# GEOLIFT: GROUNDING VIDEO SPATIAL REASONING IN EXPLICIT 3D MAPS

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

## ABSTRACT

Video spatial reasoning requires multimodal large language models (MLLMs) to understand how objects are arranged and related across viewpoints. While these models can recognize objects in individual frames, integrating these observations to solve complex spatial tasks remains challenging. We examine route planning as a representative task and observe that existing models become substantially less accurate as the number of required turns increases, with several models achieving accuracy close to random guessing on routes requiring three or more turns. This finding suggests that recognizing landmarks alone is insufficient: models must also connect observations across viewpoints and track how landmarks are arranged in a shared space. Motivated by this observation, we introduce GeoLift, a training-free framework that uses feed-forward 3D reconstruction to map objects detected in individual frames into a shared metric 3D scene. GeoLift combines a frozen MLLM for 2D grounding with pretrained geometry models for scene reconstruction. It then uses shared geometric regions, called superpoints, to associate observations of the same object across views without a learned matcher. The reconstructed 3D map enables explicit computation of spatial relations and multistep routes, and provides geometric evidence for estimating distances and sizes. With all model weights frozen, GeoLift improves Qwen3-VL-4B's average score on VSI-Bench from 55.4% to 64.1%, including an increase in route-planning accuracy from 30.9% to 53.6%, and achieves 59.8% accuracy on MindCube. These results show that connecting frame-level object grounding through a shared 3D map can improve complex spatial reasoning without task-specific training.

## 1 INTRODUCTION

Spatial reasoning requires more than recognizing what appears in an image. An agent must understand where objects are, how they relate to one another, and how these relationships change with the observer's viewpoint. These abilities are essential for navigation, manipulation, and interaction with the physical environment. Human spatial cognition integrates observations from different viewpoints into stable representations of places and spatial relations (Orban, 2011; Kravitz et al., 2011; Epstein & Kanwisher, 1998; Maguire et al., 1998). Developing similar capabilities in multimodal large language models (MLLMs) is a central challenge in video and multiview spatial reasoning.

A video presents a sequence of partial views rather than a complete description of a scene. As the camera moves, the same object appears at different image locations, while objects observed in separate frames may never be visible together. Answering spatial questions therefore requires connecting observations across viewpoints and understanding their relationships in a shared space. Modern MLLMs can recognize and ground objects in individual frames, but these capabilities do not necessarily translate into reliable scene-level spatial reasoning (Wang et al., 2025).

Existing approaches often improve spatial reasoning by training MLLMs to incorporate geometric information, either through feature fusion or geometric supervision (Zheng et al., 2025; Wu et al., 2025; Fan et al., 2026; Yeh et al., 2026). As illustrated in Figure 1 (a), representative geometry-augmented methods fuse visual and geometric features and infer spatial answers from learned representations. However, our route-planning analysis reveals that several direct and spatially trained MLLMs perform close to random guessing as the number of required turns increases (section 3.1). This finding suggests that geometric awareness alone may not be sufficient for complex tasks:



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.

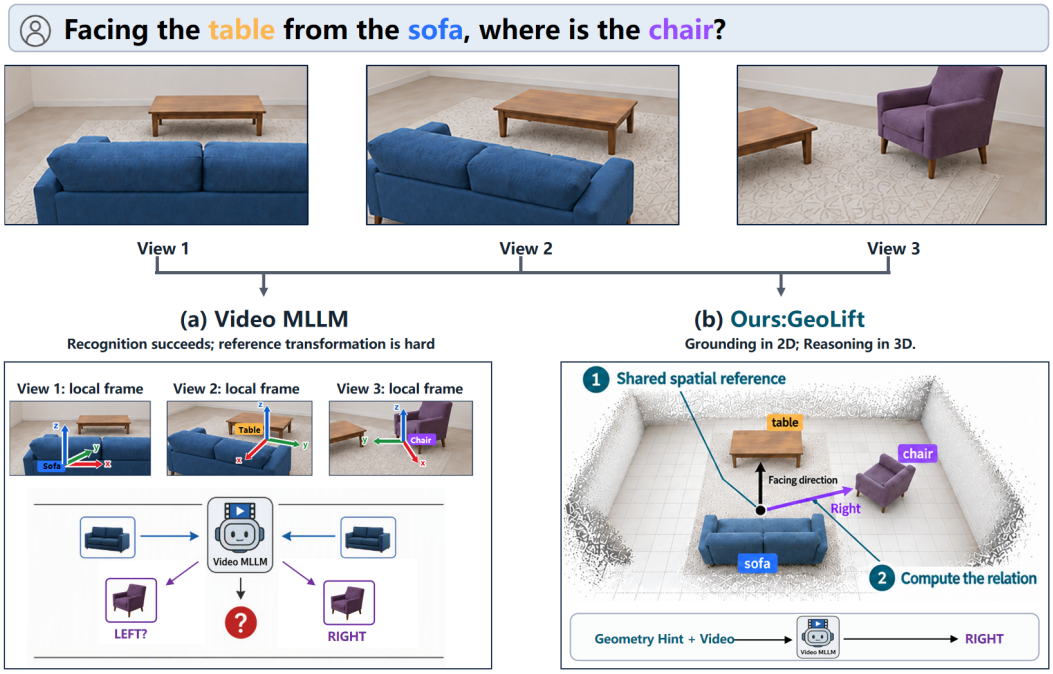



Figure 1: **From local 2D observations to explicit 3D spatial reasoning.** Given the same RGB video and spatial query, **(a) Video MLLM** grounds objects in individual frames, but each observation remains tied to its local coordinate frame. Without a shared reference, the cross-view relation between the sofa and chair is ambiguous. **(b) GeoLift** lifts the grounded objects into a shared spatial reference and explicitly computes the relation. For this query, the sofa-to-table direction defines the facing orientation, which places the chair on the right.

models also need to maintain a consistent spatial reference and compose relations across multiple observations.

Our key insight is that feed-forward 3D reconstruction can provide this shared reference by mapping frame-level 2D groundings into a common 3D space. As illustrated in Figure 1 (b), grounding objects in an explicit 3D map allows their spatial relationships to be computed directly. Building on this idea, we introduce GeoLift, a training-free framework that combines frozen MLLM grounding with reconstructed metric geometry. GeoLift uses geometric superpoints (Landrieu & Simonovsky, 2018; Robert et al., 2023) to associate observations of the same object across viewpoints, producing a view-consistent map for explicit spatial computation and geometry-guided reasoning. All neural components remain frozen, with no benchmark-specific parameter updates.

We evaluate GeoLift on VSI-Bench and MindCube, covering egocentric videos and sparse multiview observations. With Qwen3-VL-4B, GeoLift improves the average VSI-Bench score from 55.4% to 64.1%, including an increase in route-planning accuracy from 30.9% to 53.6%, and achieves 59.8% overall accuracy on MindCube. These results support the value of connecting local visual groundings through a shared 3D map, particularly for tasks that require combining spatial relations across viewpoints. Our contributions are:

- We identify a weakness in existing MLLMs on route-planning tasks requiring multiple turns, motivating a shared spatial reference for connecting observations across viewpoints and composing spatial relations.
- We introduce GeoLift, a training-free framework that lifts 2D groundings into a reconstructed metric 3D map and associates observations across viewpoints through geometric superpoint agreement, enabling explicit spatial computation and geometry-guided reasoning.
- With all model weights frozen, GeoLift improves Qwen3-VL-4B's average score on VSI-Bench from 55.4% to 64.1%, with gains of 28.3 and 22.7 percentage points in relative-

direction and route-planning accuracy, respectively. On MindCube, it improves overall accuracy from 23.4% to 59.8%, a gain of 36.4 percentage points.

# 2 Related Work

## 2.1 Video and Multiview Spatial Reasoning Benchmarks

Video and multiview spatial reasoning requires models to integrate evidence distributed across frames and viewpoints, especially when no single observation contains the full spatial context. VSI-Bench (Yang et al., 2025) evaluates egocentric video reasoning through object counting, relative and absolute distance, relative direction, object and room size, route planning, and appearance order. MindCube (Wang et al., 2025) complements this setting by studying cognitive mapping, perspective taking, and counterfactual spatial inference from limited multiview observations. Together, these benchmarks test whether a model can maintain a coherent spatial state across observations for metric comparison, viewpoint transformation, object relations, and multistep reasoning, revealing the gap between strong frame-level recognition and reliable scene-level spatial intelligence.

## 2.2 Learning-Based Spatial Reasoning

Learning-based approaches improve spatial reasoning through task-specific spatial supervision, post-training, or learned geometric interfaces. Video-R1, Cambrian-S, and InternSpatial advance this direction through spatially annotated data and supervised or reinforcement learning (Feng et al., 2025; Yang et al., 2026; Deng et al., 2026). Geometry-augmented methods, including VG-LLM, Spatial-MLLM, and VLM-3R, align depth, camera, or multiview geometric features with visual-language representations through learned fusion (Zheng et al., 2025; Wu et al., 2025; Fan et al., 2026). Structured representations provide another route: LASAR and EagleVision introduce latent cognitive maps and BEV representations, respectively (Tang et al., 2026; Wan et al., 2026), while VoxAnchor explicitly aligns cross-view visual features in a voxel-semantic space (Li et al., 2026). These approaches demonstrate the value of spatially organized representations, but generally require additional learning to align geometric cues with MLLM reasoning. Their spatial states therefore remain latent or are accessed through a learned query interface, motivating a complementary design based on explicit geometric computation.

## 2.3 Training-Free Spatial Reasoning

Training-free methods improve spatial reasoning at inference time by composing frozen MLLMs with prompting, rendering, or pretrained perception modules. VideoAnchor reinforces consistent cross-frame visual cues (Wang et al., 2026c), TRACE organizes spatial evidence into textual representations (Hua et al., 2026), and Reason, then Re-reason uses rendered novel views to verify and revise spatial hypotheses (Ma et al., 2026). Other approaches construct explicit spatial representations using pretrained perception models. GeoLift combines frozen semantic and geometric models with geometric superpoints, lifting 2D groundings into a fused point cloud and associating cross-view observations without a learned matcher. No components are updated and no benchmark-specific spatial-QA supervision is used. Unlike text, latent-feature, or rendered-view interfaces, GeoLift uses a shared metric 3D workspace for explicit spatial computation.

# 3 Method

## 3.1 Problem Setup and Motivation

**Problem setup.** Let $V = \{I_t\}_{t=1}^{T}$ denote an RGB video or a sparse collection of views, let $q$ be a spatial query, and let $y$ be its answer. The goal is to predict $y$ from $(V, q)$ without updating model parameters on the target benchmark. The central challenge is to transform partial observations from changing viewpoints into a spatial representation that supports consistent reasoning.

**Motivation.** We observe that current video MLLMs remain weak on complex spatial reasoning tasks that require combining information across viewpoints and reasoning over multiple steps. In this

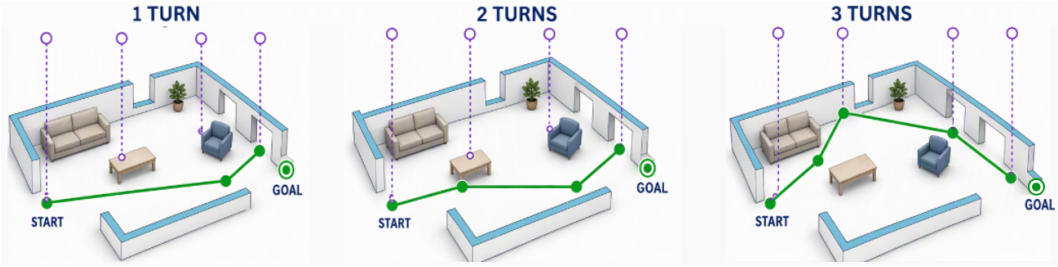



Figure 2: **Route-planning complexity as the number of turns increases.** The green path connects the start and goal through an ordered sequence of intermediate landmarks. A higher turn count requires more direction changes and stronger consistency across observations from different viewpoints, making the underlying global spatial arrangement increasingly important.

section, we examine this limitation through route planning, where a model must connect landmarks observed in different frames and determine an ordered sequence of turns. To study how performance changes with reasoning complexity, we group route-planning questions from VSI-Bench by the number of required turns: one, two, and three or more, as examples shown on Figure 2. We compare direct MLLM inference, spatially trained models, and GeoLift within each group. Because one-turn questions have three answer choices while multi-turn questions have four, we also report the corresponding random-choice accuracy of 33.3% and 25.0%, respectively.

Table 1 shows that existing models generally become less accurate as more turns are required. On questions with three or more turns, Qwen3-VL-4B achieves 23.7% accuracy, while GS-Reasoner and Cambrian-S with 32 frames both achieve 26.3%, close to random guessing. These results suggest that recognizing individual landmarks and incorporating geometric cues do not by themselves ensure reliable multistep spatial reasoning. A model must also associate the same landmark across viewpoints, maintain its position relative to other landmarks, and update the reference direction after each turn. Errors in these operations may accumulate across successive steps, making routes with more turns particularly challenging.

GeoLift addresses these requirements by organizing frame-level 2D groundings in a shared reconstructed 3D map, where landmark observations can be associated and spatial relations explicitly computed and composed. On questions requiring three or more turns, GeoLift achieves 47.4% accuracy, exceeding the direct Qwen3-VL-4B baseline by 23.7 percentage points. Although this task remains challenging, the improvement supports the value of a shared geometric reference for multistep spatial reasoning. The following sections describe how GeoLift constructs and uses this reference.

Table 1: **Longer routes expose the need for a persistent spatial state.** Route-planning accuracy is grouped by the number of required turns. Values are accuracies (%), with differences from the corresponding random-guessing baseline shown in parentheses in percentage points. Red entries are below the baseline, while orange entries are at most five percentage points above it.

| **Model** | **Protocol** | **1 turn** 3 choices | **2 turns** 4 choices | **≥3 turns** 4 choices |
|---|---|---|---|---|
| Qwen3-VL-4B | frozen | 43.4 (+10.1) | 22.5 (−2.5) | 23.7 (−1.3) |
| GPT-5.6-Sol | frozen | 48.7 (+15.4) | 36.3 (+11.3) | 28.9 (+3.9) |
| EgoMind 7B | learned | 34.2 (+0.9) | 28.8 (+3.8) | 28.9 (+3.9) |
| Cambrian-S 7B (32f) | learned | 44.7 (+11.4) | 41.3 (+16.3) | 26.3 (+1.3) |
| Cambrian-S 7B (64f) | learned | 53.9 (+20.6) | 35.0 (+10.0) | 31.6 (+6.6) |
| GS-Reasoner | learned | 55.3 (+22.0) | 30.0 (+5.0) | 26.3 (+1.3) |
| VLM-3R 7B | learned | 50.0 (+16.7) | 45.0 (+20.0) | 39.5 (+14.5) |
| **GeoLift (ours)** | **frozen** | **57.9** (+24.6) | **52.5** (+27.5) | **47.4** (+22.4) |
| Random guessing | – | 33.3 | 25.0 | 25.0 |

The three groups contain 76, 80, and 38 questions, respectively; the ≥3-turn group combines questions requiring 3 and 4 turns. Differences are computed from the displayed accuracies and baselines.

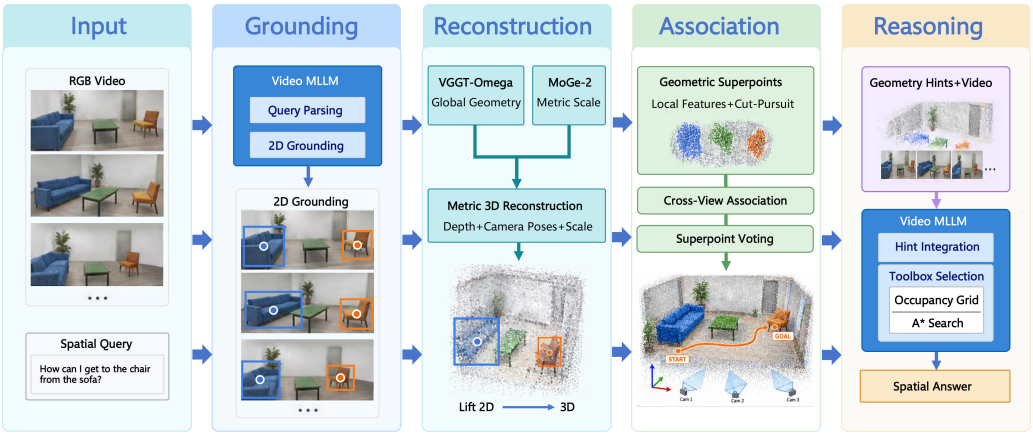



Figure 3: **GeoLift framework.** A frozen Video MLLM parses the query and grounds query-relevant entities in the RGB video. VGGT-Omega and MoGe-2 recover global geometry and metric scale, allowing the 2D groundings to be lifted into a shared 3D reference. Geometric superpoints and superpoint agreement associate observations of the same instance across views without a learned matcher. The resulting view-consistent map supports a toolbox of deterministic spatial operations, including occupancy-grid construction and A* route search. Geometry can also be provided as a hint together with the video for queries that require visual interpretation.

### 3.2 GeoLift Overview

Figure 3 illustrates the pipeline of GeoLift, which consists of four stages: grounding, reconstruction, association, and reasoning. Given an RGB video and a spatial query, a frozen video MLLM first identifies and localizes query-relevant objects in individual frames. VGGT-Omega reconstructs the scene geometry, while MoGe-2 calibrates its metric scale, allowing these 2D groundings to be mapped into a shared 3D space. GeoLift then uses geometric superpoint agreement to associate observations of the same object across viewpoints, producing a 3D map with consistent object identities. Finally, the reconstructed map supports spatial reasoning in two ways: directly computing spatial relations and routes, or providing geometric evidence alongside the original video to guide the MLLM's answer. All neural components remain frozen throughout the pipeline. The following sections describe each stage in detail.

### 3.3 Grounding

The grounding module localizes query-relevant objects in individual frames for subsequent mapping into 3D. A frozen video MLLM first searches a grid of sampled frames with frame indices, then refines the target bounding boxes in the selected original frames. Missing targets trigger a denser frame search. For sparse multiview inputs, objects are localized independently in each view, with separate prompts for individual objects when needed. Each observation records the object name, frame or view index, normalized 2D bounding box, and confidence score, with an optional 2D point for navigation landmarks. Low-confidence observations and invalid boxes are discarded before 3D lifting and cross-view association.

### 3.4 Reconstruction

The reconstruction module builds a shared metric 3D space and maps the grounded 2D observations into it. It consists of two steps: reconstructing the scene geometry and lifting object observations from individual frames into 3D.

**3D reconstruction.** We use frozen VGGT-Omega (Wang et al., 2026a) to predict a depth map $d_f^{\mathrm{V}}$, camera intrinsics $K_f$, and a camera-to-world pose $[R_f \mid t_f]$ for each frame $f$. These predictions place observations from different views in a common coordinate system, but the reconstruction has an

unknown scale: its distances do not directly correspond to real-world units. To recover metric scale, we use depth estimates $d_f^{\mathrm{M}}$ from frozen MoGe-2 as a reference. After resizing the MoGe-2 (Wang et al., 2026b) depth maps to match the VGGT-Omega resolution, we estimate a single scale factor $s$ that minimizes the squared differences between the scaled VGGT-Omega depths and the MoGe-2 depths:

$$s = \frac{\sum_{(f,p)\in\Omega} d_f^{\mathrm{V}}(p) d_f^{\mathrm{M}}(p)}{\sum_{(f,p)\in\Omega} \left(d_f^{\mathrm{V}}(p)\right)^2}, \tag{1}$$

where $\Omega$ contains valid, high-confidence pixel correspondences between the two depth predictions across frames. We apply $s$ to both the predicted depths and camera translations:

$$\tilde{d}_f(p) = s d_f^{\mathrm{V}}(p), \qquad \tilde{t}_f = s t_f. \tag{2}$$

Camera rotations and intrinsics remain unchanged.

**2D-to-3D lifting.** We then use the calibrated geometry to map each grounded bounding box into 3D. Since a box may include background pixels, we slightly shrink it and retain only valid pixels whose depths are close to the median depth within the box. Each retained pixel is first back-projected into the camera coordinate system using its depth and camera intrinsics, then transformed into the shared world coordinate system:

$$\tilde{\mathbf{X}}_f(p) = R_f \left(\tilde{d}_f(p) K_f^{-1} \bar{p}\right) + \tilde{t}_f, \tag{3}$$

where $\bar{p} = (u, v, 1)^\top$ denotes the homogeneous pixel coordinates. We combine these points across frames while retaining their source-frame indices. The resulting metric point cloud places the grounded observations in a common 3D space for subsequent cross-view association and spatial reasoning.

### 3.5 Association

After reconstruction, observations from different frames are placed in the same 3D coordinate system, but it remains unclear which observations correspond to the same physical instance. This problem is further complicated by imperfect 2D grounding: bounding boxes predicted by the Video MLLM can be loose and include substantial background regions. Directly lifting and aggregating all points inside these boxes therefore introduces background points that vary across views and can make cross-view matching unreliable.

To obtain more robust geometric evidence, we partition the fused point cloud into geometrically coherent *superpoints* following Robert et al. (2023). Superpoints group nearby points with similar local geometry, allowing object surfaces to be separated from surrounding background structures. We construct a $k$-nearest-neighbor graph and apply cut-pursuit optimization (Landrieu & Obozinski, 2017) using point coordinates and local geometric features such as planarity, surface orientation, and curvature.

For each grounded observation, its filtered 3D points vote for the superpoints they occupy. This converts a noisy set of box-level points into a compact set of geometric regions. Weak votes are discarded, and a one-hop dilation is applied over adjacent superpoints to connect partial observations of the same surface. To further suppress common background regions, we downweight superpoints that are shared by many observations. Let $c_{i,s}$ denote the vote count of observation $i$ for superpoint $s$, $\mathrm{df}(s)$ the number of observations voting for $s$, and $N$ the total number of observations:

$$w_{i,s} = c_{i,s} \left(\log \frac{1+N}{1+\mathrm{df}(s)} + 1\right).$$

We then associate observations from different frames using their weighted superpoint evidence. Two observations are matched when their vote vectors have high cosine similarity and their 3D centers satisfy an extent-aware distance constraint. Compatible matches are merged using union-find, while merges causing excessive center drift are rejected to avoid connecting nearby but distinct instances.

For each query entity, the largest consistent component defines a view-consistent track, 3D anchor, and spatial footprint. Competing separated components are marked as conflicts, while insufficient

Table 2: **Comparative results on full VSI-Bench** (Yang et al., 2025). The numerical and multiple-choice columns report MRA and exact-match accuracy, respectively. Training-free means no benchmark-specific updates; † marks quoted results. Direct Qwen3-VL-4B and GeoLift use the same frozen backbone and matched evaluation protocol. All result values are percentages rounded to one decimal place.

| Methods | Overall | Numerical Questions | | | | Multiple-Choice Questions | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Obj. Cnt. | Abs. Dist. | Obj. Size | Room Size | Rel. Dist. | Rel. Dir. | Route Plan | Appr. Order |
| *General-Purpose MLLMs* | | | | | | | | | |
| GPT-4o† (Yang et al., 2025) | 34.0 | 46.2 | 5.3 | 43.8 | 38.2 | 37.0 | 41.3 | 31.5 | 28.5 |
| Gemini-1.5 Flash† (Yang et al., 2025) | 42.1 | 49.8 | 30.8 | 53.5 | 54.4 | 37.7 | 41.0 | 31.5 | 37.8 |
| Gemini-1.5 Pro† (Yang et al., 2025) | 45.4 | 56.2 | 30.9 | 64.1 | 43.6 | 51.3 | 46.3 | 36.0 | 34.6 |
| InternVL2-8B† (Yang et al., 2025) | 37.5 | 31.3 | 29.0 | 48.9 | 44.2 | 38.0 | 33.4 | 28.9 | 46.4 |
| InternVL3-78B† (Zhao et al., 2026) | 48.5 | 71.2 | 53.7 | 44.4 | 39.5 | 55.9 | 39.5 | 28.9 | 55.4 |
| Qwen2.5-VL-7B† (Zhao et al., 2026) | 33.0 | 40.9 | 14.8 | 43.4 | 10.7 | 38.6 | 38.5 | 33.0 | 29.8 |
| LLaVA-OneVision-72B† (Zhao et al., 2026) | 40.2 | 43.5 | 23.9 | 57.6 | 37.5 | 42.5 | 39.9 | 32.5 | 44.6 |
| LLaVA-NeXT-Video-72B† (Zhao et al., 2026) | 40.9 | 48.9 | 22.8 | 57.4 | 35.3 | 42.4 | 36.7 | 35.0 | 48.6 |
| Qwen3-VL-8B-Instruct† (Huang et al., 2026) | 57.4 | 66.7 | 38.9 | 74.1 | 53.5 | 59.0 | 65.4 | 32.0 | 69.4 |
| Qwen3-VL-4B-Instruct (Bai et al., 2025) | 55.4 | 65.9 | 43.1 | 73.9 | 61.0 | 53.1 | 46.1 | 30.9 | 69.1 |
| *Training-Based Spatial MLLMs* | | | | | | | | | |
| VG-LLM-4B† (Zheng et al., 2025) | 47.3 | 66.0 | 37.8 | 55.2 | 59.2 | 44.6 | 45.6 | 33.5 | 36.4 |
| VG-LLM-8B† (Zheng et al., 2025) | 50.7 | 67.9 | 37.7 | 58.6 | 62.0 | 46.6 | 40.7 | 32.4 | 59.2 |
| Spatial-MLLM† (Wu et al., 2025) | 48.4 | 65.3 | 34.8 | 63.1 | 45.1 | 41.3 | 46.2 | 33.5 | 46.3 |
| 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 |
| VLM-3R† (Fan et al., 2026) | 60.9 | 70.2 | 49.4 | 69.2 | 67.1 | 65.4 | 80.5 | 45.4 | 40.1 |
| GS-Reasoner (w/d)† (Chen et al., 2026a) | 64.7 | 69.1 | 61.9 | 70.0 | 65.7 | 65.4 | 88.9 | 44.3 | 52.3 |
| SpaceMind† (Zhao et al., 2026) | 69.6 | 73.3 | 61.4 | 77.3 | 74.2 | 67.2 | 88.4 | 44.3 | 70.6 |
| EgoMind (Qwen2.5-VL-7B)† (Chen et al., 2026b) | 50.2 | 55.4 | 37.9 | 67.1 | 40.4 | 44.1 | 47.2 | 32.0 | 58.4 |
| SpatioLM (SenseNovaSI-8B)† (Wu et al., 2026) | 71.6 | 81.9 | 55.7 | 75.2 | 71.5 | 78.2 | 74.4 | 47.9 | 87.8 |
| ConsiSpace (UC-SSRL)† (Huang et al., 2026) | 76.6 | 77.9 | 78.9 | 82.2 | 76.3 | 72.1 | 90.1 | 54.7 | 79.9 |
| SpaR3D-MoE† (Feng et al., 2026) | 63.5 | 71.3 | 48.0 | 72.8 | 69.6 | 58.7 | 70.1 | 44.0 | 73.5 |
| VoxAnchor-4B† (Li et al., 2026) | 66.8 | 66.3 | 52.6 | 72.5 | 71.0 | 67.3 | 80.0 | 46.4 | 78.0 |
| *Training-Free* | | | | | | | | | |
| VideoAnchor† (InternVL2-8B) (Wang et al., 2026c) | 37.8 | 39.3 | 30.3 | 50.5 | 38.8 | 38.6 | 31.0 | 36.1 | 37.8 |
| TRACE† (Qwen2.5-VL-72B-Instruct) (Hua et al., 2026) | 39.4 | 22.1 | 28.0 | 60.0 | 39.0 | 40.9 | 37.4 | 32.0 | 42.6 |
| Reason, Then Re-reason† (Qwen3-VL-4B) (Ma et al., 2026) | 36.5 | 21.1 | 26.7 | 50.2 | 35.8 | 37.2 | 43.1 | 28.9 | 34.1 |
| **Ours (Qwen3-VL-4B)** | **64.1** | **68.0** | **49.4** | **74.4** | **65.9** | **57.8** | **74.4** | **53.6** | **69.1** |

superpoint evidence falls back to a robust bounding-box depth center marked as unverified. All association decisions rely on reconstructed geometry and fixed rules, without a learned cross-view matcher or benchmark-specific association training.

### 3.6 Reasoning

After reconstruction and association, GeoLift provides a view-consistent 3D map together with the original video to the frozen Video MLLM. The Video MLLM acts as the reasoning controller: it integrates geometric hints with visual evidence and selects appropriate tools from a geometric toolbox according to the spatial query.

The toolbox supports explicit spatial operations in the shared 3D space, such as computing relative directions and distances. Route planning is one representative example, where the model invokes occupancy-grid construction and A* search (Hart et al., 1968) to connect successive landmarks and derive the required turns.

For queries that cannot be fully resolved by deterministic geometry, the reconstructed map provides metric anchors, object extents, distances, or floor footprints as additional hints to the Video MLLM. A reliability gate checks the consistency of the geometric evidence; when it is unreliable, the model relies more heavily on the original visual evidence or the direct Video MLLM prediction.

## 4 Experiments

### 4.1 Experimental Setup

We evaluate GeoLift on the official full splits of VSI-Bench (Yang et al., 2025) and MindCube (Wang et al., 2025). VSI-Bench contains 5,130 questions from 288 egocentric videos; its overall score is the unweighted mean of eight category scores, using MRA for numerical tasks and exact-match accuracy for multiple-choice tasks. MindCube contains 21,154 sparse-multiview questions; we report Rotation, Among, and Around accuracies together with the pooled micro-average over all questions.

Table 3: **Results on the full MindCube benchmark** (Wang et al., 2025). We report accuracy for the Rotation, Among, and Around settings. The Around setting includes the AroundNew subset. The overall score is the micro-averaged accuracy over all 21,154 questions. GeoLift is training-free: the Qwen3-VL-4B variant uses Qwen3-VL-4B for grounding and answering. Bold numbers indicate the best value in each column. All values are percentages rounded to one decimal place.

| Method | Overall (micro) | Rotation | Among | Around |
|---|---|---|---|---|
| *Baseline* | | | | |
| Chance | 32.4 | 36.4 | 32.3 | 30.7 |
| *Open-Weight Multi-Image Models* | | | | |
| LLaVA-OneVision-7B | 47.4 | 36.5 | 48.4 | 44.1 |
| Qwen2.5-VL-7B-Instruct | 29.3 | 38.8 | 29.5 | 21.4 |
| Gemma-3-12B-it | 46.7 | 38.4 | 48.4 | 34.6 |
| Mantis-8B (SigLip) | 41.1 | 37.7 | 40.2 | 51.0 |
| DeepSeek-VL2-Small | 47.6 | 37.0 | 50.4 | 26.9 |
| Qwen3-VL-4B | 23.4 | 32.0 | 21.9 | 33.2 |
| *Proprietary Models* | | | | |
| GPT-5 | 47.6 | **93.3** | 34.2 | 41.6 |
| Claude-4-Sonnet | 44.8 | 48.4 | 44.2 | 47.6 |
| GPT-5.6-Sol | 54.1 | 67.1 | 52.6 | 61.5 |
| *Spatial Models* | | | | |
| RoboBrain | 37.4 | 35.8 | 38.3 | 29.5 |
| SpaceMantis | 22.8 | 37.7 | 21.3 | 29.3 |
| Space-Qwen | 33.3 | 38.0 | 33.7 | 26.3 |
| **Ours (Qwen3-VL-4B)** | **59.8** | 56.2 | **58.5** | **74.6** |

All neural components remain frozen, with no benchmark-specific updates or spatial-QA supervision. Published baselines retain their reported protocols; the direct Qwen3-VL-4B baseline and a matched no-superpoint variant provide controlled comparisons, with implementation and evaluation details given in the supplementary material.

### 4.2 Main Results on VSI-Bench

Table 2 shows that GeoLift produces task-dependent rather than uniform improvements. Relative to direct Qwen3-VL-4B-Instruct inference, GeoLift raises the VSI-Bench category macro-average from 55.4 to 64.1, with the largest gains in relative direction (+28.3 percentage points) and route planning (+22.7 percentage points). These gains are concentrated in tasks that require integrating evidence across multiple views, maintaining a persistent global spatial state, and composing several spatial relations, particularly in long-horizon route planning. By contrast, object counting and object-size estimation are comparatively local: an informative frame or a small number of frames may be sufficient, and size judgments may additionally draw on local visual cues and category-level priors acquired during pretraining. Their weaker dependence on cross-view integration limits the marginal benefit of explicit 3D reconstruction. GeoLift also exceeds several listed training-based spatial baselines, including VG-LLM-8B (50.7), Spatial-MLLM (48.4), VLM-3R (60.9), and SpaR3D-MoE (63.5), as well as the direct Qwen3-VL-8B-Instruct baseline (57.4) and the reported training-free baselines TRACE (39.4), VideoAnchor (37.8), and Reason, Then Re-reason (36.5).

### 4.3 Results on MindCube

To assess transfer beyond egocentric video, Table 3 reports GeoLift's performance on MindCube, which uses sparse multiview observations. We use the same frozen semantic and geometric components as on VSI-Bench.

On the full MindCube benchmark, GeoLift achieves 59.8% micro-averaged accuracy, improving over direct Qwen3-VL-4B inference (23.4%) by 36.4 percentage points. Performance is highest on Around (74.6%), followed by Among (58.5%) and Rotation (56.2%), a pattern consistent with the

Table 4: **Ablation of metric geometry and superpoint association on VSI-Bench.** All rows use the frozen Qwen3-VL-4B Video MLLM. A checkmark indicates that the component is enabled, while a cross indicates that it is disabled. The first row is direct Qwen3-VL-4B inference without VGGT-Omega, MoGe-2, or superpoint association. Geometric variants use the fixed 64-frame protocol. Scores are percentages.

| VGGT+MoGe | SP Association | Overall | Numerical Questions | | | | Multiple-Choice Questions | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Obj. Cnt. | Abs. Dist. | Obj. Size | Room Size | Rel. Dist. | Rel. Dir. | Route Plan. | Appr. Order |
| × | × | 55.4 | 65.9 | 43.1 | 73.9 | 61.0 | 53.1 | 46.1 | 30.9 | 69.1 |
| ✓ | × | 62.8 | 68.0 | 48.4 | 74.4 | 65.5 | 57.8 | 72.1 | 46.9 | 69.1 |
| ✓ | ✓ | **64.1** | **68.0** | **49.4** | **74.4** | **65.9** | **57.8** | **74.4** | **53.6** | **69.1** |

value of organizing sparse observations into a shared spatial representation. GeoLift also exceeds the GPT-5.6-Sol baseline in overall accuracy (54.1%) and on Around (61.5%). GPT-5 nevertheless retains the highest Rotation score (93.3%). Together, these results complement VSI-Bench by supporting the usefulness of explicit spatial organization beyond egocentric video.

### 4.4 Analysis and Ablations

Table 4 isolates the contributions of metric geometry and superpoint association. Introducing the VGGT-Omega/MoGe-2 geometry branch raises the category macro-average from 55.4 to 62.8, with particularly pronounced improvements in relative direction (+26.0 points) and route planning (+16.0 points). Superpoint association further increases the macro-average to 64.1, contributing the largest additional gain to route planning (+6.7 points), followed by relative direction (+2.3 points), with further improvements in absolute-distance and room-size estimation. The concentration of these gains in tasks that require integrating observations across views and composing spatial relations is consistent with GeoLift's intended role: bridging frame-level visual perception and scene-level spatial reasoning through a shared spatial reference.

## 5 Limitations

GeoLift's training-free design relies on frozen visual and geometric modules, making performance sensitive to 2D grounding and the reconstructed spatial reference. The Video MLLM determines entity coverage and localization, while VGGT-Omega supplies camera poses and depth and MoGe-2 calibrates metric scale; residual errors can propagate to spatial relations, metric estimates, and route planning. Reliability checks and fallback to direct Video MLLM answers mitigate uncertain geometry but cannot recover missing or ambiguous evidence. Future work could complement this training-free design with learning-based adaptation of grounding, reconstruction, metric calibration, and spatial reasoning to improve robustness and performance across task categories while preserving the explicit 3D representation.

## 6 Conclusion

Current Video MLLMs can recognize and localize query-relevant objects in individual frames but remain less reliable when integrating spatial evidence across viewpoints and multiple steps. We introduced GeoLift, a training-free framework that complements a frozen Video MLLM with a persistent metric 3D representation; for video inputs, geometric superpoints associate observations across views, while explicit spatial operators compute relations and routes. With the same frozen Qwen3-VL-4B backbone, GeoLift raises the VSI-Bench category macro-average from 55.4% to 64.1% and MindCube overall accuracy from 23.4% to 59.8%, with the largest improvements in relative direction and route planning. Although the gains are task-dependent, these results indicate that pairing pretrained visual grounding with an explicit 3D spatial state broadens the range of tasks that current Video MLLMs can address reliably, providing a practical path toward stronger cross-view and multistep reasoning without benchmark-specific parameter updates.

### AI USE STATEMENT

Generative AI tools were used to assist with literature search and organization, language editing, code development, and the generation and refinement of illustrative figures. All AI-assisted outputs, including references, code, numerical results, claims, and visual artifacts, were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

### REPRODUCIBILITY STATEMENT

GeoLift is evaluated as a frozen inference procedure: no model weights are updated and no benchmark-specific spatial-QA supervision is used. The main paper and supplementary material specify the frozen Video MLLM and geometry pipeline and preprocessing settings, fixed geometric thresholds, output format, evaluation protocols, and ablation configurations. All settings are fixed before benchmark evaluation, and our experiments use the official splits and metrics of the public VSI-Bench and MindCube benchmarks.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Yiming Chen, Zekun Qi, Wenyao Zhang, Xin Jin, Li Zhang, and Peidong Liu. Reasoning in space via grounding in the world. In *International Conference on Learning Representations*, 2026a.

Zhenghao Chen, Huiqun Wang, and Di Huang. EgoMind: Activating spatial cognition through linguistic reasoning in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Jinhui Yin, Qi Wei, Tianshuo Yang, et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 50773–50814, 2026.

Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. 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.

Haida Feng, Hao Wei, Haolin Wang, Shiwei Li, Chade Li, and Yihong Wu. Spar3d-moe: Adaptive 3d spatial reasoning from sparse views meets geometry-inductive mixture-of-experts. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in mllms. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

Jiacheng Hua, Yishu Yin, Yuhang Wu, Tai Wang, Yifei Huang, and Miao Liu. Unleashing spatial reasoning in multimodal large language models via textual representation guided reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Ting Huang, Zhenyu Zhang, Wenyuan Huang, Jian Yang, and Hao Tang. Consispace: Learning geometric consistency matters for video spatial reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, and Mortimer Mishkin. A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, 12(4):217–230, 2011.

Loic Landrieu and Guillaume Obozinski. Cut pursuit: Fast algorithms to learn piecewise constant functions on general weighted graphs. *SIAM Journal on Imaging Sciences*, 10(4):1724–1766, 2017.

Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4558–4567. IEEE, 2018.

Xinglin Li, TingTing Long, Jingzhi Zhou, Chuxuan Zeng, Jiajing Chen, Jian Yang, and Jin Xie. Voxanchor: Explicit voxel-semantic grounding for spatial understanding in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. URL https://eccv.ecva.net/virtual/2026/poster/4848.

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. *arXiv preprint arXiv:2606.11683*, 2026.

Eleanor A Maguire, CD Frith, Neil Burgess, James G Donnett, and John O'Keefe. Knowing where things are: Parahippocampal involvement in encoding object locations in virtual large-scale space. *Journal of cognitive neuroscience*, 10(1):61–76, 1998.

Guy A Orban. The extraction of 3d shape in the visual system of human and nonhuman primates. *Annual review of neuroscience*, 34(1):361–388, 2011.

Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17149–17158. IEEE, 2023.

Jinzhou Tang, Sidi Liu, Waikit Xiu, Weixing Chen, and Keze Wang. LASAR: Towards spatio-temporal reasoning with latent cognitive map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jiaxu Wan, Xu Wang, Mengwei Xie, Hang Zhang, Mu Xu, Yang Han, Ding Yuan, Hong Zhang, and Yifan Yang. EagleVision: A dual-stage framework with bev-grounding-based chain-of-thought for spatial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-\omega. *arXiv preprint arXiv:2605.15195*, 2026$a$.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Li Fei-Fei, Jiajun Wu, and Manling Li. MindCube: Spatial mental modeling from limited views. *arXiv preprint arXiv:2506.21458*, 2025.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026b.

Zhaozhi Wang, Tong Zhang, Mingyue Guo, Yaowei Wang, and Qixiang Ye. VideoAnchor: Reinforcing subspace-structured visual cues for coherent visual-spatial reasoning. In *International Conference on Learning Representations*, 2026c.

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

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. Spatiolm: Towards general physical spatial intelligence in vision-language models. *arXiv preprint arXiv:2608.01899*, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, June 2025.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *International Conference on Learning Representations*, volume 2026, pp. 78185–78225, 2026.

Chun-Hsiao Yeh, Shengyi Qian, Manchen Wang, Yi Ma, Joseph Tighe, and Fanyi Xiao. Beyond 3d vqas: Injecting 3d spatial priors into vision-language models for enhanced geometric reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16811–16822, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhao_SpaceMind_Camera-Guided_Modality_Fusion_for_Spatial_Reasoning_in_Vision-Language_Models_CVPR_2026_paper.html.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. In *Advances in Neural Information Processing Systems*, 2025.

# APPENDIX

## .1 INFERENCE PROTOCOL

GeoLift is evaluated as a frozen inference procedure. The Video MLLM and the geometry models are loaded from their released checkpoints, and no parameter is updated using either benchmark. The Video MLLM receives the query together with the sampled RGB observations and is used in two places: it first produces open-vocabulary 2D observations, and it may later provide the final answer when the geometric evidence is incomplete or the query is not directly executable from the map. This reuse keeps the semantic interface identical to the direct baseline. Given a sampled clip, the geometric representation is constructed as part of the inference procedure before the query-specific spatial solver is applied.

The settings below specify the reference inference configuration and are shared across scenes within each task branch. The additional experiments in Section D vary the stated parameters or depth inputs while holding the remaining configuration fixed. These diagnostic sweeps do not introduce any neural weight updates.

## .2 GROUNDING AND OBSERVATION FILTERING

For each query, the Video MLLM parses the landmark names and returns, for each candidate observation, a frame index, a normalized bounding box, a confidence score, and, when useful for navigation, a point on the object. Frames are examined in contact sheets of eight images. A contact-sheet pass retains at most two observations per landmark, after which the six strongest candidates are refined on their original frames. If a required landmark is not found, a dense retry over 64 uniformly sampled frames is used. The grounding response is required to be JSON so that the subsequent geometric stages do not depend on free-form parsing.

Boxes with confidence below $0.25$, width or height below $5/1000$ of the image, or area larger than $0.60$ of the image are discarded. Before lifting, each retained box is eroded by $8\%$ of its width and height to reduce depth bleed from object boundaries. At least 30 valid lifted points are required for an observation to participate in superpoint voting; large observations are uniformly subsampled to at most 2,000 points. Depth samples are retained only when they are finite, lie in the valid metric range, and are within $\max(0.20\,\mathrm{m}, 0.20\,\tilde{z})$ of the observation's median depth $\tilde{z}$.

## .3 METRIC RECONSTRUCTION AND POINT-CLOUD FUSION

VGGT-Omega produces a per-frame depth map $d_t$, camera intrinsics $K_t$, and camera-to-world pose $(R_t, t_t)$ in an arbitrary reconstruction scale. MoGe-2 provides a metric depth estimate $m_t$ for the same frames. We estimate one positive scale from valid, confidence-filtered pixel pairs by the fixed least-squares relation

$$s^* = \arg\min_{s>0} \sum_t \sum_{u\in\Omega_t} \big(s\, d_t(u) - m_t(u)\big)^2, \qquad d_t^{\mathrm{m}} = s^* d_t, \tag{4}$$

where $\Omega_t$ retains valid depth pairs after confidence filtering and trimming extreme metric-to-reconstruction depth ratios. Confidence is used to select pairs; the retained pairs enter the least-squares fit with equal weight. The closed-form solution is $s^* = \sum dm / \sum d^2$. The ratio quantile bounds control outlier filtering, rather than clipping the estimated scale. Applying this scale to the camera translations gives

$$t_t^{\mathrm{m}} = s^* t_t, \qquad R_t^{\mathrm{m}} = R_t, \qquad K_t^{\mathrm{m}} = K_t; \tag{5}$$

rotations and intrinsics are therefore unchanged. We estimate a single global scale factor by least-squares fitting the VGGT-Omega depths to the metric depths predicted by MoGe-2, using confidence-filtered valid depth pairs within 65 m; MoGe-2 is run at resolution level 7.

For a pixel $u = (x, y)$ with scaled depth $z$, the corresponding camera-frame point and world point are

$$X_t(u) = R_t^{\mathrm{m}} \left[ z\frac{x - c_{x,t}}{f_{x,t}}, z\frac{y - c_{y,t}}{f_{y,t}}, z \right]^\top + t_t^{\mathrm{m}}. \tag{6}$$

For cloud fusion, the production superpoint path retains pixels above the 10th percentile of the VGGT confidence values. We then unproject valid pixels from all frames, retain points with depth below

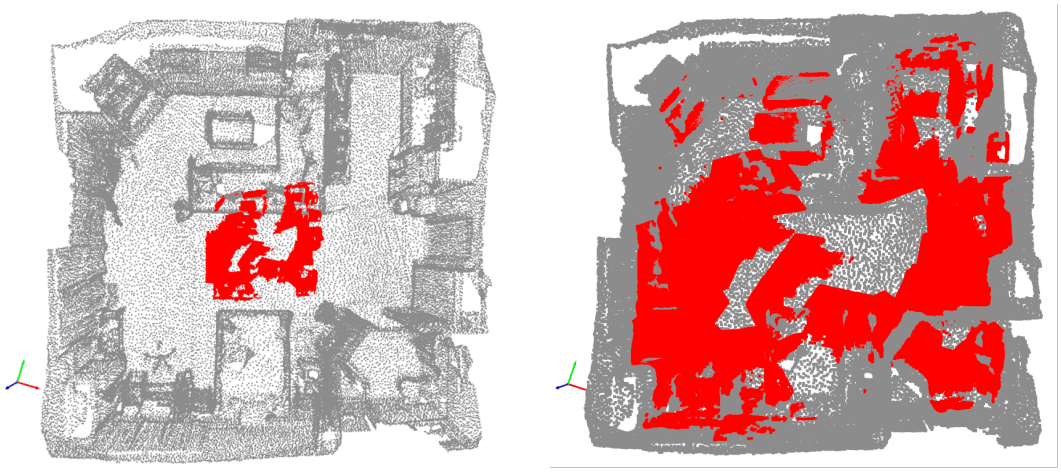


Figure 4: **Metric-scale calibration.** Left: VGGT-Omega reconstruction before metric calibration, overlaid with the reference scene. Right: the same reconstruction after the global scale estimated from MoGe-2 is applied to both depth and camera translation. Gray denotes the reference cloud and red denotes the reconstructed cloud; the calibrated reconstruction matches the reference extent substantially better.

Table 5: Fixed reconstruction and grounding configuration used throughout evaluation.

| Component | Fixed setting |
|---|---|
| Video sampling | uniform clip sampling; geometry stride 3, at least 8 frames |
| VGGT-Omega | resolution 256, patch size 16, latter-12 feature preset |
| MoGe-2 | resolution level 7; depth-based global scale |
| Scale filtering | confidence percentile 50; calibration depth $\leq 65$ m |
| Grounding contact sheet | 8 frames per sheet; at most 2 candidates per sheet |
| Box filtering | confidence $\geq 0.25$; minimum side $5/1000$; area $\leq 0.60$ |
| Box lifting | erosion $0.08$; minimum 30 points; maximum 2,000 samples |
| Cloud fusion | pixel step 2; confidence percentile 10; depth $\leq 10$ m; voxel size $0.02$ m |
| Outlier removal | radius $0.04$ m; minimum 3 neighbors |

$10\,\mathrm{m}$ for the dense planning cloud, and downsample by $0.02\,\mathrm{m}$ voxels. A radius-outlier pass removes a point unless it has at least three neighbors within $0.04\,\mathrm{m}$. The retained point cloud stores its RGB value, source-frame count, and frame provenance for later visibility checks.

Figure 4 visualizes the role of metric calibration on the same scene. The gray points are the reference reconstruction and the red points are the VGGT-Omega reconstruction after rigid registration for comparison. Before calibration, VGGT-Omega recovers the scene layout but its arbitrary global scale produces a visibly contracted spatial extent. After the MoGe-2 scale is applied jointly to the depth maps and camera translations, the reconstructed extent agrees with the reference. For the illustrative ScanNet scene `scene0000_01`, an independent Umeyama Sim3 alignment of 32 paired predicted and ground-truth camera trajectories gives a residual scale error of 1.57%, rounded to 1.6%.

# A Superpoint Mapping and Association

## A.1 Superpoint Partition

The fused cloud is partitioned into geometric superpoints by a symmetric, deduplicated $k$-nearest-neighbor graph. Each point is represented by local geometric features computed from its neighborhood: eigenvalue-based linearity, planarity, scattering, local scale and curvature statistics, verticality, and an oriented surface normal. Normals are made sign-consistent within the cloud. We then solve a cut-pursuit piecewise-constant partition problem over the graph. The production configuration uses $k = 12$, spatial-coordinate weight $5.0$, regularization $0.6$, and at most 15 cut-pursuit iterations. RGB

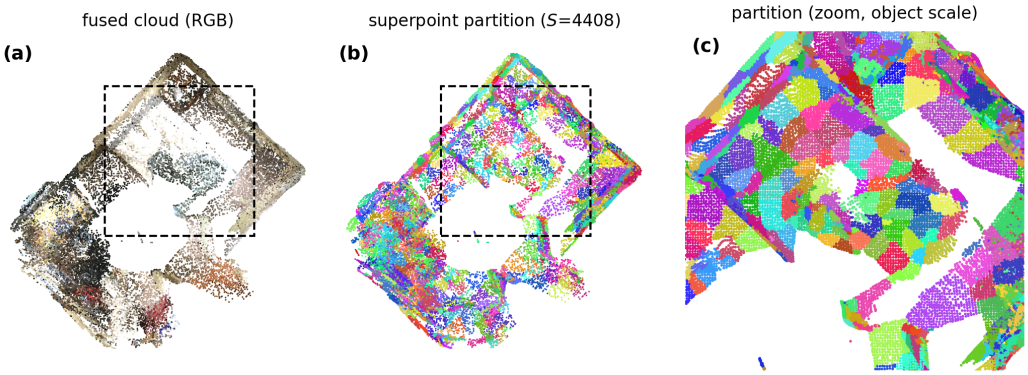



Figure 5: **Geometry-only superpoint partition.** (a) Fused RGB point cloud. (b) The same cloud colored by superpoint assignment. (c) An object-scale crop showing locally coherent geometric regions. Colors denote superpoint identities rather than semantic classes.

or Lab features are disabled in this configuration, so the partition is determined by geometry rather than by appearance or a learned descriptor. Figure 5 visualizes the resulting decomposition at both scene and object scales.

### A.2 Observation Voting and Cross-View Association

An eroded 2D observation is unprojected into the metric cloud. Each lifted point votes for its nearest cloud point within $0.05\,\mathrm{m}$, and the corresponding superpoint vote is kept only when it reaches at least two votes and at least $15\%$ of the dominant vote count. For route association, a one-hop dilation adds half weight to adjacent superpoints to connect visible fragments. The evaluated route configuration disables the optional surface-normal gate on this dilation.

Let $c_{i,s}$ be the vote count of observation $i$ for superpoint $s$, and let $\mathrm{df}(s)$ be the number of observations that vote for $s$. We use the fixed inverse-document-frequency weight

$$w_{i,s} = c_{i,s}\left(\log\frac{1+N}{1+\mathrm{df}(s)}+1\right), \tag{7}$$

where $N$ is the number of observations for the query. Two observations from different frames are linked when the cosine similarity of their weighted vote vectors is at least $\tau = 0.3$ and their vote-weighted 3D centers satisfy the center gate. The base gate is $0.5\,\mathrm{m}$; for extended instances it scales with the mean vote radius and is clipped at $1.5\,\mathrm{m}$. A union-find procedure propagates accepted links, while rejecting a union that would move the running component center beyond the same gate. Same-frame observations are never linked.

For each queried entity, the component with the greatest observation support (with vote mass as the tie-breaker) defines the view-consistent track, its 3D anchor, and its spatial footprint. A second supported but spatially separated component is retained as a conflict flag. If an observation has insufficient superpoint evidence, the system uses a robust bounding-box depth center and marks the resulting anchor as unverified. These rules make association an inference-only geometric operation; no learned matcher or benchmark-specific association training is used.

Relative-direction questions use a complementary consensus filter. For each landmark, we sum the per-view superpoint vote vectors and retain views whose cosine similarity to this consensus is at least $\theta = 0.1$. The retained bounding-box depth centers are averaged to obtain the landmark anchor. If no view passes the filter, the available per-view centers are averaged. This view-to-consensus threshold differs from the pairwise association threshold $\tau$ used by the route branch. Figure 6 illustrates how query-conditioned observations vote over superpoint identities and form a cross-view consensus for relative-direction reasoning.

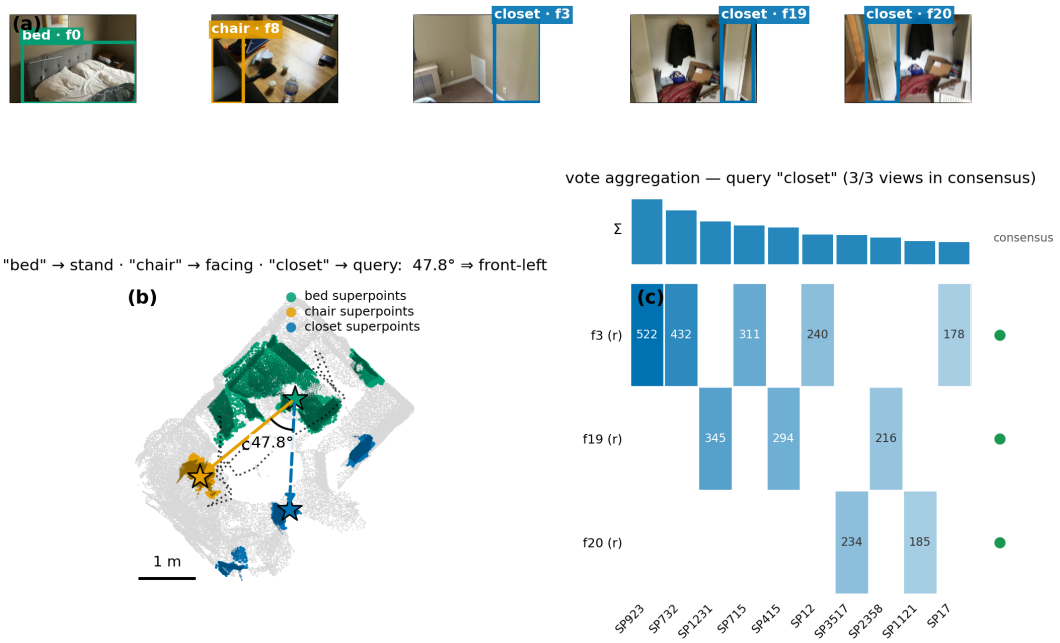



Figure 6: **Cross-view superpoint voting and aggregation.** (a) Query-conditioned 2D observations from sampled frames. (b) Lifted superpoints and recovered object anchors used for relative-direction reasoning. (c) Per-view votes over superpoint identities and their aggregated consensus; numbers indicate retained vote counts, and green markers denote views retained by the consensus filter.

Table 6: Fixed superpoint and association configuration.

| Component | Fixed setting |
|---|---|
| Superpoint graph | $k = 12$ symmetric neighbors |
| Cut-pursuit | spatial weight 5.0; regularization 0.6; 15 iterations |
| Appearance features | disabled (color weight 0) |
| Lifted-point assignment | nearest-cloud radius 0.05 m |
| Vote retention | at least 2 votes and 15% of the maximum |
| Route surface dilation | one hop; weight 0.5; normal gate disabled |
| Route association | IDF-cosine threshold $\tau = 0.3$ |
| Direction consensus | view-to-consensus cosine threshold $\theta = 0.1$ |
| Center gate | base 0.5 m; adaptive cap 1.5 m |

# B Spatial Solvers and Reliability Gates

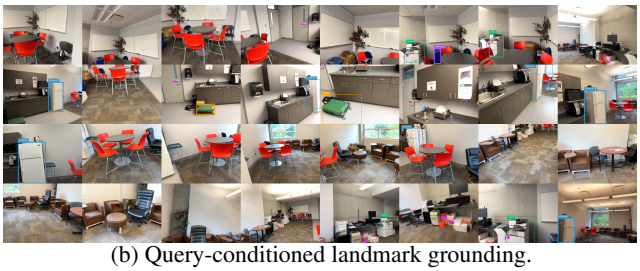
## B.1 Route Planning

For route questions, the recovered points are projected onto a robust RANSAC ground plane. The plane is fitted with 200 samples, an inlier band of one grid cell, and a normal-alignment requirement of 0.57 with the camera-down direction. Obstacles are formed from reconstructed points and grounded object footprints. We rasterize the plane at 0.10 m resolution, dilate observed floor support by 0.50 m and the camera trajectory corridor by 0.40 m, and use a 0.15 m robot radius. Up to 512 visibility rays per frame are used to mark observed free space.

Object centers, or a grounded object-edge point when the query specifies one, are snapped to the nearest free cell within 0.80 m. An 8-connected A* search then connects the ordered route anchors; diagonal corner cutting is forbidden. The resulting path is converted into left, right, forward, or turn-back commands. A turn is recorded only above 15°, and a direct segment is used to disambiguate turns below 45°; angles above 150° are treated as turn-backs.



(a) Sampled video frames.



(b) Query-conditioned landmark grounding.




(c) Occupancy-grid route and ordered landmarks.

Figure 7: **Qualitative route-planning example.** The query asks the robot to navigate from the refrigerator to the recycling bin while passing the suitcase and printer. The three panels show the sampled observations, the grounded landmarks, and the corresponding ordered route evidence. The predicted option is A (turn left, turn left, turn right), matching the ground truth.

The geometry answer is certified only when the planned answer is stable under the fixed perturbation check using robot radii $0.10$, $0.15$, and $0.25\,\mathrm{m}$ and tangent windows of 6, 12, and 18 path cells. The option margin must also be at least $10°$. If A* fails, the perturbation check is unstable, or the geometry and Video MLLM answers disagree under an unclear margin, GeoLift falls back to the direct Video MLLM path.

### B.2 Qualitative Route-Planning Example

Figure 7 shows a correctly answered three-turn VSI-Bench example (question id 5110, ScanNet++ scene `5f99900f09`). The robot starts at the refrigerator while facing it, passes the suitcase and the printer, and terminates at the recycling bin. The grounding visualization shows that the four query-relevant landmarks are localized across the sampled video frames, while the route visualization organizes the corresponding scene evidence into three consecutive segments: refrigerator $\rightarrow$ suitcase, suitcase $\rightarrow$ printer, and printer $\rightarrow$ recycling bin. The resulting command sequence is "turn left, turn left, turn right," i.e., option A, which agrees with the ground truth. The reconstructed 3D map and the A* planner provide the ordered route evidence as a geometric hint. The frozen Video MLLM combines this evidence with the original video and produces the final answer. If the geometric evidence fails the reliability checks, GeoLift falls back to the direct Video MLLM prediction.

### B.3 Relations and Distance Comparisons

For relative direction, the consensus-filtered object anchors are projected onto the recovered ground plane and compared with the estimated standing-to-facing heading. The geometric solver requires usable anchors for the queried entities; otherwise the direct Video MLLM answer is retained. Relative-distance questions use the closest supported surface points when spatial footprints are available, and otherwise use distances between the reconstructed object anchors. The resulting geometric measurements are provided to the frozen Video MLLM for final answer prediction.

### B.4 Metric Estimation and Video MLLM Hints

Absolute-distance questions use the recovered metric anchors and camera-motion evidence as structured numerical hints, while the Video MLLM produces the final numerical readout. Room-size questions use the reconstructed floor footprint and its metric extent as a numerical hint. When this

estimate is unavailable, the system uses the camera-trajectory hull or falls back to the direct visual prompt. Object counting and object-size estimation benefit from the view-consistent object tracks and reconstructed spatial extents, respectively. For appearance-order questions, frame provenance retained by the reconstructed object tracks provides additional cross-view evidence. The frozen Video MLLM integrates these geometric cues with the original visual observations to produce the final answer.

### B.5 Sparse-View MindCube

MindCube supplies two to four sparse views rather than a long egocentric clip. We run the same frozen Video MLLM grounding and VGGT-Omega/MoGe-2 geometry components on the available views. Viewpoint relations use camera poses and object bearings; an entity's 3D point is the median of its valid per-view MoGe lifts. The video superpoint association is not invoked for this sparse view branch. The reported MindCube result uses the official three settings, with "aroundnew" included in "around", and computes the overall value as micro accuracy over all 21,154 questions. A direct Video MLLM answer is used when the sparse evidence does not support a reliable geometric decision.

## C Evaluation Protocol

VSI-Bench is evaluated with its official metrics: threshold mean relative accuracy for numerical categories and exact-match accuracy for multiple-choice categories. The reported overall score is the unweighted average of the eight category scores. MindCube is evaluated with the official Rotation, Among, and Around settings and the micro overall accuracy described above. All reported baselines retain their published backbone and protocol, whereas GeoLift uses the frozen Qwen3-VL-4B-Instruct Video MLLM and the fixed geometry pipeline described here.

## D Additional Experiments

The main-paper ablation evaluates the contributions of metric geometry and superpoint association. Here we examine two complementary questions: how performance varies with the superpoint parameters, and how errors in the reconstructed depth affect downstream reasoning. All neural components remain frozen, and every tested value in the sweeps described below is reported.

**Diagnostic scoring.** The additional experiments use a fixed scoring protocol within each sweep. Route planning and relative direction use pooled question-level accuracy over 194 and 968 questions, respectively. For relative direction, this differs from averaging the three difficulty-level accuracies in the main-paper evaluation. Numerical results use the official VSI-Bench MRA, which averages relative-error correctness over 10 thresholds from 0.50 to 0.95 with an interval of 0.05, consistent with the main-paper evaluation protocol. All tabulated scores are percentages rounded to one decimal place, and differences are calculated before rounding.

### D.1 Hyperparameter Sensitivity

**Design and coverage.** We vary one parameter at a time while keeping the upstream grounding, reconstruction, and remaining inference settings fixed. The partition regularization $\lambda \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ controls superpoint granularity and is evaluated on route planning and relative direction. The pairwise association threshold $\tau \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$ is evaluated on route planning and absolute distance. Relative direction instead uses the view-to-consensus threshold $\theta \in \{0.05, 0.10, 0.20, 0.30, 0.40\}$; these are separate parameters and are reported on their own scales. The reference settings are $\lambda = 0.6$, $\tau = 0.3$, and $\theta = 0.1$. The route and direction reference configurations reproduce their respective unmodified diagnostic predictions with 100% agreement.

The absolute-distance threshold sweep covers a fixed subset of 802 of the 834 questions. The remaining 32 questions are excluded because the required scene partitions were unavailable. All five settings use the same subset. We report both the full 802-question subset and the 748-question subset with available geometric evidence, rather than restricting the main comparison to evidence-supported questions. The conclusions for this sweep are limited to this coverage. These experiments do not include a joint parameter grid, a room-size parameter sweep, or a MindCube sensitivity study.

Table 7: **Superpoint parameter sensitivity.** All five tested values are shown for each parameter. Shading identifies the reference configuration. Geometry use is the percentage of questions whose final answer is supplied by the geometric solver. Absolute-distance results use diagnostic MRA; route and direction results use pooled accuracy. W/L counts questions whose MRA improves/decreases relative to $\tau = 0.3$ on the 802-question subset; ties are omitted.

(a) Partition regularization; $\tau = 0.3$, $\theta = 0.1$ fixed

| $\lambda$ | Route acc. | Geometry use | Rel. dir. acc. | Geometry use |
|---|---|---|---|---|
| 0.2 | 51.6 | 87.1 | 73.4 | 98.3 |
| 0.4 | 50.5 | 88.7 | 73.5 | 98.3 |
| 0.6 | 53.6 | 94.3 | 73.5 | 98.3 |
| 0.8 | 49.5 | 91.2 | 73.4 | 98.3 |
| 1.0 | 52.1 | 90.7 | 73.2 | 98.3 |

(b) Pairwise association threshold; $\lambda = 0.6$ fixed

| $\tau$ | Route acc. | Geometry use | Abs. dist. (802) | Abs. dist. (748) | W/L |
|---|---|---|---|---|---|
| 0.2 | 55.7 | 93.8 | 49.3 | 49.3 | 10/16 |
| 0.3 | 53.6 | 94.3 | 49.4 | 49.4 | – |
| 0.4 | 53.1 | 94.8 | 49.3 | 49.3 | 11/16 |
| 0.5 | 52.1 | 95.4 | 49.3 | 49.3 | 34/32 |
| 0.6 | 51.6 | 95.4 | 49.4 | 49.4 | 49/39 |

(c) View-to-consensus threshold; $\lambda = 0.6$ fixed

| $\theta$ | Rel. dir. acc. | Geometry use |
|---|---|---|
| 0.05 | 73.0 | 98.3 |
| 0.10 | 73.5 | 98.3 |
| 0.20 | 73.2 | 98.3 |
| 0.30 | 71.3 | 98.3 |
| 0.40 | 69.4 | 98.3 |

**Results.** Table 7 shows that sensitivity depends on both the parameter and the task. Across the partition sweep, relative-direction accuracy changes by less than 0.3 percentage points, whereas route accuracy ranges from 49.5% to 53.6%. All tested route settings remain above the 46.9% no-superpoint result in the main ablation. Loosening the route association threshold to $\tau = 0.2$ improves accuracy to 55.7%, 2.1 points above the reference; the reference is therefore not the best setting in this sweep. Relative direction is comparatively stable for $\theta \in \{0.05, 0.10, 0.20\}$, but falls to 69.4% at $\theta = 0.40$, consistent with a cost to discarding too many views. On the absolute-distance subset, the largest MRA difference from the reference is less than 0.2 points. Together, these results support stability in a neighborhood of the reference settings, while also identifying sensitivity to stricter association filters.

Geometry use remains between 87.1% and 95.4% for route planning and at 98.3% for relative direction. The observed performance is thus not accompanied by a wholesale switch to direct visual answers, although usage alone does not establish that the geometric answers are correct. One changed answer corresponds to approximately 0.5 points on route planning and 0.1 points on pooled relative-direction accuracy. We report descriptive differences without assigning statistical significance to small changes.

### D.2 Sensitivity to Depth Perturbations

**Perturbations and task coverage.** We perturb calibrated depth before point-cloud fusion and recompute the downstream geometric representation. The two perturbation families are

$$\tilde{d}_f^{\text{pixel}}(u) = d_f^{\text{m}}(u)\exp(\sigma\epsilon_{f,u}), \qquad \tilde{d}_f^{\text{frame}}(u) = d_f^{\text{m}}(u)\exp(\sigma\eta_f), \tag{8}$$

where $\epsilon_{f,u}$ are independent standard normal samples for each pixel and frame, and $\eta_f$ is one standard normal sample shared by all pixels in frame $f$. We test $\sigma \in \{0.02, 0.05, 0.10, 0.20\}$ and include $\sigma = 0$ as the unperturbed reference. Camera poses, intrinsics, grounding, and the reference superpoint parameters are held fixed for these two families. Within each scene and perturbation family, all

strengths share one underlying noise realization, seeded by the scene identifier and family. These curves represent one realization per scene and family, not averages over repeated random seeds. The log-depth standard deviation $\sigma$ does not specify a bounded percentage error or an empirically calibrated VGGT-Omega error level.

The experiment covers all four tasks that consume reconstructed depth in the evaluated pipeline: route planning ($N = 194$), relative direction ($N = 968$), absolute distance ($N = 834$), and room size ($N = 288$). Object counting, object size, relative distance, and appearance order use branches without depth inputs and are not rerun under this intervention. We do not count their unchanged inputs as additional measured robustness results. Route and direction answers are recomputed with fixed upstream visual observations and fallback answers; for absolute distance and room size, the frozen Video MLLM is evaluated with the corresponding recomputed geometric hints.

**Within-experiment references.** Absolute-distance hints use a reconstructed evidence-generation routine for both the unperturbed and perturbed conditions. It uses closest bounding-box surface distance, with center distance as a fallback for overlapping boxes. This reconstruction of the evidence yields an unperturbed diagnostic MRA of 49.3%, compared with 49.1% in the earlier diagnostic run. Room-size evaluation preserves the floor-hull, camera-hull, and direct-answer fallback chain. Its unperturbed rerun scores 66.2%, compared with 66.2% in the earlier diagnostic run, and matches 258 of 288 earlier predictions. The rerun used a different video-decoding/pixel-budget path, so it is not an exact reproduction of those earlier answers. All perturbation differences below use the unperturbed reference from the same experiment, rather than the earlier diagnostic or main-paper scores.

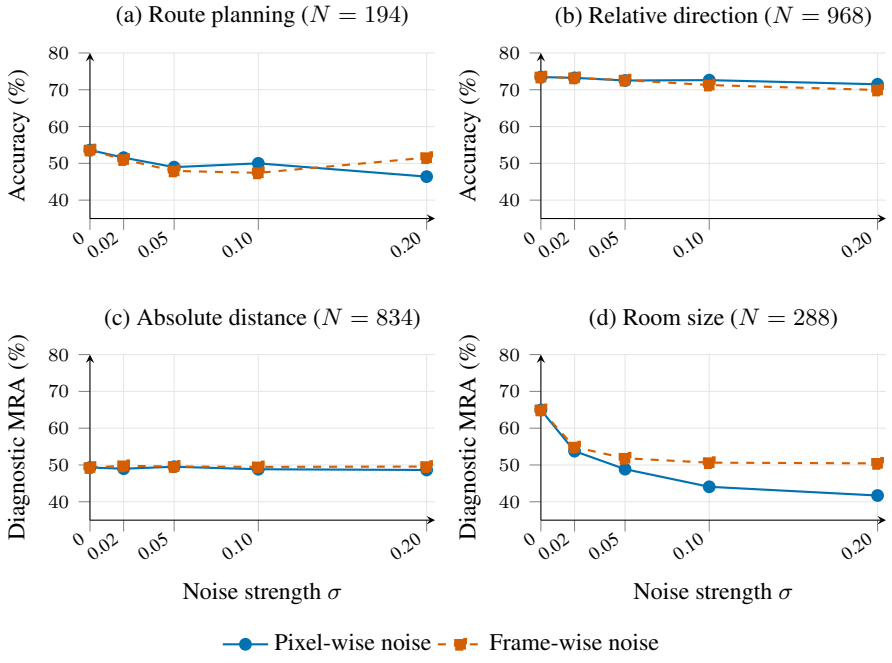



Figure 8: **Sensitivity to depth perturbations across all four depth-dependent task branches.** Both perturbation families are evaluated at every noise strength using one fixed realization per scene. Shared vertical limits preserve relative score changes; metrics differ as labeled. Lines connect measured points without smoothing; Table 8 reports all values.

Table 8: **Complete depth-perturbation results.** All four tested strengths are reported for both noise families. Route and relative direction use accuracy; absolute distance and room size use diagnostic MRA. Every column retains the same question set across conditions. The unperturbed row is shared by both families.

| Perturbation | $\sigma$ | Route | Rel. dir. | Abs. dist. | Room size |
|---|---|---|---|---|---|
| Unperturbed | 0 | 53.6 | 73.5 | 49.3 | 66.2 |
| Pixel-wise | 0.02 | 51.6 | 73.2 | 49.0 | 53.8 |
| Pixel-wise | 0.05 | 49.0 | 72.5 | 49.5 | 48.9 |
| Pixel-wise | 0.10 | 50.0 | 72.6 | 48.8 | 44.1 |
| Pixel-wise | 0.20 | 46.4 | 71.5 | 48.6 | 41.7 |
| Frame-wise | 0.02 | 51.0 | 73.2 | 49.8 | 54.9 |
| Frame-wise | 0.05 | 47.9 | 72.6 | 49.6 | 51.8 |
| Frame-wise | 0.10 | 47.4 | 71.3 | 49.5 | 50.6 |
| Frame-wise | 0.20 | 51.6 | 69.9 | 49.6 | 50.4 |

**Results and interpretation.** Figure 8 and Table 8 show task-dependent sensitivity. At $\sigma = 0.20$, pixel-wise perturbations reduce route and relative-direction accuracy by 7.2 and 2.0 points, respectively. Frame-wise perturbations cause a larger direction decrease of 3.5 points, consistent with sensitivity to disagreement between views. The route curve is non-monotonic: under frame-wise perturbations, accuracy increases from 47.4% at $\sigma = 0.10$ to 51.6% at $\sigma = 0.20$. Absolute-distance MRA varies by less than one point across the tested noise conditions. This is an observation about the complete evidence-guided answering branch, not a claim that its underlying geometric distances are unchanged.

Room-size estimation is the most sensitive branch: pixel-wise and frame-wise perturbations at $\sigma = 0.20$ reduce MRA by 24.5 and 15.8 points, respectively. Even at $\sigma = 0.02$, the decreases are 12.4 and 11.3 points. This branch depends on a reconstructed floor-area hint, making the result consistent with sensitivity to changes in the recovered floor support and its convex hull. The experiment does not isolate which geometric operation causes the decline: both floor selection and its spatial extent can change under these perturbations.

The diagnostic records also report geometry-use rates of approximately 85–98% across the route and direction noise conditions. Sustained use shows that the geometric branch remains active; it does not demonstrate that the reliability checks detect all perturbation-induced errors. Likewise, the small absolute-distance score changes should not be interpreted as immunity to other reconstruction errors, occlusions, or grounding failures.

**Removing metric calibration.** We separately replace calibrated depths and camera translations with their uncalibrated VGGT-Omega counterparts and recompute the downstream geometry, while leaving other settings unchanged. This is a component ablation, rather than another noise-strength setting, and is therefore reported separately in Table 9.

Table 9: **Metric-scale calibration under the diagnostic protocol.** Both conditions use the same question sets and scorer as the depth experiment. Removing calibration reverts both depth and camera translation; fixed distance thresholds and planner settings are retained.

| Configuration | Route | Rel. dir. | Abs. dist. | Room size |
|---|---|---|---|---|
| With calibration | 53.6 | 73.5 | 49.3 | 66.2 |
| Without calibration | 47.9 | 73.8 | 48.4 | 36.1 |
| Change (points) | $-5.7$ | $+0.3$ | $-0.9$ | $-30.1$ |

Removing calibration reduces route accuracy by 5.7 points and room-size MRA by 30.1 points, while relative direction changes by $+0.3$ points and absolute-distance MRA by $-0.9$ points. These results measure the role of metric scale in the complete pipeline, including fixed metric thresholds, grid resolution, and robot size. They do not isolate a change in landmark coordinates alone. Together, the

diagnostic experiments identify a useful range of superpoint settings and a pronounced dependence of floor-area estimation on reconstruction quality.