (71.9 vs. 62.9). Together with the ReVSI ablations, these results support our strategy of learning multi-view geometry through reconstruction pretraining and training its use in explicit spatial reasoning through CoT with visual compensation.

**VSI-Bench** (normal training setting). Following VG-LLM (Zheng et al., 2025a), we adopt the same training sets sampled from the LLaVA-Hound split of LLaVA-Video-178K (Zhang et al., 2024b) and SPAR-7M (Zhang et al., 2025), and additionally incorporate the spatial CoT data constructed as described in Sec. 3.3. As shown in Tab. 7, our method achieves an average score of 63.3, the highest among the compared methods. It exceeds Omni-View-7B (55.4) by **7.9 points**.

Table 7: **Comparison with state-of-the-art methods on VSI-Bench** (normal training setting).

| Method | Avg. | Obj. Count | Abs. Dist. | Obj. Size | Room Size | Rel. Dist. | Rel. Dir. | Route Plan | Appr. Order |
|---|---|---|---|---|---|---|---|---|---|
| | | Numerical Answer | | | | Multiple-Choice Answer | | | |
| *Proprietary Models (API)* | | | | | | | | | |
| GPT-4o | 34.0 | 46.2 | 5.3 | 43.8 | 38.2 | 37.0 | 41.3 | 31.5 | 28.5 |
| Gemini-1.5-Flash | 42.1 | 49.8 | 30.8 | 53.5 | 54.4 | 37.7 | 41.0 | 31.5 | 37.8 |
| Gemini-1.5-Pro | 45.4 | 56.2 | 30.9 | 64.1 | 43.6 | 51.3 | 46.3 | 36.0 | 34.6 |
| *Open-source Models* | | | | | | | | | |
| InternVL2-8B (Chen et al., 2024b) | 34.6 | 23.1 | 28.7 | 48.2 | 39.8 | 36.7 | 30.7 | 29.9 | 39.6 |
| InternVL2-40B (Chen et al., 2024b) | 36.0 | 34.9 | 26.9 | 46.5 | 31.8 | 42.1 | 32.2 | 34.0 | 39.6 |
| LongVILA-8B (Chen et al., 2025) | 21.6 | 29.1 | 9.1 | 16.7 | 0.0 | 29.6 | 30.7 | 32.5 | 25.5 |
| VILA-1.5-40B (Lin et al., 2024) | 31.2 | 22.4 | 24.8 | 48.7 | 22.7 | 40.5 | 25.7 | 31.5 | 32.9 |
| LongVA-7B (Zhang et al., 2024a) | 29.2 | 38.0 | 16.6 | 38.9 | 22.2 | 33.1 | 43.3 | 25.4 | 15.7 |
| LLaVA-Video-72B (Zhang et al., 2024b) | 40.9 | 48.9 | 22.8 | 57.4 | 35.3 | 42.4 | 36.7 | 35.0 | 48.6 |
| LLaVA-OneVision-72B (Li et al., 2025a) | 40.2 | 43.5 | 23.9 | 57.6 | 37.5 | 42.5 | 39.9 | 32.5 | 44.6 |
| *Spatial-Enhanced Models* | | | | | | | | | |
| SAT-LLaVA-Video-7B (Ray et al., 2025) | - | - | - | - | 47.3 | 41.1 | 37.1 | 36.1 | 40.4 |
| SPAR-8B (Zhang et al., 2025) | 41.1 | - | - | - | - | - | - | - | - |
| UniUGG-3B (Xu et al., 2025) | 42.2 | - | - | - | - | - | - | - | - |
| SpaceR-7B (SG-RLVR) (Ouyang et al., 2025) | 45.6 | - | - | - | - | - | - | - | - |
| 3DRS-7B (Huang et al., 2025) | 45.9 | 68.7 | 34.8 | 53.6 | 56.6 | 40.9 | 43.2 | 30.4 | 39.2 |
| SpatialLadder-3B (Li et al., 2026b) | 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) | 47.0 | 65.3 | 34.8 | 63.1 | 45.1 | 41.3 | 46.9 | 33.5 | 46.3 |
| VG-LLM-4B (Zheng et al., 2025a) | 47.3 | 66.0 | 37.8 | 55.2 | 59.2 | 44.6 | 45.6 | 33.5 | 36.4 |
| VG-LLM-8B (Zheng et al., 2025a) | 50.7 | 67.9 | 37.7 | 58.6 | 62.0 | 46.6 | 40.7 | 32.4 | 59.2 |
| GeoThinker-7B (Li et al., 2026a) | 50.5 | 69.5 | 38.5 | 57.9 | 62.2 | 45.2 | 46.2 | 31.4 | 52.6 |
| Omni-View-7B (Hu et al., 2026a) | 55.4 | 70.3 | 46.4 | 68.6 | 54.7 | 65.9 | 54.4 | 33.5 | 49.0 |
| SpatialSpeak-4B (Ours) | **63.3** | 72.2 | 56.9 | 67.6 | 62.8 | 62.5 | 86.2 | 43.3 | 55.2 |

Table 8: **Comparison with state-of-the-art methods on VSI-Bench** (scaled training setting).

| Method | Avg. | Obj. Count | Abs. Dist. | Obj. Size | Room Size | Rel. Dist. | Rel. Dir. | Route Plan | Appr. Order |
|---|---|---|---|---|---|---|---|---|---|
| | | Numerical Answer | | | | Multiple-Choice Answer | | | |
| Qwen3-VL-8B (Bai et al., 2025a) | 59.8 | 67.5 | 52.6 | 76.2 | 62.3 | 60.6 | 52.5 | 32.5 | 73.8 |
| VLM-3R-7B (Fan et al., 2026) | 60.9 | 70.2 | 49.4 | 69.2 | 67.1 | 65.4 | 80.5 | 45.4 | 40.1 |
| Map2Thought-7B (Gao et al., 2026) | 61.0 | 70.8 | 55.0 | 70.1 | 69.4 | 56.9 | 69.8 | 38.1 | 57.4 |
| VST-7B (Yang et al., 2025c) | 61.2 | - | - | - | - | - | - | - | - |
| VG-LLM-8B (Zheng et al., 2025a) | 62.2 | 71.4 | 56.8 | 69.0 | 69.1 | 67.9 | 83.2 | 47.4 | 32.5 |
| 3DThinker-7B (Chen et al., 2026) | 63.7 | - | - | - | - | - | - | - | - |
| SpatialStack-4B (Zhang et al., 2026a) | 67.5 | 71.0 | 55.6 | 69.1 | 68.2 | 67.3 | 84.1 | 41.2 | 83.5 |
| Cambrian-S-7B (Yang et al., 2026) | 67.5 | 73.2 | 50.5 | 74.9 | 72.2 | 71.1 | 76.2 | 41.8 | 80.1 |
| GeoThinker-7B (Li et al., 2026a) | 68.5 | - | - | - | - | - | - | - | - |
| SenseNova-SI-8B (Cai et al., 2026b) | 68.8 | 72.0 | 53.5 | 76.8 | 72.8 | 69.6 | 80.8 | 48.5 | 76.4 |
| GeoThinker-8B (Li et al., 2026a) | 72.6 | - | - | - | - | - | - | - | - |
| SpatialSpeak-4B (Ours) | **73.0** | 73.8 | 60.7 | 76.2 | 79.9 | 70.7 | 88.0 | 47.9 | 86.9 |

**VSI-Bench** (scaled training setting). We extend the training data to include SPAR (Zhang et al., 2025), VLM-3R (Fan et al., 2026), and VSI-590K (Yang et al., 2026), again augmented with our spatial CoT data (Sec. 3.3). As shown in Tab. 8, SpatialSpeak achieves 73.0 on VSI-Bench with a 4B VLM backbone, compared with 72.6 for GeoThinker-8B.

**SPAR-Bench.** We further evaluate on SPAR-Bench (Zhang et al., 2025), with detailed results provided in Tab. 9 of the appendix. SpatialSpeak achieves an average score of 76.0, exceeding SpatialStack-4B (72.0) and GeoThinker-8B (68.2) by 4.0 and 7.8 points, respectively. These results complement the ReVSI and VSI-Bench evaluations, supporting the effectiveness of our reconstruction-to-reasoning training strategy across a broader range of spatial tasks.

**Qualitative Examples.** Fig. 3 presents an example from ReVSI (Zhang et al., 2026c). Given the question "How many blackboards are in the scene?", SpatialSpeak enumerates two blackboard instances and expresses their estimated 3D centers, correctly predicting a count of 2, whereas the variant with neither QA-RP nor CoT-VC predicts 3. Additional examples appear in Appendix Sec. A.2.

## 5 Conclusion

We have presented SpatialSpeak, a two-stage training framework that combines multi-view reconstruction pretraining for local geometry and global scene context with explicit spatial reasoning supervision. QA-RP uses local point and global object-center queries to supervise geometry through the VLM's standard QA outputs. Spatial CoT finetuning then teaches the model to express question-relevant geometric estimates and use them to derive answers, with visual compensation supporting refinement when needed. SpatialSpeak achieves state-of-the-art performance on ReVSI, VSI-Bench and SPAR-Bench. ReVSI ablations show that local and global reconstruction supervision both contribute to spatial reasoning, and that QA-RP increases the gains from the subsequent CoT-VC stage. These findings support the complementary roles of learning multi-view geometry and explicitly supervising its use in spatial reasoning within our framework. We hope this work motivates exploration of reconstruction-augmented reasoning in vision-language models.

### AI USE STATEMENT

We used generative AI tools to check grammar, improve the clarity and organization of the manuscript, and assist with developing selected scripts for tasks such as data processing. The research ideas originated from the authors, who also conducted the experiments. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## 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. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory P. Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. In *ICLR*, 2026a.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, Qingping Sun, Hui En Pang, Jiaqi Li, Oscar Qian, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. In *CVPR*, 2026b.

Yang Cao, Yihan Zeng, Hang Xu, and Dan Xu. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection. In *NeurIPS*, 2023.

Yang Cao, Feize Wu, Dave Zhenyu Chen, Yingji Zhong, Lanqing Hong, and Dan Xu. Vggt-det: Mining vggt internal priors for sensor-geometry-free multi-view indoor 3d object detection. In *CVPR*, 2026.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. In *ICLR*, 2025.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. In *CVPR*, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024b.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018.

Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantsplat: Sparse-view gaussian splatting in seconds. *arXiv preprint arXiv:2403.20309*, 2024a.

Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, Marco Pavone, and Yue Wang. Large spatial model: End-to-end unposed images to semantic 3d. In *NeurIPS*, 2024b.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *CVPR*, 2026.

Xiangjun Gao, Zhensong Zhang, Dave Zhenyu Chen, Songcen Xu, Long Quan, Eduardo Pérez-Pellitero, and Youngkyoon Jang. Map2thought: Explicit 3d spatial reasoning via metric cognitive maps. *arXiv preprint arXiv:2601.11442*, 2026.

Gemini Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

JiaKui Hu, Shanshan Zhao, Qing-Guo Chen, Xuerui Qiu, Jialun Liu, Zhao Xu, Weihua Luo, Kaifu Zhang, and Yanye Lu. Omni-view: Unlocking how generation facilitates understanding in unified 3d model based on multiview images. In *ICLR*, 2026a.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. $G^2$vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. In *CVPR*, 2026b.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3drs: Mllms need 3d-aware representation supervision for scene understanding. In *NeurIPS*, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *3DV*, 2026.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024.

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. *TMLR*, 2025a.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025b.

Haoyuan Li, Yanpeng Zhou, Yufei Gao, Tao Tang, Jianhua Han, Yujie Yuan, Dave Zhenyu Chen, Jiawang Bian, Hang Xu, and Xiaodan Liang. Does your 3d encoder really work? when pretrain-sft from 2d vlms meets 3d vlms. *arXiv preprint arXiv:2506.05318*, 2025c.

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. In *ICML*, 2026a.

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. In *ICLR*, 2026b.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

Zhenxing Mi, Yuxin Wang, and Dan Xu. One4d: Unified 4d generation and reconstruction via decoupled lora control. In *ECCV*, 2026.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models. In *COLM*, 2025.

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.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021.

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. *arXiv preprint arXiv:2406.16860*, 2024.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, 2021.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024b.

Yuxin Wang, Lei Ke, Boqiang Zhang, Tianyuan Qu, Hanxun Yu, Zhenpeng Huang, Meng Yu, Dan Xu, and Dong Yu. N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. *arXiv preprint arXiv:2512.16561*, 2025c.

Zipeng Wang and Dan Xu. Flashvggt: Efficient and scalable visual geometry transformers with compressed descriptor attention. In *CVPR*, 2026.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. In *NeurIPS*, 2025.

Yueming Xu, Jiahui Zhang, Ze Huang, Yurui Chen, Yanpeng Zhou, Zhenyu Chen, Yu-Jie Yuan, Pengxiang Xia, Guowei Huang, Xinyue Cai, et al. Uniugg: Unified 3d understanding and generation via geometric-semantic encoding. *arXiv preprint arXiv:2508.11952*, 2025.

Chi Yan and Dan Xu. Progressive gaussian transformer with anisotropy-aware sampling for open vocabulary occupancy prediction. In *ICLR*, 2026.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. *arXiv preprint arXiv:2501.13928*, 2025a.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025b.

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. *arXiv preprint arXiv:2511.05491*, 2025c.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. In *ICLR*, 2026.

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. In *NeurIPS*, 2025.

Jian Zhang, Shijie Zhou, Bangya Liu, Achuta Kadambi, and Zhiwen Fan. Spatialstack: Layered geometry-language fusion for 3d vlm spatial reasoning. In *CVPR*, 2026a.

Jiaxin Zhang, Junjun Jiang, Haijie Li, Youyu Chen, Kui Jiang, and Dave Zhenyu Chen. Gap-mllm: Geometry-aligned pre-training for activating 3d spatial perception in multimodal large language models. In *ECCV*, 2026b.

Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928*, 2020.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024a.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning. In *ICML*, 2026c.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *NeurIPS*, 2025a.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *CVPR*, 2025b.

Yingji Zhong, Kaichen Zhou, Zhihao Li, Lanqing Hong, Zhenguo Li, and Dan Xu. Empowering sparse-input neural radiance fields with dual-level semantic guidance from dense novel views. In *AAAI*, 2026.

Zhe Zhu, Yiheng Zhang, Peng Li, Zixing Zhao, Honghua Chen, Yaqing Zhang, Le Wan, Zhiyang Dou, Cheng Lin, Yuan Liu, Mingqiang Wei, and Wenping Wang. Partllm: A unified multimodal foundation for 3d part segmentation. *arXiv preprint arXiv:2609.25832*, 2026.

# A APPENDIX

## A.1 EVALUATION ON SPAR-BENCH

Table 9: **Comparison with state-of-the-art models on SPAR-Bench** (Zhang et al., 2025). Baselines include InternVL2 (Chen et al., 2024b), InternVL2.5 (Chen et al., 2024a), LLaVA-OV (Li et al., 2025a), Qwen2-VL (Wang et al., 2024a), Qwen2.5-VL (Bai et al., 2025b), LLaVA-v1.5 (Liu et al., 2024a), LLaVA-v1.6 (Liu et al., 2024b), Spatial-MLLM (Wu et al., 2025), VLM-3R (Fan et al., 2026), UniUGG-3B (Xu et al., 2025), G$^2$VLM-SR (Hu et al., 2026b), GeoThinker (Li et al., 2026a), SenseNova-SI (Cai et al., 2026b) and SpatialStack (Zhang et al., 2026a). Baseline results are primarily sourced from the supplementary material of G$^2$VLM (Hu et al., 2026b).

| Method | *Avg.* | *Low* | *Depth-OC* | *Depth-OC-MV* | *Depth-OO* | *Depth-OO-MV* | *Dist-OC* | *Dist-OC-MV* | *Dist-OO* | *Dist-OO-MV* | *Medium* | *PoseMatch* | *CamMotion* | *ViewChgI* | *High* | *Distl-OO* | *Distl-OO-MV* | *ObjRel-OC-MV* | *ObjRel-OO* | *ObjRel-OO-MV* | *SpImag-OC* | *SpImag-OC-MV* | *SpImag-OO* | *SpImag-OO-MV* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| InternVL2-2B | 28.1 | 21.7 | 18.1 | 24.8 | 23.2 | 21.0 | 19.5 | 20.0 | 26.8 | 20.6 | 22.8 | 39.7 | 23.0 | 5.8 | 35.4 | 51.2 | 56.0 | 46.0 | 31.6 | 23.8 | 36.0 | 34.3 | 17.6 | 22.4 |
| InternVL2-4B | 32.0 | 28.9 | 23.9 | 27.2 | 20.0 | 18.1 | 42.6 | 40.2 | 31.3 | 28.2 | 29.2 | 49.9 | 21.0 | 16.6 | 35.7 | 56.8 | 55.4 | 40.3 | 36.8 | 25.2 | 28.8 | 32.3 | 21.2 | 24.7 |
| InternVL2.5-2B | 30.1 | 25.8 | 39.7 | 39.7 | 12.1 | 15.0 | 30.9 | 29.6 | 20.2 | 19.0 | 22.9 | 37.9 | 24.3 | 6.6 | 36.4 | 51.5 | 56.9 | 50.3 | 33.8 | 24.1 | 27.2 | 35.2 | 26.5 | 22.4 |
| InternVL2.5-4B | 30.6 | 25.7 | 29.1 | 33.0 | 21.8 | 16.8 | 20.8 | 26.9 | 28.1 | 28.8 | 29.8 | 47.1 | 33.3 | 8.9 | 35.2 | 54.1 | 58.9 | 35.5 | 29.7 | 34.6 | 24.7 | 31.4 | 19.2 | 28.3 |
| InternVL2.5-8B | 36.3 | 29.5 | 25.8 | 29.3 | 23.8 | 18.8 | 46.8 | 42.7 | 22.6 | 25.9 | 31.9 | 61.3 | 28.0 | 6.3 | 43.8 | 59.7 | 56.9 | 51.8 | 44.2 | 41.6 | 36.6 | 41.6 | 22.5 | 39.5 |
| LLaVA-OV-0.5B | 29.5 | 30.1 | 49.2 | 42.7 | 18.0 | 14.9 | 31.5 | 25.7 | 29.0 | 30.1 | 15.9 | 24.4 | 21.8 | 1.5 | 33.4 | 50.9 | 50.0 | 32.0 | 27.8 | 26.0 | 30.9 | 34.0 | 24.5 | 24.7 |
| LLaVA-OV-7B | 31.2 | 21.8 | 30.3 | 26.9 | 18.6 | 13.9 | 10.4 | 13.6 | 31.2 | 29.3 | 26.1 | 38.7 | 30.3 | 9.5 | 40.1 | 56.5 | 55.1 | 37.3 | 48.6 | 38.2 | 30.4 | 33.7 | 26.5 | 35.0 |
| Qwen2-VL-2B | 24.6 | 19.4 | 38.0 | 40.6 | 18.8 | 14.1 | 7.8 | 7.1 | 17.8 | 11.1 | 27.6 | 26.2 | 25.3 | 31.2 | 28.2 | 54.1 | 49.1 | 21.8 | 25.3 | 12.5 | 23.9 | 27.6 | 24.8 | 14.9 |
| Qwen2-VL-7B | 30.7 | 27.5 | 36.0 | 35.2 | 20.8 | 12.9 | 28.7 | 30.0 | 28.2 | 28.5 | 20.4 | 35.4 | 20.3 | 5.7 | 37.0 | 59.7 | 52.4 | 30.3 | 38.5 | 41.0 | 22.0 | 28.5 | 22.5 | 38.4 |
| Qwen2.5-VL-3B | 29.4 | 26.7 | 31.7 | 34.2 | 32.1 | 17.5 | 18.4 | 22.7 | 32.1 | 24.8 | 24.9 | 39.2 | 27.3 | 8.1 | 33.3 | 55.6 | 60.7 | 37.5 | 32.1 | 20.2 | 21.0 | 27.0 | 20.9 | 24.6 |
| Qwen2.5-VL-7B | 33.1 | 28.8 | 31.3 | 33.7 | 22.0 | 15.0 | 42.9 | 37.7 | 23.8 | 23.6 | 23.0 | 33.3 | 28.8 | 6.8 | 40.3 | 58.2 | 51.5 | 44.8 | 50.0 | 32.1 | 33.9 | 32.9 | 27.2 | 31.9 |
| LLaVA-v1.5-7B | 23.7 | 10.9 | 5.2 | 12.5 | 17.4 | 11.3 | 7.3 | 5.3 | 18.7 | 9.1 | 26.5 | 24.4 | 26.8 | 28.3 | 34.1 | 51.2 | 52.4 | 34.3 | 24.2 | 26.9 | 34.7 | 29.9 | 22.5 | 30.8 |
| LLaVA-v1.6-7B | 13.2 | 8.5 | 12.1 | 0.0 | 20.4 | 0.3 | 10.8 | 0.4 | 24.3 | 0.0 | 4.8 | 6.6 | 7.8 | 0.0 | 20.2 | 51.8 | 7.7 | 6.3 | 32.1 | 6.4 | 39.5 | 10.5 | 21.5 | 5.9 |
| Spatial-MLLM-7B | 32.2 | 29.9 | 31.9 | 22.9 | 22.8 | 16.4 | 35.9 | 38.7 | 35.5 | 34.9 | 20.3 | 34.1 | 26.8 | 0.0 | 38.1 | 54.7 | 50.9 | 39.0 | 34.6 | 24.7 | 38.7 | 41.3 | 28.8 | 30.5 |
| VLM-3R-7B | 43.2 | 39.8 | 47.8 | 45.6 | 40.1 | 20.6 | 42.2 | 44.3 | 40.1 | 37.5 | 28.4 | 42.0 | 30.0 | 13.3 | 51.2 | 55.9 | 59.2 | 58.8 | 53.0 | 54.6 | 47.3 | 50.6 | 30.5 | 50.7 |
| SenseNova-SI-8B | 45.8 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| UniUGG-3B | 50.6 | 50.8 | - | - | - | - | - | - | - | - | 49.1 | - | - | - | 51.9 | - | - | - | - | - | - | - | - | - |
| G$^2$VLM-SR-2B | 54.9 | 60.0 | 80.3 | 73.8 | 21.4 | 18.9 | 78.4 | 75.2 | 68.4 | 63.6 | 36.3 | 27.0 | 28.2 | 53.3 | 56.5 | 53.5 | 49.1 | 76.8 | 50.0 | 68.7 | 50.5 | 52.6 | 44.4 | 63.0 |
| GeoThinker-8B | 68.2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| SpatialStack-4B | 72.0 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| SpatialSpeak-4B (Ours) | **76.0** | 68.0 | 89.2 | 84.6 | 39.4 | 34.0 | 87.2 | 86.3 | 71.1 | 52.2 | 75.0 | 86.8 | 81.8 | 56.6 | 83.3 | 89.1 | 90.2 | 93.5 | 86.8 | 88.1 | 76.9 | 79.7 | 64.9 | 81.0 |

### A.2 More qualitative examples

In this section, we show additional reasoning examples in Fig. 5 and Fig. 6, together with reconstruction examples in Fig. 7. In Fig. 5, the model uses estimated object centers and dimensions to derive an approximate closest-point distance between a trash bin and a toilet. In Fig. 6, it enumerates four chair instances with distinct estimated 3D centers and predicts a count of 4, matching the ground truth, whereas the variant with neither QA-RP nor CoT-VC predicts 6. Fig. 7 extends the reconstruction comparison in Fig. 4 with enlarged visualizations of the same scene and a comparison on an additional scene. For both illustrated scenes, the distances marked in SpatialSpeak's Stage I reconstructions are closer to ground truth than those of CUT3R and MapAnything.

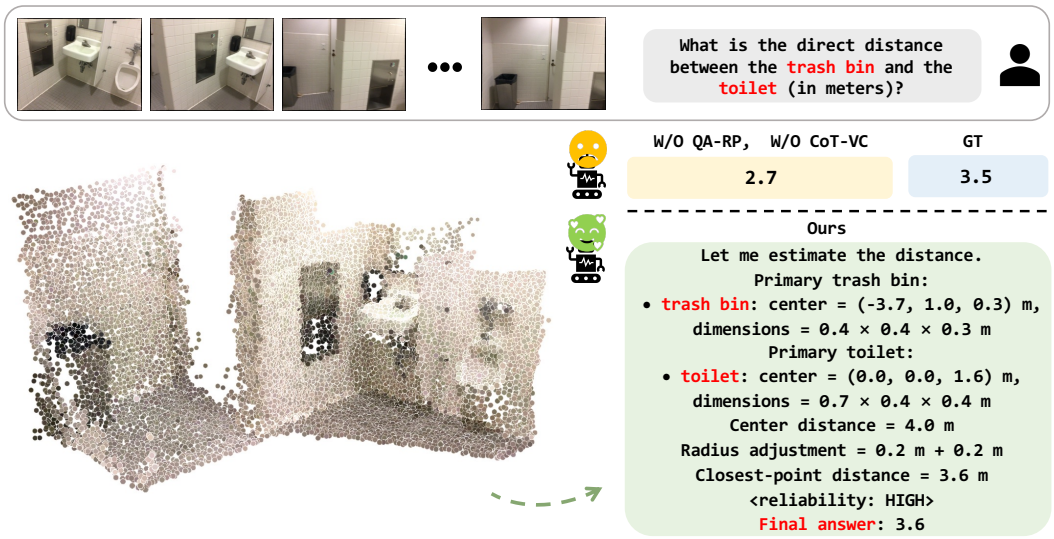


Figure 5: **Qualitative example**. The bottom-left panel shows the Stage I (QA-RP) reconstruction, and the bottom-right panel shows the Stage II (CoT-VC) reasoning output. SpatialSpeak estimates the 3D centers and dimensions of the trash bin and toilet, then approximates their closest-point distance as 3.6 m, compared with 2.7 m from the variant with neither QA-RP nor CoT-VC and a ground truth of 3.5 m.

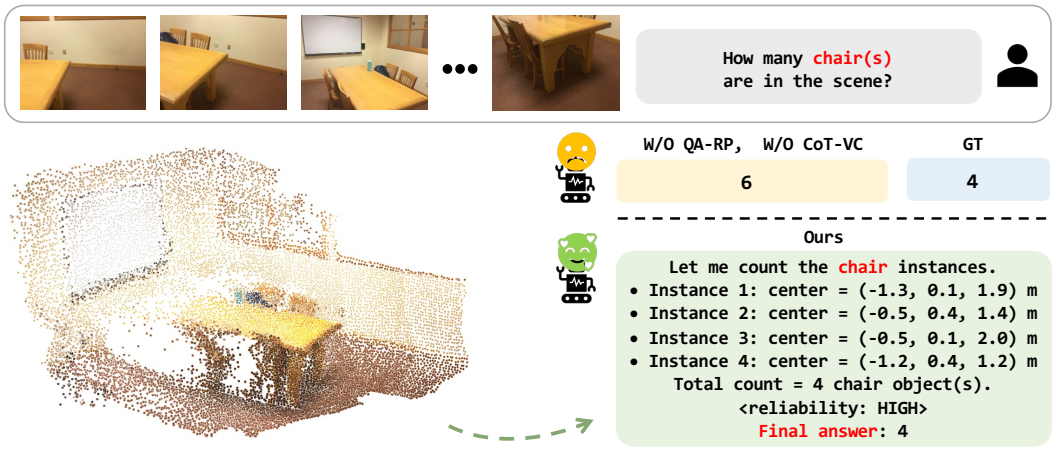


Figure 6: **Qualitative example**. The bottom-left panel shows the Stage I (QA-RP) reconstruction, and the bottom-right panel shows the Stage II (CoT-VC) response. SpatialSpeak enumerates four chair instances with distinct estimated 3D centers and predicts a count of 4, matching the ground truth, whereas the variant with neither QA-RP nor CoT-VC predicts 6.

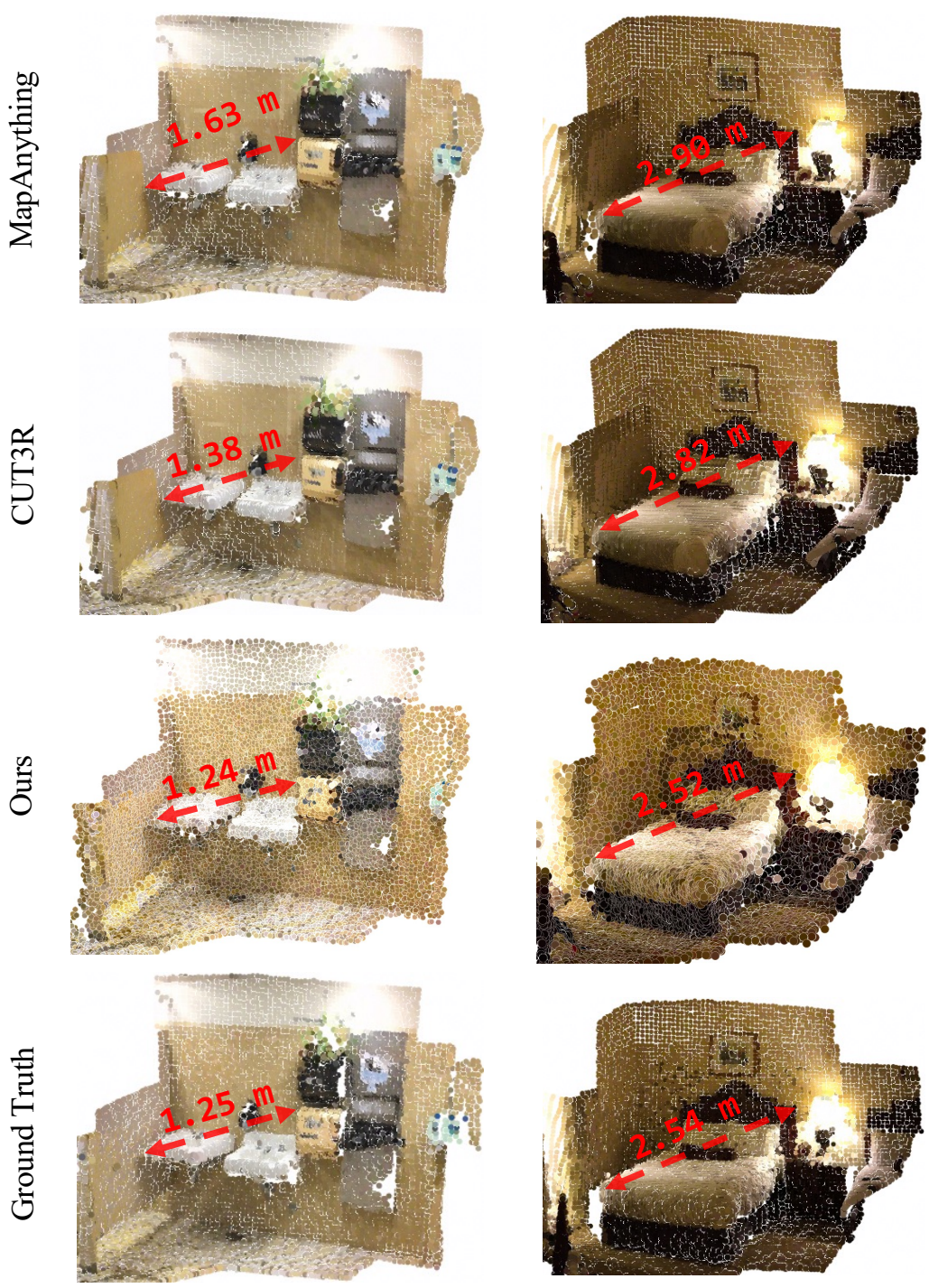


Figure 7: **Extended qualitative reconstruction comparisons**. Rows show MapAnything (Keetha et al., 2026), CUT3R (Wang et al., 2025b), SpatialSpeak after Stage I, and ground truth from top to bottom. The left column provides enlarged visualizations of the scene in Fig. 4, while the right column shows an additional scene. Red dashed arrows mark corresponding distances in meters. SpatialSpeak's marked distances (1.24/2.52 m, left/right) are closer to the ground-truth values (1.25/2.54 m) than those of CUT3R (1.38/2.82 m) and MapAnything (1.63/2.90 m).

### A.3 MORE IMPLEMENTATION DETAILS

In this section, we present the training hyperparameters. The configurations for Stage I and Stage II are summarized in Tab. 10 and Tab. 11, respectively.

Table 10: **Training hyperparameters in Training Stage I**.

| Hyperparameter | Value |
|---|---|
| accelerator model | H800 GPUs |
| accelerator count | 8 |
| global batch size | 32 |
| tune_mm_llm | True |
| tune_mm_vision | False |
| tune_mm_mlp | False |
| bf16 | True |
| per_device_train_batch_size | 1 |
| gradient_accumulation_steps | 4 |
| learning_rate | 5e-6 |
| optim | adamw_torch |
| num_train_epochs | 1 |
| warmup_ratio | 0.03 |
| lr_scheduler_type | 'cosine' |
| weight_decay | 0.01 |

Table 11: **Training hyperparameters in Training Stage II**.

| Hyperparameter | Value |
|---|---|
| accelerator model | H800 GPUs |
| accelerator count | 8 |
| global batch size | 64 |
| tune_mm_llm | True |
| tune_mm_vision | False |
| tune_mm_mlp | False |
| bf16 | True |
| per_device_train_batch_size | 1 |
| gradient_accumulation_steps | 8 |
| learning_rate | 1e-5 |
| optim | adamw_torch |
| num_train_epochs | 1 |
| warmup_ratio | 0.03 |
| lr_scheduler_type | 'cosine' |
| weight_decay | 0.01 |