DESPATIAL: THE DEVIL OF AGENTIC SPATIAL REASONING LIES IN 3D ITSELF

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial agents have access to reconstruction and segmentation tools, yet inconsistent predictions and fragmented object identities can prevent their outputs from becoming reliable 3D scene information. A simple oracle experiment exposes this perception bottleneck: replacing tool-predicted perception with ground-truth 3D bounding boxes raises an agent to nearly full score across five object-centric VSI-Bench tasks, while retaining the same reasoning tools. Motivated by this finding, we believe the devil of spatial reasoning lies in 3D itself, and introduce DeSpatial(**De**vil of **Spatial**), which explicitly decomposes spatial understanding into perception and reasoning. DeSpatial has a shared 3D perception tool to reconcile independent perception outputs into a unified, metric, gravity-aligned 3D scene. This perception tool has two core components: a) a *joint segmentation–depth co-refinement* module that consolidates masks of the same object and removes projection artifacts caused by inconsistent segmentation and depth within a single frame; b) a *cross-view instance fusion* module that consolidate voxel-overlapping instances but not allowed to merge same-frame instances. This fusion module additionally fuses spatially disjoint fragments caused by insufficient multi-view observations, by checking their 3D structure and source mask appearance. After receiving the complete and unified 3D scene information, the agent chooses reasoning or computational tools to solve the spatial problem. DeSpatial achieves leading performance on VSI-Bench, MindCube-Tiny, and ViewSpatial, with additional long-video results on VSI-SUPER, revealing the bottleneck of spatial understanding lies in a precise 3D perception tool.

1 INTRODUCTION

Spatial understanding requires a consistent account of a scene: how many objects it contains, how large they are, and where they lie relative to one another or to an observer. Videos make this problem particularly demanding because the same object can appear repeatedly, with different visible surfaces and intervening occluders. Recent benchmarks have exposed these challenges through various types of visual questions including object measurements, multi-view questions, and changes of perspective (Yang et al., 2025a; Wang et al., 2025b; Li et al., 2025a).

Spatial agents address these demands by combining vision-language models (VLMs) with perception tools, such as 3D reconstruction (Wang et al., 2025a), segmentation (Kirillov et al., 2023), and detection (Liu et al., 2025). Access to these tools, however, does not automatically provide reliable 3D scene understanding. Their independently predicted masks, depths, and reconstructed surfaces need to be reconciled into objects with consistent geometry and identity. Existing agents organize this evidence through visual programs, rendered views, or structured scene memories (Luo et al., 2026; Zhang et al., 2026b; Mou et al., 2026; He et al., 2026). Yet even a correct sequence of tool calls can leave an agent with conflicting evidence. For example, inconsistencies between a chair's 2D mask and independently predicted depth can introduce floor points into the chair's 3D point cloud during back-projection, as illustrated in Figure 2(b). Observations of different parts of the same table may also be treated as separate instances. The resulting errors propagate into otherwise valid counting, measurement, and relational computations. We therefore conjecture that effective agentic spatial understanding depends on both how to organize intermediate perception outputs to a 3D representation and how accurately this representation describes the scene.

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.

To understand this problem better, we do a simple oracle experiment that reveals how much the quality of this 3D representation matters. We use ground-truth (GT) annotated 3D bounding boxes as the outputs of a pseudo 3D perception tool and pair the agent with suitable spatial reasoning and computational tools to solve tasks. On VSI-Bench, the agent can achieve a surprising score of 98.23% across five object-centric tasks covering 4,030 questions, as shown in Table 1, while the performance drops greatly when replacing the oracle 3D tool with real perception tools. The comparison identifies perception as a major bottleneck in these tasks and motivates improving the 3D evidence available to the reasoner. Section 2 details the annotation interface and scope of this diagnostic.

Motivated by this observation, we introduce **DeSpatial**, a framework that explicitly decomposes spatial understanding into perception and reasoning. It shares the same 3D perception tool across all spatial tasks, taking responsibility for reconciling independent 2D perception and 3D reconstruction outputs. This tool returns a unified, metric, gravity-aligned scene representation containing object identities, oriented bounding boxes, and source observations. A separate set of reasoning and computation tools can be dynamically chosen to operate on this shared representation to answer the question. This division gives the agent a consistent set of objects and spatial attributes to work with, reducing the need to reconcile intermediate perception outputs during task reasoning.

Our perception design addresses two coupled problems. *Joint segmentation–depth co-refinement* first consolidates duplicate masks and identifies parts of an occluded object within each frame. It then combines semantic membership with geometric reliability to remove erroneous projections by using a sequence of geometric and component filters. *Cross-view instance fusion* establishes identity across frames. Voxel overlap provides positive association evidence, while distinct instances observed in the same frame prohibit incompatible merges. For cross-view observations of different object parts that fully disjoint in space, a visual judge combines multiple source views with their 3D arrangement to decide whether they belong to one physical object. These operations consolidate observed evidence into a scene that downstream tools can consistently measure and query.

Our contributions are threefold:

- We identify the accuracy of 3D perception as a major bottleneck in agentic spatial understanding through an oracle experiment: given GT 3D boxes, an agent achieves strong performance with the same reasoning and computational tools. This locates the "devil" in the accuracy of the underlying 3D information: improving perception can unlock the agent's existing reasoning capability without a more elaborate dynamic tool-calling mechanism.
- We develop a shared 3D perception tool that addresses inconsistencies between independently predicted segmentation and depth. Its cross-view fusion uses voxel-overlap matching to recover associations between partial observations, while same-frame exclusions guard against merging distinct objects, jointly targeting high recall and precision. Visual identity verification further identifies fragments of the same object from spatially non-overlapping observations.
- DeSpatial achieves leading agentic spatial-understanding performance on VSI-Bench, MindCube-Tiny, and ViewSpatial, and extends the evaluation to long-video spatial understanding on VSI-SUPER.

2 Diagnosing the Perception Bottleneck

We test how well a spatial agent can use reasoning tools when the object geometry needed for a question is supplied by ground truth annotations. At the perception–reasoning interface, we supply an oracle 3D tool that returns ground-truth 3D oriented bounding boxes (OBBs). These boxes are the only 3D information provided by the oracle tool, and the agent still requests object classes, selects instances, calls spatial computation tools, and derives the answer through ordinary tool calls by itself. The provided reasoning and computational tools are the same as in Section 3.4.

Evaluation setup. We compare the oracle 3D tool with a raw-perception baseline that directly lifts SAM3 (Carion et al., 2025) masks using Pi3 (Wang et al., 2025c) geometry, with geometric filtering and component selection disabled. Both conditions use Qwen3-VL-8B-Instruct, the same system prompt, tool registrations, iteration limits, and 32-frame sampling policy, and all reasoning

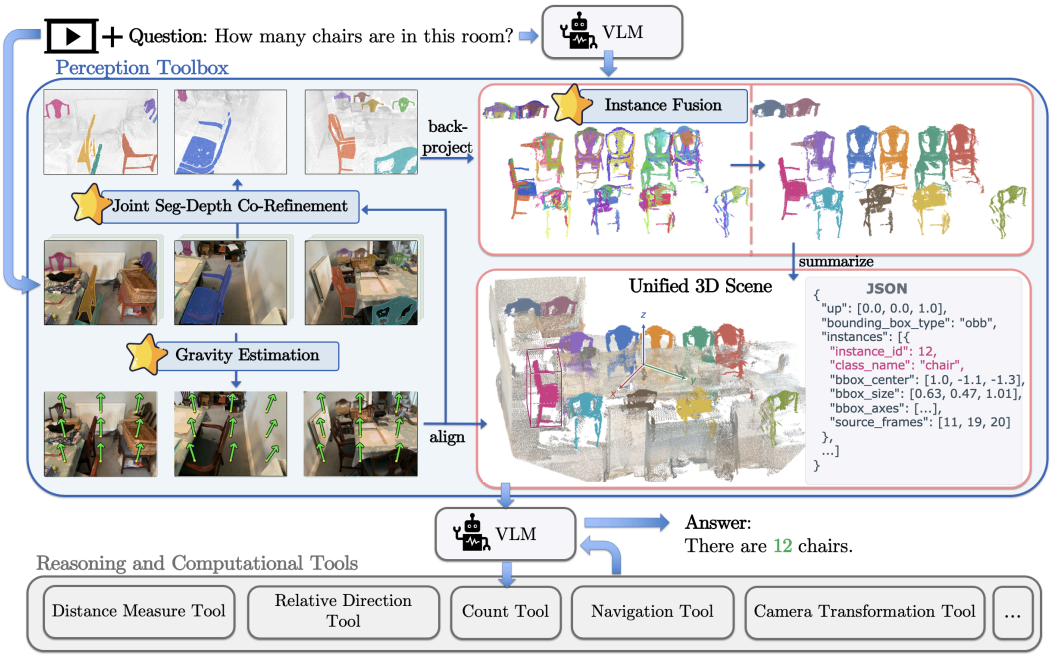

Figure 1: **Overview of DeSpatial.** The VLM receives the video and question and calls our shared perception tool. Joint segmentation–depth co-refinement cleans per-frame observations; cross-view instance fusion associates them into distinguished individual objects. Colors identify individual masks before fusion and persistent instances afterward. Gravity estimation establishes a common vertical direction. The VLM then reasons and answers using reasoning and computational tools.

Table 1: Oracle and raw-perception results on the same 4,030 questions across five VSI-Bench categories.

3D Tool	Obj. Count	Abs. Dist.	Obj. Size	Rel. Dist.	Rel. Dir.	Avg.
SAM3 + Pi3 (raw)	42.70	32.54	49.68	47.62	79.37	50.38
GT 3D Boxes	98.76	96.73	99.37	98.59	97.70	**98.23**

and computational tools are kept the same. Both the oracle and raw-perception conditions evaluate the same 4,030 questions across five VSI-Bench categories: object counting, absolute distance, object size, relative distance, and relative direction. Numerical questions use mean relative accuracy (MRA), and multiple-choice questions use accuracy. Appendix A details the experiment settings.

What the oracle reveals. Table 1 shows that with the oracle 3D tool, scores on all five tasks are close to perfect. This result should not be over-explained because the annotations in VSI-Bench are not perfect (Zhang et al., 2026a), and the annotations align with the benchmark's questions' answers. However, it suffices to indicate that given the proper 3D information, the agent's spatial reasoning ability is little constrained by how to utilize the spatial measurement, analysis, or relative direction tools, but rather identifies perception as a major bottleneck. It motivates improving the scene representation consumed by those tools, rather than simply providing the agent with a large list of perception and reasoning tools and expecting it to magically know how to fully reconcile the inconsistency between various perception results. Next, we introduce our framework focusing on improving the 3D perception ability inspired by this oracle experiment.

3 DeSpatial

DeSpatial separates spatial understanding into perception and reasoning (Figure 1). Its perception tool constructs a unified, gravity-aligned 3D scene through geometry estimation, joint segmentation–

depth co-refinement, and cross-view instance fusion; reasoning tools then operate on this representation.

3.1 Metric and Gravity-Aligned Geometry

We obtain dense geometry and camera poses with reconstruction models (Wang et al., 2025a;c; Lin et al., 2025), and rescale them to approximate metric scale using Depth Anything 3 (Lin et al., 2025). We uniformly sample up to 12 frames and estimate their camera-coordinate up directions with GeoCalib (Veicht et al., 2024). Let R_t denote the camera-to-world rotation of frame t, and let $\hat{\mathbf{u}}_t^c$ be its predicted unit up vector, opposite to gravity. We express these estimates in the common reconstruction coordinates as

$$\hat{\mathbf{u}}_t^w = R_t \hat{\mathbf{u}}_t^c. \tag{1}$$

To reject inconsistent estimates, we select the direction with the most neighbors within an angular threshold $\theta = 20°$:

$$\mathcal{I}(j) = \left\{ t : \arccos\left((\hat{\mathbf{u}}_t^w)^\top \hat{\mathbf{u}}_j^w \right) \le \theta \right\}, \qquad j^\star = \arg\max_j |\mathcal{I}(j)|. \tag{2}$$

Ties are resolved by the smaller median inlier angle. We use $\mathcal{I} = \mathcal{I}(j^\star)$ for joint calibration; if fewer than $\min(4, N)$ frames agree, where N is the number sampled, we fall back to all sampled frames and flag the estimate as less reliable.

GeoCalib then jointly calibrates the selected views with shared intrinsics and fixed camera rotations. Their refined camera-coordinate directions $\tilde{\mathbf{u}}_t^c$ are constrained by a single unit scene-up vector $\mathbf{u}_{\text{raw}}$:

$$\tilde{\mathbf{u}}_t^c = R_t^\top \mathbf{u}_{\text{raw}}, \qquad \|\mathbf{u}_{\text{raw}}\|_2 = 1, \qquad t \in \mathcal{I}. \tag{3}$$

We recover the shared direction by normalizing $\sum_{t\in\mathcal{I}} R_t \tilde{\mathbf{u}}_t^c$. Finally, we construct a rotation R such that $R\mathbf{u}_{\text{raw}} = \mathbf{e}_z$, and apply the same metric scale s and rotation to points and camera centers, $\mathbf{p}' = sR\mathbf{p}$ and $\mathbf{c}_t' = sR\mathbf{c}_t$; camera rotations become $R_t' = RR_t$. This places the scene in metric coordinates with $+z$ pointing upward, providing a common frame for reconciling the independently predicted masks and depths.

3.2 Joint Segmentation–Depth Co-Refinement

The module addresses inconsistencies within a single frame. It first consolidates same-class masks within each frame, then filters their geometric support (Figure 2). Consolidation addresses two sources of repeated object identities. For highly overlapping masks, we take their union when the intersection exceeds half the smaller mask's area. Occlusion can instead split one object into disjoint visible parts. We use their gravity-aligned horizontal voxel projections to identify overlapping candidates, then ask a VLM judge whether the marked parts in the source image belong to one physical object. The judge considers visible structure, the occluder, appearance, and scene context. Only a positive decision permits mask union. Resolving these identities before cross-view fusion is essential: otherwise, the same-frame exclusion constraint in the fusing stage would incorrectly keep parts of one object apart.

A consolidated mask can still include incorrect 3D points. Independent segmentation and depth predictions often disagree around object boundaries, assigning background geometry to foreground masks, as shown in the chair back region in Figure 2(b). Even a small contaminated region can inflate the estimated object extent and distort size or distance measurements. We therefore combine semantic membership with reconstruction confidence, depth-discontinuity checks, and 3D outlier filtering. For consolidated mask $M_{t,i}$, the geometry-filtered point set is

$$P_{t,i} = \{ sR\mathbf{p}_t(\mathbf{x}) \mid \mathbf{x} \in M_{t,i} \cap V_t \}, \tag{4}$$

where $\mathbf{p}_t$ is the reconstructed world-coordinate point map, sR applies metric scaling and gravity alignment, and V_t is the set of pixels passing the geometric filters detailed in Appendix B.2.

These filters remove many outliers but cannot reliably reject an entire coherent background patch admitted by a mask. In Figure 2(b), disagreement around the openings in a chair back leaves detached floor points even after geometry filtering. We therefore select connected components within each proposal. A mask comprising k visible parts, confirmed in the mask consolidation phase, retains its k largest 3D components; ordinary masks use $k = 1$. Duplicate-mask union alone does not increase k. Selecting k components allows multiple confirmed parts to remain while removing smaller detached contamination.

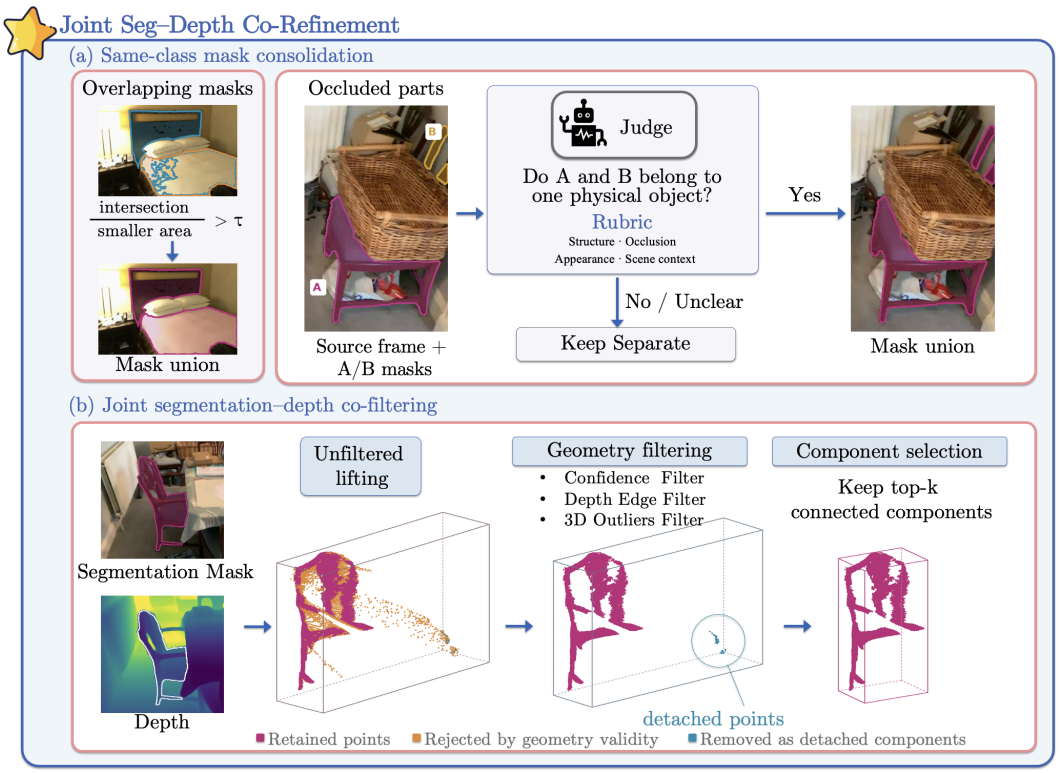

Figure 2: **Joint segmentation–depth co-refinement.** (a) Same-class mask consolidation unifies highly overlapping masks and visually verified parts of an occluded object. Negative or uncertain identity judgments keep masks separate. (b) Confidence, depth-edge, and 3D outlier filters remove unreliable geometry. Component selection then removes residual detached points that distort the bounding box, retaining the top-k components for k confirmed object parts, usually one.

3.3 Cross-View Instance Fusion

Clean per-frame observations still need persistent global object identities. One way is to use spatial clustering such as DBSCAN (Ruan et al., 2026) or centroid-distance clustering (Mou et al., 2026), but is prone to merging nearby objects. We also tried to merge based on 3D bounding-box IoU, but find it is sensitive to partial visibility and may wrongly connect objects whose bounding boxes overlap but geometry don't. We instead associate occupied voxels while preserving evidence that two observations represent different objects.

Each refined observation becomes a node carrying its class, source frame t_i, and point set. Let A_i and A_j be the occupied voxel sets of two same-class observations from different frames. We measure their overlap by

$$w_{ij} = \frac{|A_i \cap A_j|}{\min(|A_i|, |A_j|)}. \tag{5}$$

Normalizing by the smaller occupied set accommodates observations that cover different portions of an object. We retain candidate edges with $w_{ij} \geq \tau_{\text{ov}}$ and process them in descending score order.

To prevent distinct objects from merging, we impose *same-frame exclusion* on entire groups of observations. For a group G, let $\mathcal{F}(G) = \{t_i : i \in G\}$ be its set of source frames. An edge can join groups G_a and G_b only when

$$\mathcal{F}(G_a) \cap \mathcal{F}(G_b) = \emptyset. \tag{6}$$

Thus, distinct consolidated observations from one frame cannot enter the same instance, including through indirect chains of overlapping observations. In Figure 3(a), joining A with C blocks a later B–C merge because A and B were observed as distinct objects. Appendix B.3 gives the voxel size, overlap threshold, and implementation details.

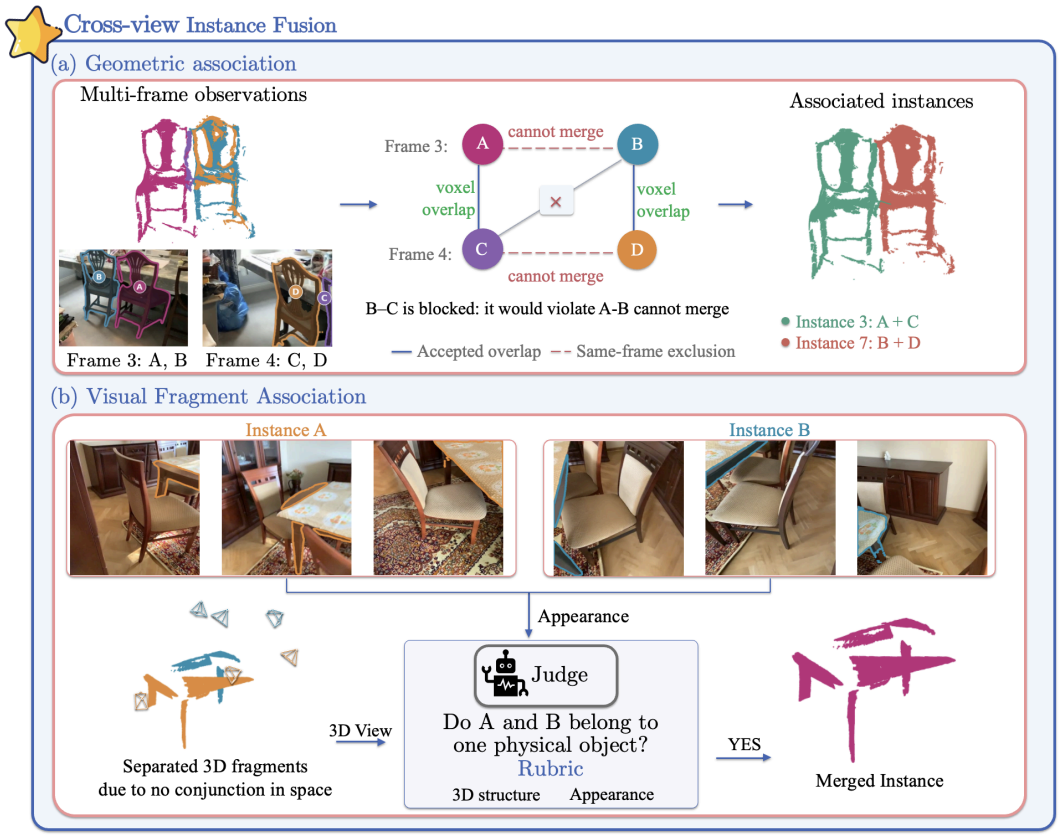

Figure 3: **Cross-view instance fusion.** (a) Voxel overlap links partial observations, while same-frame exclusions prevent distinct objects from merging, including through indirect connections. (b) Visual fragment association handles observations without shared voxels. Multiple source views and a common 3D view provide complementary appearance and structural evidence to an identity judge. A positive decision combines the observed fragments and their provenance.

Voxel overlap cannot connect all observations of one object. Different viewpoints may reveal separate portions of a surface, producing disjoint 3D fragments even without occlusion. After geometric fusion, we retrieve remaining same-class candidates using the union of neighbors in 3D surface distance and gravity-aligned horizontal projection distance. A VLM judge receives several masked source views for each candidate together with their 3D visualization in the common coordinate frame. Its criteria are *3D structure*—compatible position, orientation, and structural arrangement—and *appearance*—materials, textures, and corresponding visible details. The candidates are merged if they are judged to be the same object. Figure 3(b) illustrates this visual fragment association, and more details are in Appendix B.3.

The final scene packages these fused instances in one consistent representation. Each instance has a persistent identifier, semantic class, centroid, gravity-aligned oriented bounding box (center, side lengths, and axes), and source frame observations. Together with the scene up direction and camera geometry, these attributes let downstream tools refer to the same objects without independently reconciling masks, depths, or cross-view identities.

3.4 Reasoning with the Unified Scene

The VLM uses structured tool calls to count objects, measure dimensions and distances, determine relative directions, transform reference frames, plan routes, and identify appearance order from the shared scene representation and source observations. Appendix B.4 describes these operations and their reference-frame conventions.

Table 2: VSI-Bench results (%): MRA for numerical questions and accuracy for multiple choice. Avg. averages eight categories (four multiple-choice categories for Think3D); relative direction averages three difficulty levels. Unmarked methods use the full benchmark. *Subset evaluations from the original papers.

		Numerical (MRA)				Multiple choice (accuracy)				
Method	Backbone	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	Avg.
Trained models										
Cambrian-S	Qwen2.5-7B	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1	67.5
VLM-3R	LLaVA-NV-7B	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1	60.9
CaMo-3B	Qwen2.5-VL-3B	64.8	18.6	59.2	43.2	45.2	45.3	25.8	49.9	44.0
Training-free inference pipelines (Li et al., 2025c; Ma et al., 2026)										
See&Trek	InternVL3-8B	65.2	32.9	46.9	46.7	45.9	40.2	30.4	37.4	43.2
ReRe	InternVL3-8B	38.8	31.1	44.4	30.0	37.2	37.3	30.9	23.6	34.16
Spatial reasoning agents										
ViSRA	InternVL3-8B	36.6	34.6	58.8	52.7	47.3	62.4	46.9	59.7	49.88
AlloSpatial*	Gemini-3-Pro	51.8	36.8	57.7	62.6	62.0	67.7	65.3	83.7	61.0
Think3D*	Gemini-2.5-Pro	–	–	–	–	54.00	37.30	46.93	68.24	51.61
DeSpatial	Qwen3-VL-8B	65.43	55.08	66.86	73.58	67.93	82.53	64.43	77.90	69.22

4 EXPERIMENTS

4.1 EVALUATION SETUP

Qwen3-VL-8B-Instruct serves as the agent backbone. We evaluate video-based spatial understanding on VSI-Bench, multi-view mental modeling on MindCube-Tiny, and perspective-dependent reasoning on ViewSpatial (Yang et al., 2025a; Wang et al., 2025b; Li et al., 2025a). VSI-SUPER further tests spatial recall and continual counting over long videos (Yang et al., 2025b).

4.2 MAIN RESULTS

DeSpatial achieves an average score of 69.22 on VSI-Bench (Table 2), demonstrating that its shared scene representation supports diverse spatial computations. It obtains 87.19 overall accuracy on MindCube-Tiny and 68.35 on ViewSpatial (Table 3), extending its performance to multi-view reasoning and changes of perspective. On VSI-SUPER (Table 4), DeSpatial maintains high recall across increasing video durations and improves continual counting at longer horizons. These benchmarks extend the evaluation beyond object measurements to changes of viewpoint and observations accumulated over long videos. Together, these results show the utility of consistent scene information across object-centric, viewpoint-dependent, and long-video spatial reasoning.

4.3 ABLATION STUDY

On a fixed 504-question VSI-Bench subset (Table 5), co-refinement improves Avg by 11.27 points over raw lifting, demonstrating the combined benefit of geometric cleanup, mask consolidation, and part-aware component selection. Geometric filtering removes unreliable local contributions, while component selection removes coherent detached geometry that survives those filters. Consolidation and part-aware retention address a complementary risk: valid parts of one object should remain available to downstream computation. Voxel-based fusion outperforms AABB IoU; same-frame exclusion guards against false merges, and visual fragment association further improves results by connecting spatially disjoint observations. Here, voxel overlap supports associating partial observations, whereas the exclusion constraint prevents those links from collapsing distinct objects observed together. Visual association handles observations that geometric overlap alone cannot connect. For gravity estimation, we visualize GeoCalib joint calibration and a RANSAC plane-based estimator on the same 156 scenes for human assessment (Table 6). GeoCalib improves visually judged correctness by 31.41 percentage points.

Table 3: Results on MindCube-Tiny (1,050 questions) and ViewSpatial (5,712) (Wang et al., 2025b; Li et al., 2025a; Cai et al., 2025; Dai et al., 2026; Xu et al., 2026; Huang et al., 2026; Li et al., 2025b). Scores are accuracy percentages; Avg. denotes overall accuracy. *Evaluation on Think3D's MindCube-1K split.

(a) MindCube

Method	Among	Around	Rotation	Avg.
Qwen2.5-VL-3B	36.00	45.20	34.00	37.81
GPT-4o	43.00	46.40	35.00	42.29
Gemini-2.5-Pro	43.80	67.20	88.50	57.90
Gemini-3-Pro	68.20	77.20	93.00	75.10
$\text{VADAR}_{\text{GPT-4o}}$	40.67	46.80	33.50	40.76
$\text{pySpatial}_{\text{GPT-4o}}$	66.33	71.20	41.00	62.67
$\text{GCA}_{\text{Qwen3-VL-Thinking}}$	59.80	61.80	82.00	64.20
$\text{AlloSpatial}_{\text{Gemini-3-Pro}}$	75.80	86.00	93.50	81.60
$\text{Think3D}_{\text{Gemini-2.5-Pro}}$	49.67	62.75	87.00	59.97*
$\text{DeSpatial}_{\text{Qwen3-VL-8B}}$	85.08	86.80	94.00	87.19

(b) ViewSpatial

	Camera		Person			
Method	Rel. Dir.	Obj. Ori.	Rel. Dir.	Obj. Ori.	Scene Sim.	Avg.
Gemini-2.5-Pro	54.30	24.30	47.27	59.84	39.28	46.09
Qwen3-VL-8B	54.20	29.70	40.30	47.30	31.10	42.20
Qwen2.5-VL-7B	46.64	29.72	35.04	37.05	28.78	36.85
InternVL2.5-8B	49.41	41.27	42.04	46.79	32.85	43.24
$\text{SenseNova-SI}_{\text{InternVL3-8B}}$	66.30	43.20	43.10	38.40	70.00	54.70
$\text{S-Agent}_{\text{GPT-5.4}}$	62.50	55.50	81.10	42.20	60.60	60.00
$\text{MV-STRIDE}_{\text{Qwen3-VL-8B}}$	62.55	33.13	39.43	48.09	46.33	48.35
$\text{CaMo}_{\text{Qwen2.5-VL-3B}}$	47.50	30.90	31.50	53.90	35.10	41.00
$\text{SpatialLadder}_{\text{Qwen2.5-VL-3B}}$	48.30	24.10	34.40	71.10	38.90	44.20
$\text{DeSpatial}_{\text{Qwen3-VL-8B}}$	65.71	81.22	69.00	78.01	51.76	68.35

Table 4: Results on VSI-SUPER (Yang et al., 2025b). Recall (VSR) uses accuracy and Count (VSC) uses mean relative accuracy (MRA), both in percent; durations are in minutes. Published values follow Dai et al. (2026).

	Recall (VSR): accuracy					Count (VSC): MRA			
Method	10	30	60	120	240	10	30	60	120
MovieChat	18.3	21.7	16.7	26.7	25.6	0.0	0.0	0.0	0.0
Flash-VStream	28.3	33.3	23.3	28.3	31.7	0.0	0.0	0.0	0.0
Cambrian-S-7B	38.3	35.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0
Cambrian-S-7B-LFP	45.0	41.7	40.0	40.0	40.0	40.6	42.0	35.0	34.0
S-Agent	75.0	55.0	63.3	66.1	77.2	10.6	4.2	0.0	0.0
DeSpatial	90.3	88.3	90.1	88.3	86.7	33.4	53.4	49.4	47.2

5 Related Work

Spatial understanding with VLMs requires recovering object properties, spatial relationships, and changes of viewpoint from visual observations. VSI-Bench evaluates these abilities through video-based spatial questions, MindCube examines mental modeling from limited views, and ViewSpatial studies reasoning across camera and human perspectives (Yang et al., 2025a; Wang et al., 2025b; Li et al., 2025a). Existing non-agentic approaches improve either the learned visual representations or the spatial evidence available during inference. Cambrian-S strengthens video representations for spatial understanding, while VLM-3R aligns reconstructive representations with language instructions to incorporate geometric information into VLMs (Yang et al., 2025b; Fan et al., 2026). Complementary approaches introduce spatial cues without updating model parameters. See&Trek

Table 5: Ablations on a fixed stratified VSI-Bench subset (504 questions, 63 per category); Avg. averages eight categories. (a) Full cross-view fusion is fixed. (b) Full co-refinement is fixed; Exclusion: same-frame exclusion; FragAssoc.: visual fragment association. All runs share the agent settings. IoU/voxel thresholds: 0.3/0.1; voxel size: 0.1 m.

(a) Joint seg–depth co-refinement

Configuration	Avg (%)
Raw mask lifting	57.56
+ Geometry filtering	62.90
+ Largest component ($k = 1$)	65.22
+ Same-class mask consolidation	66.45
+ Part-aware top-k selection	68.83

(b) Cross-view instance fusion

Overlap	Exclusion	FragAssoc.	Avg (%)
AABB IoU	Yes	No	59.90
DBSCAN	Yes	No	61.37
Voxel overlap	No	No	64.30
Voxel overlap	Yes	No	66.82
Voxel overlap	Yes	Yes	68.83

Table 6: Human assessment of projected up directions on the same 156 scenes.

Gravity estimator	Correct / Total	Correctness (%)
RANSAC plane-based	107 / 156	68.59
GeoCalib joint calibration	156 / 156	**100.00**

supplies spatial prompting cues, and ReRe revisits initial answers using synthesized complementary views (Li et al., 2025c; Ma et al., 2026). These methods demonstrate different ways to expose spatial structure that may be difficult to infer from the original observations alone. Their emphasis ranges from learning geometry-aware representations to improving the evidence presented at inference time. Agentic approaches extend this direction by making perception and spatial computation available as executable operations, enabling models to gather and process evidence over multiple steps.

Agentic spatial reasoning augments VLMs with specialist perception and geometric computation tools. Beyond selecting an operation, an agent must compose these tools and interpret intermediate outputs such as images, point clouds, camera poses, and object geometry. Existing methods approach this process through program generation, interactive exploration, geometric constraints, and reusable skills. VADAR synthesizes a dynamic Python API, while pySpatial generates programs for scene reconstruction, camera manipulation, and novel-view rendering (Marsili et al., 2025; Luo et al., 2026). Think3D supports interactive exploration of reconstructed 3D space, allowing the agent to gather visual evidence from additional viewpoints (Zhang et al., 2026b). GCA formalizes reference frames and geometric objectives to guide tool selection and computation (Chen et al., 2025). Skill-3D distills scene-aware skills from execution trajectories, whereas NeSy-Spatial evolves reusable tool-use and geometric skills (Li et al., 2026; Tian et al., 2026). SpatialCLI learns tool-use policies and internalizes capabilities acquired through tool interaction (Zhou et al., 2026).

Other agents emphasize organizing spatial evidence for downstream reasoning. MSSR curates a minimal sufficient set through interacting perception and reasoning agents (Guo et al., 2025), while AlloSpatial constructs allocentric landmark and route maps from reconstructions filtered by confidence and point density (Ruan et al., 2026). These approaches highlight the importance of making perception outputs usable by an agent. DeSpatial focuses on the accuracy and consistency of the underlying 3D information: independent segmentation and reconstruction outputs can disagree geometrically, and observations of one object can remain fragmented within or across views. Our perception tool resolves these issues before returning a unified scene representation. We retain a direct structured tool-calling interface for downstream reasoning; the oracle diagnostic examines its capacity when predicted object information is replaced with annotated geometry.

6 Conclusion

DeSpatial treats reliable 3D scene construction as a central requirement for spatial agents. Its oracle study shows that an agent with the same reasoning tools can solve object-centric spatial questions substantially more accurately when given GT 3D bounding boxes. This motivates a shared perception tool that resolves inconsistencies between segmentation and geometry, consolidates object observations within and across frames, and returns unified scene information for subsequent computation. Joint segmentation–depth co-refinement and cross-view instance fusion address geometric contamination and fragmented identity together. The resulting framework places the emphasis on the quality and consistency of the information over which an agent reasons.

Limitations. The representation depends on detections, reconstructed geometry, metric scale, and gravity estimates. Component selection cannot recover unobserved surfaces or missing detections, and incorrect part counts or visual identity judgments can discard valid geometry or merge different objects. The oracle results concern five object-centric categories and answer-consistent annotations; they do not establish that perception is the only limitation across all spatial tasks. The component ablations cover a 504-question VSI-Bench subset.

AI USE STATEMENT

We used generative AI tools for some of the experiment execution, including feedback and implementation of code and interpretation of results, for finding related work, and for drafting and polishing the text. We did not use them to generate data or to explore specific ideas. All AI-assisted content was reviewed by the authors, who take responsibility for the final content of this work.

REFERENCES

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, 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, 2025. URL https://arxiv.org/abs/2511.13719v4.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning. *arXiv preprint arXiv:2511.22659*, 2025. URL https://arxiv.org/abs/2511.22659.

Yalun Dai, Hao Li, Shulin Tian, Runmao Yao, Yuhao Dong, Fangzhou Hong, Zhaoxi Chen, Fangfu Liu, Tao Wang, Kim-Hui Yap, and Ziwei Liu. S-Agent: Spatial tool-use elicits reasoning for spatial intelligence, 2026. URL https://arxiv.org/abs/2606.20515v3.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Fan_VLM-3R_Vision-Language_Models_Augmented_with_Instruction-Aligned_3D_Reconstruction_CVPR_2026_paper.pdf.

Yejie Guo, Yunzhong Hou, Wufei Ma, Meng Tang, and Ming-Hsuan Yang. Pursuing minimal sufficiency in spatial reasoning. *arXiv preprint arXiv:2510.16688*, 2025. URL https://arxiv.org/abs/2510.16688.

Jixuan He, Xueting Li, Chieh Hubert Lin, and Ming-Hsuan Yang. Reasmory: 3D reconstruction as explicit memory for VLMs spatial reasoning, 2026. URL https://arxiv.org/abs/2606.00963.

Hsiang-Wei Huang, Junbin Lu, Kuang-Ming Chen, Jianxu Shangguan, Cheng-Yen Yang, and Jenq-Neng Hwang. CaMo: Camera motion grounded evaluation and training for vision-language models, 2026. URL https://arxiv.org/abs/2605.20165.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating multi-perspective spatial localization in vision-language models, 2025a. URL https://arxiv.org/abs/2505.21500v2.

Haoyuan Li, Zhengdong Hu, Jun Wang, Hehe Fan, and Yi Yang. Skill-3d: Evolving scene-aware skills for agentic 3d spatial reasoning, 2026. URL https://arxiv.org/abs/2606.07436.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models, 2025b. URL https://arxiv.org/abs/2510.08531.

Pengteng Li, Pinhao Song, Wuyang Li, Huizai Yao, Weiyu Guo, Yijie Xu, Dugang Liu, and Hui Xiong. See&trek: Training-free spatial prompting for multimodal large language model. In *Advances in Neural Information Processing Systems*, volume 38, 2025c. URL https://papers.neurips.cc/paper_files/paper/2025/hash/59a9cc95f046e9125d8816ef971873e7-Abstract-Conference.html.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. URL https://arxiv.org/abs/2511.10647.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pp. 38–55, Cham, 2025. Springer Nature Switzerland. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

Zhanpeng Luo, Ce Zhang, Silong Yong, Cunxi Dai, Qianwei Wang, Haoxi Ran, Guanya Shi, Katia Sycara, and Yaqi Xie. pySpatial: Generating 3d visual programs for zero-shot spatial reasoning. *arXiv preprint arXiv:2603.00905*, 2026. URL https://arxiv.org/abs/2603.00905.

Chaofan Ma, Zhenjie Mao, Yuhuan Yang, Fanqin Zeng, Yue Shi, Yingjie Zhou, Xiaofeng Cao, and Jiangchao Yao. Reason, then re-reason: Cross-view revisiting improves spatial reasoning. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.11683.

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic AI for spatial reasoning with a dynamic API. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19446–19455, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Marsili_Visual_Agentic_AI_for_Spatial_Reasoning_with_a_Dynamic_API_CVPR_2025_paper.html.

Tingshu Mou, Jiabo He, Renying Wang, Ce Liu, Hao Yang, Tiehua Zhang, Jingjing Chen, and Xingjun Ma. ViSRA: A video-based spatial reasoning agent for multi-modal large language models, 2026. URL https://arxiv.org/abs/2605.10106.

Shouwei Ruan, Bin Wang, Zhenyu Wu, Qihui Zhu, Yuxiang Zhang, Jingzhi Li, Yubin Wang, and Xingxing Wei. AlloSpatial: Agentic harness framework for spatial reasoning in foundation models, 2026. URL https://arxiv.org/abs/2606.08952.

Shi-Yu Tian, Zhuo-Xia Wang, Xuan-Yi Zhu, Zhi Zhou, Xinwei Yang, Kun-Yang Yu, Ming Yang, Yang Chen, and Yu-Feng Li. Self-evolving neuro-symbolic skills for tool-augmented spatial reasoning, 2026. URL https://arxiv.org/abs/2608.07955.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Single-image calibration with geometric optimization. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2409.06704.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025a. URL

https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Fei-Fei Li, and Manling Li. MindCube: Spatial mental modeling from limited views, 2025b. URL https://arxiv.org/abs/2506.21458v2.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025c. URL https://arxiv.org/abs/2507.13347.

Jin Xu, Xiaojian Huang, Zhuodong Luo, Zhihong Zhang, Xin Liu, Jiansheng Wei, Xinzhi Wang, Jie Zhao, and Xuejin Chen. MV-STRIDE: Enabling MLLMs to master multi-view spatial reasoning via hierarchical capability modeling, 2026. URL https://arxiv.org/abs/2609.07258v1.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a. URL https://arxiv.org/abs/2412.14171.

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

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, 2026a. URL https://arxiv.org/abs/2604.24300.

Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Yizhuang Peng, Zhenfei Yin, Lijun Wang, and Huchuan Lu. Think3D: Thinking with space for spatial reasoning, 2026b. URL https://arxiv.org/abs/2601.13029.

Yang Zhou, Zixuan Huang, Sunzhu Li, Zhuo Yang, Chen Zhang, Shunian Chen, Caijun Yan, Jianyao Xu, Shunyu Liu, Weijie Fu, Peiliang Li, Xiaozhi Chen, and Yuxiang Cai. SpatialCLI: Learning to reason with spatial tools, then without them. *arXiv preprint arXiv:2607.27703*, 2026. URL https://arxiv.org/abs/2607.27703.

A ORACLE PROTOCOL

Information supplied. The oracle replaces the perception module's object output with annotated semantic classes and 3D OBBs. Each box supplies a center, side lengths, and orientation axes in metric coordinates with positive-z up. These boxes are the only ground-truth 3D representation supplied to the agent. Object counts are computed from the returned instances rather than read from a separate count annotation. The adapter supplies no question answers, question categories, ground-truth target-instance selections, dataset camera poses, or temporal annotations. A question identifier retrieves its scene; it does not select the objects or computation needed to answer that question. The agent retains responsibility for requesting classes, choosing instances and tools, and producing its answer.

Controlled inference and geometric adaptation. Both conditions use the same 32-frame policy, Qwen3-VL-8B-Instruct, system prompt, continuation instruction, model-visible tool schemas, decoding settings, and iteration limits. The oracle common-policy audit passed for all 4,030 selected questions, and all are retained in scoring. Two questions did not request the scene tool and remain included. The reasoning tool set is unchanged. Local-frame computation is unchanged.

Scope of the intervention. The oracle response lists the available annotated classes and explains the coordinate convention and annotation coverage. It cannot provide source-frame correspondences, appearance descriptions, or detection confidence, so it omits confidence-based default-selection fields. Room, navigation, and temporal operations return a uniform unsupported response while their schemas remain visible to every question. These differences mean the experiment diagnoses access to annotated object information, including identity and coverage, rather than changing only box coordinates. Annotation errors documented by ReVSI (Zhang et al., 2026a) further mean that oracle accuracy measures compatibility with annotation-derived answers, not perfect recovery of physical geometry. The five evaluated categories and their scores are given in Table 1; room size, appearance order, and route planning are outside this diagnostic.

Both conditions evaluate the same 4,030 questions: 565 object-counting, 834 absolute-distance, 953 object-size, 710 relative-distance, and 968 relative-direction questions. Relative direction is averaged across its three official difficulty levels before computing the unweighted five-category average in Table 1. This diagnostic average is distinct from the eight-category average in Table 2.

B Perception and Spatial Computation Details

B.1 Metric Geometry and Co-refinement

Metric scale is estimated using robust depth-ratio statistics within frames, aggregated over reliable frames. Camera rotations transform the per-frame gravity predictions into reconstruction coordinates; cross-frame agreement removes inconsistent estimates before estimating a common up direction. If $\mathbf{u}_{\text{raw}}$ is this direction, we choose R satisfying $R\mathbf{u}_{\text{raw}} = \mathbf{e}_z$ and transform points by sR. Camera geometry receives the same transformation, and the resulting scene uses $\mathbf{u} = \mathbf{e}_z$.

For same-class masks M_i and M_j in one frame, duplicate consolidation uses

$$\rho(M_i, M_j) = \frac{|M_i \cap M_j|}{\min(|M_i|, |M_j|)} > 0.5. \tag{7}$$

Masks satisfying this rule are united by pixelwise OR. For disjoint masks that may be parts of an occluded object, we retrieve same-class candidates using overlap between their gravity-aligned horizontal voxel projections. The judge then sees the source image with the candidate parts marked A/B and considers their structure, occlusion, appearance, and surrounding context. A positive judgment combines the masks and records the number of confirmed visible parts. Duplicate masks covering the same part do not increase this count.

After mask consolidation, the geometry-validity mask combines reconstruction confidence, depth-discontinuity checks, and global point-outlier filtering. Intersecting it with the consolidated mask yields the merged proposal's point candidates. Component selection is applied to this merged point set and retains the k largest connected 3D components for k confirmed parts, or all available components if fewer than k remain. Ordinary masks use $k = 1$. This choice preserves validated disconnected parts; it cannot distinguish arbitrary foreground and background components of equal size, nor recover points rejected by preceding filters.

B.2 Geometry Filtering

The geometric validity set V_t in Equation 4 records which pixels in frame t pass reconstruction-confidence, depth-edge, and 3D outlier filtering. Let Ω_t be the image grid, $D_t(\mathbf{x})$ the reconstructed depth in camera coordinates, and $\mathbf{p}_t(\mathbf{x})$ the corresponding point in the common reconstruction coordinates. The thresholds below are the defaults of the Pi3-based perception implementation.

Reconstruction-confidence filtering. For the confidence logit $\ell_t(\mathbf{x})$ predicted by Pi3, we retain pixels whose sigmoid confidence exceeds τ_c:

$$V_t^{\text{conf}} = \{\mathbf{x} \in \Omega_t : \sigma(\ell_t(\mathbf{x})) > \tau_c\}, \qquad \tau_c = 0.08. \tag{8}$$

This confidence describes the reconstructed geometry; segmentation-mask confidence is not used in this test.

Depth-edge filtering. Let $\mathcal{N}_3(\mathbf{x})$ denote the 3×3 neighborhood centered at $\mathbf{x}$, restricted to the image grid. We compute the local depth range relative to the center pixel's depth and reject pixels exceeding τ_d:

$$e_t(\mathbf{x}) = \frac{\max_{\mathbf{y} \in \mathcal{N}_3(\mathbf{x})} D_t(\mathbf{y}) - \min_{\mathbf{y} \in \mathcal{N}_3(\mathbf{x})} D_t(\mathbf{y})}{D_t(\mathbf{x})},$$
$$V_t^{\text{depth}} = \{\mathbf{x} \in \Omega_t : e_t(\mathbf{x}) \le \tau_d\}, \qquad \tau_d = 0.02. \tag{9}$$

This removes pixels near sharp depth transitions, where foreground and background projections can mix. The depth is taken from the reconstruction model's camera-coordinate point map.

Global 3D outlier filtering. We pool the reconstructed points passing both preceding filters across all frames into $\mathcal{Q} = \{\mathbf{q}_i\}_{i=1}^{L}$, retaining one sample per surviving pixel. From this pool, we estimate

the empirical mean $\boldsymbol{\mu}$ and covariance Σ, and calculate squared Mahalanobis distances:

$$a_i = (\mathbf{q}_i - \boldsymbol{\mu})^\top \Sigma^\dagger (\mathbf{q}_i - \boldsymbol{\mu}),$$
$$\bar{a} = \frac{1}{L}\sum_{i=1}^{L} a_i, \qquad s_a = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(a_i - \bar{a})^2}. \tag{10}$$

Here $\Sigma^\dagger$ denotes the covariance pseudoinverse, equal to its inverse when nonsingular. We retain points satisfying $a_i < \bar{a} + \lambda s_a$, with $\lambda = 5$. Thus, with $a(\mathbf{p}) = (\mathbf{p} - \boldsymbol{\mu})^\top \Sigma^\dagger (\mathbf{p} - \boldsymbol{\mu})$, the final valid-pixel set is

$$V_t = \left\{ \mathbf{x} \in V_t^{\text{conf}} \cap V_t^{\text{depth}} : a(\mathbf{p}_t(\mathbf{x})) < \bar{a} + 5 s_a \right\}. \tag{11}$$

The mean, covariance, and outlier threshold are shared across the scene rather than estimated separately for each object. Intersecting V_t with the consolidated instance mask $M_{t,i}$ selects that instance's geometrically filtered points. A coherent patch of background points can still pass these tests; the subsequent instance-level component selection addresses such residual contamination.

B.3 Cross-view Instance Fusion

Geometric association settings. For the overlap score in Equation 5, the base configuration uses 0.1 m voxels and $\tau_{\text{ov}} = 0.1$. Union–find maintains each group's source-frame set to implement Equation 6; an accepted merge updates that set by union. Small-instance filtering follows geometric fusion and precedes visual fragment association; the latter does not restore discarded instances.

Fragment candidate retrieval. For two sampled point sets, we calculate each point's nearest-neighbor distance to the other set in both directions. We use the smaller of the two directional fifth-percentile distances as a robust surface-proximity measure. The same computation on the gravity-aligned horizontal projections provides a second retrieval channel. Each fragment contributes its top-three same-class neighbors from each channel, and their union defines the candidate pairs.

Judge inputs and group verification. Each candidate pair is represented by up to three masked source views per instance, together with oblique and top-down renderings of the pair. Available source-camera poses are included in the 3D visualization. The fragments remain fixed in their shared coordinate frame: neither is independently translated, rotated, or rescaled to improve their apparent structural fit. Mask and point-cloud colors identify the candidates and are not evidence of their true appearance. The judge returns `same`, `different`, or `unclear` using the structure and appearance criteria described in Section 3.3.

Group verification requires explicit `same` decisions for every pair across two groups before joining them. Thus A–B and B–C positive decisions do not establish A–C automatically; a missing, negative, or uncertain judgment blocks that group merge. The source-frame constraint in Equation 6 continues to apply.

B.4 Reasoning Tools

The tool suite counts instances, measures object and room dimensions, computes inter-object distances from OBB geometry, determines spatial relations, plans routes, and locates object appearance times using source observations. The VLM selects objects and operations through structured tool calls. If a query requires object facing or an observer pose, these attributes are estimated through perception and supplied to the corresponding computation. View-dependent queries use the explicit observer frames described below.

Observer reference frames. For a horizontal directional query, an observer at $\mathbf{a}$ faces an object at $\mathbf{b}$, and the unit scene-up vector is $\mathbf{u}$. We compute

$$\mathbf{f} = \text{normalize}\big[(\mathbf{b} - \mathbf{a}) - \mathbf{u}\,\mathbf{u}^\top (\mathbf{b} - \mathbf{a})\big], \qquad \mathbf{r} = \text{normalize}(\mathbf{f} \times \mathbf{u}). \tag{12}$$

The observer basis is $B = [\mathbf{r}, \mathbf{f}, \mathbf{u}]$. For a target at $\mathbf{c}$, its observer-frame coordinates are $\mathbf{q} = B^\top(\mathbf{c} - \mathbf{a})$; their signs indicate right/left, front/back, and above/below.

For full 3D scene simulation, we retain the complete viewing direction: $\mathbf{f} = \mathrm{normalize}(\mathbf{b} - \mathbf{a})$. We compute $\mathbf{r} = \mathrm{normalize}(\mathbf{f} \times \mathbf{u})$ and use the orthogonalized up axis $\mathbf{u}' = \mathbf{r} \times \mathbf{f}$, yielding $B = [\mathbf{r}, \mathbf{f}, \mathbf{u}']$. A zero horizontal viewing direction is invalid for horizontal queries. For full 3D queries, coincident observer and facing locations are invalid. When the viewing direction is parallel to scene up, a fixed fallback reference axis determines the otherwise unspecified roll, and use of this convention is recorded. Observer selection, viewing-direction estimation, and coordinate transformation remain distinct steps; reference-frame ambiguity cannot be resolved by silently comparing global coordinates.