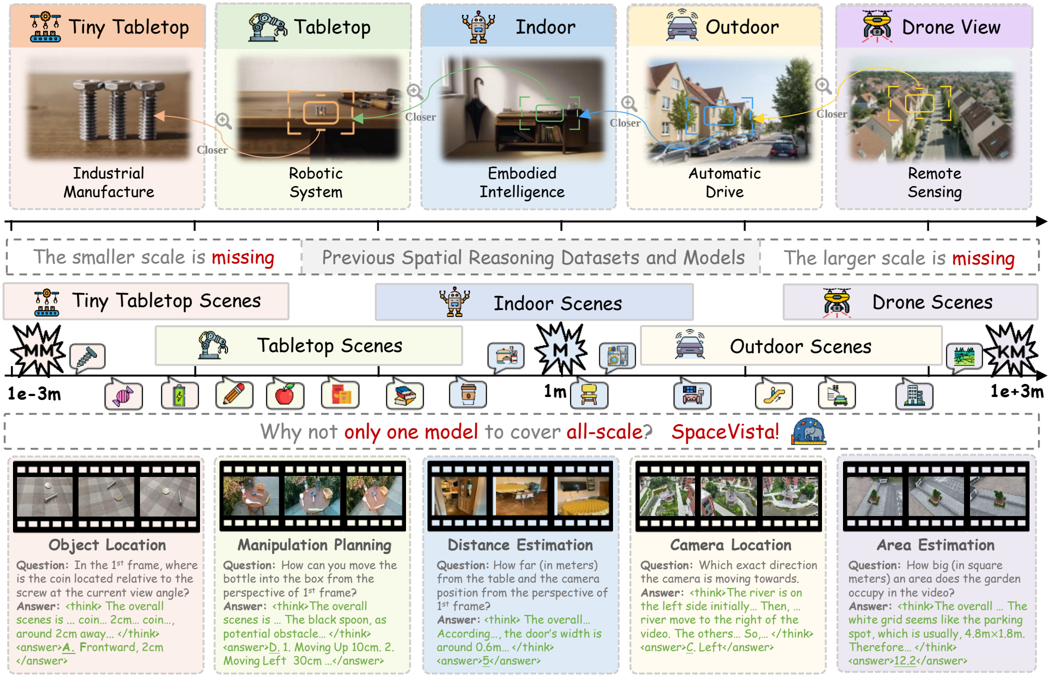
SPACEVISTA: ALL-SCALE VISUAL SPATIAL REASONING FROM mm TO km

Anonymous authors
Paper under double-blind review



Question: In the 1st frame, where is the coin located relative to the screw at the current view angle?
Answer: <think> The overall scenes is ... coin... 2cm... coin..., around 2cm away... </think> <answer>A. Frontward, 2cm </answer>

Question: How can you move the bottle into the box from the perspective of 1st frame?
Answer: <think>The overall scenes is ... The black spoon, as potential obstacle... </think> <answer>D. 1. Moving Up 10cm. 2. Moving Left 30cm ...</answer>

Question: How far (in meters) from the table and the camera position from the perspective of 1st frame?
Answer: <think> The overall... According..., the door's width is around 0.6m... </think> <answer>5</answer>

Question: Which exact direction the camera is moving towards.
Answer: <think>The river is on the left side initially... Then, ... river move to the right of the video. The others... So,... </think> <answer>C. Left</answer>

Question: How big (in square meters) an area does the garden occupy in the video?
Answer: <think>The overall ... The white grid seems like the parking spot, which is usually, 4.8m×1.8m. Therefore... </think> <answer>12.2</answer>

Figure 1: Prior works of spatial reasoning have largely focused on indoor (1-30 m) scenes, while our SpaceVista model and dataset span scales from mm (1e-3 m) to km (1e+3 m). Dotted lines represent our contribution in filling the gap. This six-order-of-magnitude range introduces not only scale variation but also rich semantics and diverse tasks. SpaceVista enables all-scale spatial reasoning by integrating cues from micro-objects to macro-scenes.

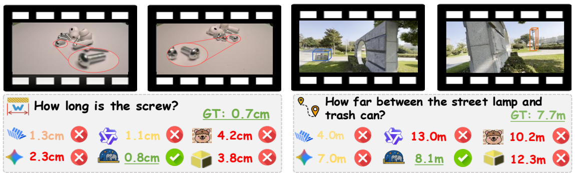
ABSTRACT

With the current surge in spatial reasoning explorations, researchers have made significant progress in understanding indoor scenes, but still struggle with diverse applications such as robotics and autonomous driving. This paper aims to advance all-scale spatial reasoning across diverse scenarios by tackling two key challenges: 1) the heavy reliance on indoor 3D scans and labor-intensive manual annotations for dataset curation; 2) the absence of effective all-scale scene modeling, which often leads to overfitting to individual scenes. In this paper, we introduce a holistic solution that integrates a structured spatial reasoning knowledge system, scale-aware modeling, and a progressive training paradigm, as the **first attempt** to broaden the all-scale spatial intelligence of MLLMs to the best of our knowledge. Using a task-specific, specialist-driven automated pipeline, we curate over 38K video scenes across 5 spatial scales to create **SpaceVista-1M**, a dataset comprising approximately 1M spatial QA pairs spanning 19 diverse task types. While specialist models can inject useful domain knowledge, they are not reliable for evaluation. We then build an all-scale benchmark with precise annotations by manually recording, retrieving, and assembling video-based data. However, naive training with SpaceVista-1M often yields suboptimal results due to the potential knowledge conflict. Accordingly, we introduce **SpaceVista-7B**, a spatial reasoning model that accepts dense inputs beyond semantics and uses scale as an anchor for scale-aware experts and progressive rewards. Finally, extensive evaluations across 5 benchmarks, including our **SpaceVista-Bench**, demonstrate competitive performance, showcasing strong generalization across all scales and scenarios. Our dataset, model, and benchmark will be released at our project page🌍.

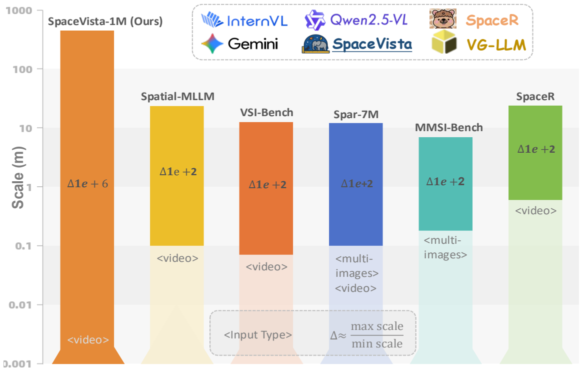
1 INTRODUCTION

Spatial reasoning, the ability to sense, interpret, and interact with environments across scales from tiny objects understanding to remote drone sensing, is crucial for next-generation intelligent systems. It significantly enhances 3D and even 4D scene understanding, enabling agents to interpret complex environments from easily obtainable videos. **All-scale reasoning** capability supports diverse applications: mm for advanced manufacturing (Song et al., 2024), cm and m for embodied intelligence (Pan et al., 2025), $10m$ for autonomous driving (Liu et al., 2022), and $100m$ for drone-based sensing (Xiao et al., 2023). Recent research (Yang et al., 2025a), especially on how Multimodal Large Language Models (MLLMs) perceive and recall space, is narrowing the gap in visual spatial reasoning.

The current works on spatial reasoning primarily focus on improvements from two perspectives: data and model. From the data perspective, pioneer works (Ouyang et al., 2025; Zhang et al., 2025e; Deng et al., 2025b) utilize more scanning-based data, or image-based data employing fully automated pipelines to acquire additional information for Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL). During modeling indoor spatial scenes, Wu et al. (2025a); Zheng et al. (2025) leverage latent features from VGGT (Wang et al., 2025a) by incorporating geometric information to enhance spatial understanding. Concurrently, a series of outstanding works (Ouyang et al., 2025; Zhang et al., 2025e) have improved the performance of existing models by refining the training and thinking approaches. Moreover, Wu et al. (2025b) employs multi-turn dialogues to enhance self-correction capabilities.



(a) Case comparison across scales on popular MLLMs



(b) Scale comparison on popular spatial datasets and benchmarks

Figure 2: (a) and (b) show model performance and dataset distribution across scales. Current models and datasets necessitate all-scale spatial reasoning.

Despite these works' advancements, their spatial perception capabilities are primarily limited to indoor settings, specific objects, and constrained scales, as shown in the the bar chart Fig.1. Moreover, current methodologies lack dedicated training frameworks for holistic all-scale scene understanding. To bridge this gap, we introduce the **first comprehensive solution** to address data, model, and evaluation dimensions for all-scale scenarios.

Previous datasets (Yang et al., 2025a;b; Ouyang et al., 2025; Zhang et al., 2025e) for spatial reasoning have primarily been constructed based on indoor scanning video data (Dai et al., 2017; Yeshwanth et al., 2023) as shown in Fig. 2(b). These indoor datasets often feature relatively simple scenes and depend on manual 3D annotations. Scaling up to build large-scale, wild datasets encompassing video scenes ranging from mm to km presents two major challenges: 1) the **high cost** of large-scale annotation from complex and wild scenes; 2) the difficulty in obtaining **precise evaluations** that align with the physical world. To address these challenges, we use an automated pipeline leveraging popular specialized models to generate structured training data across 5 different scales. Since different scales have distinct characteristics and applications, we define several scale-specific tasks for better application, i.e., manipulation planning and area estimation. Overall, we provide over 1 million QA pairs across 19 diverse tasks from around 38K wild video scenes. To adapt to different stages of training, we provide both answers with rationale for SFT and regression/multiple-choice answers for RL. To facilitate accurate evaluation, we collect a highly accurate SpaceVista-Bench through manually recording or retrieving authoritative sources, supplemented with human annotations.

Most popular reasoning models are optimized for indoor settings, which leads to clear limitations: their responses **often deviate** significantly, in tabletop and other diverse real-world scenes illustrated in Fig. 2(a). We address this by first injecting SpaceVista-1M knowledge to fine-tune existing models with the self-supervised visual encoder to make compensation for the classic semantic visual tokenizer, enabling extra geometry-based and depth-based spatial understanding. However, naive fine-tuning rarely yields optimal results, largely due to **cross-scale conflicts** between scenes and objects based on our observation. To address this, we introduce LoRA-like scale experts that cooperates with a scale router during fine-tuning. Moreover, to strengthen the model's ability to learn scale-centric spatial reasoning processes, we design a training strategy that uses scale as an anchor for progressive rewards. During evaluation, SpaceVista-7B shows superior understanding of spatial layout, size, and comparison, delivering a clear improvement on popular benchmarks and SpaceVista-Bench.

Our key contributions with this comprehensive solution are:

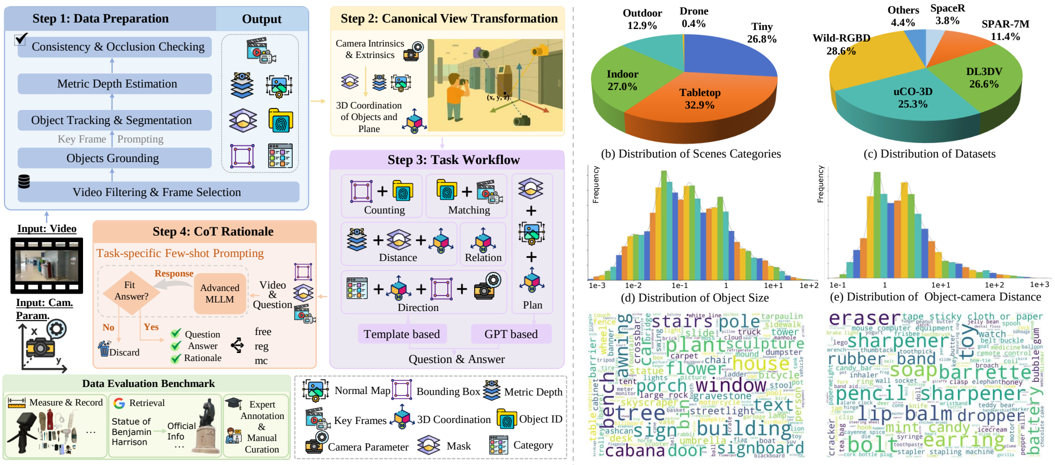
- Developing an automated pipeline to create a diverse, real-world, all-scale reasoning dataset, **SpaceVista-1M**, with 1M QA pairs across 5 scales and 19 tasks (including specific-scale tasks), and supporting both cold start with rationale and high-quality reinforced learning.
- Introducing **SpaceVista-7B**, a spatial reasoning model that integrates rich spatial information and employs scale experts with a customized training strategy to alleviate potential cross-scale conflicts during all-scale finetuning.
- Hand-crafting **SpaceVista-Bench**, an accurate video benchmark spanning all scales, by measuring and recording real-world objects, retrieving authoritative sources, and performing human annotation.

2 RELATED WORKS

Visual Reasoning. Currently, vision-based general reasoning has seen diverse developments (Tan et al., 2025; Wang et al., 2025b; Qiao et al., 2025). General MLLMs (Wang et al., 2025c; Bai et al., 2025) first provided the basic understanding ability towards video to the community. Pioneering works (Feng et al., 2025; Liao et al., 2025) started to provide reasonable rewards during model training using Group Relative Policy Optimization (GRPO) for the reasonable Chain of Thought (CoT). Then, visual reasoning (Li et al., 2025c; Chen et al., 2025; Liu et al., 2025c) was considered from broader perspectives, ranging from data to training structure. In general video reasoning, spatial claims are generally divided into two categories: 2D plane-based spatial reasoning (Han et al., 2025; Zhou et al., 2025), and 3D space-based spatial reasoning (Wu et al., 2025a; Zheng et al., 2025). This paper primarily focuses on the latter. Although these general models have achieved a certain degree of spatial ability, spatial MLLM is still in its early stages.

Spatial Reasoning. Mainstream spatial reasoning models can be categorized based on input modalities into image (Ma et al., 2025; Liu et al., 2025b; Chen et al., 2024a), multi-image (Xu et al., 2025), multi-view (Li et al., 2025b), video (Wu et al., 2025a; Zheng et al., 2025; Ouyang et al., 2025; Zhang et al., 2025b; Ghazanfari et al., 2025), and simulation (Li et al., 2025a; Tang et al., 2025; Zhang et al., 2025c; Wang et al., 2025d; Zhang et al., 2025f). Among these categories, video stands out as the challenging task due to the difficulty of data acquisition and modeling. As the first work in spatial reasoning, VSI-Bench (Yang et al., 2025a) introduced a video-based benchmark that removes linguistic shortcuts and evaluated MLLMs on spatial tasks such as counting, direction, and planning, highlighting substantial performance gaps compared to humans. InternSpatial (Deng et al., 2025b), SPAR (Zhang et al., 2025e), and SpaceR (Ouyang et al., 2025) enriched spatial supervision through extensive QA pairs spanning indoor and other limited settings. Qi et al. (2025) used the bird-view map to aid overall understanding. Then, Spatial-MLLM (Wu et al., 2025a), VG-LLM (Zheng et al., 2025), and VLM-3R (Fan et al., 2025) adopted geometry-aware dual encoders to capture geometry cues and inferred occluded structures from monocular inputs. Additionally, spatial reasoning on long (Zhang et al., 2025b), omni (Dongfang et al., 2025), ego-centric (Wu et al., 2025c) and aerial video (Zhang et al., 2025b) were also explored separately. However, the systematic data and model with all-scale video scenes remain unexplored.

All-Scale Exploration. The challenge of multi-scale in early years lay in information loss within low-resolution image patches (Zhao, 2025; Nikouei et al., 2025), which has almost no effect on spatial reasoning. In this paper, "all-scale" primarily concerns the real scales of the physical world, including distances, semantics, and object states across different scales. Deng et al. (2025a) pushed the limits of 3D perception and reconstruction from meters to kilometers; Wen et al. (2025) extended metric depth estimation from close range to infinity; and Liu et al. (2025a) curated uncommon objects, ranging from screws to airplanes, with object-centric annotations. Together, these developments underscore the need for AI to move beyond simple single-scale memorization toward robust, multiscale, and reasonable visual understanding.



(a) Data Construction Pipeline

(b) Distribution of Scenes Categories

(c) Distribution of Datasets

(d) Distribution of Object Size

(e) Distribution of Object-camera Distance

(f) Word Cloud of Outdoor Scenes

(g) Word Cloud of Tiny Tabletop Scenes

Figure 3: Fig.(a) shows our automated data construction pipeline. The pie charts (b-c) depict the composition of scenes and sources. The bar charts (d–e) show object sizes ranging mm-100m, while object-to-camera distances typically span 10-600m. Accordingly, we claim SpaceVista-1M basically covers the mm-km scale. The word clouds (f-g) provide a glimpse of the scene diversity.

3 DATASET

Due to high labeling cost, Tab.1 and Fig.2 show the clear drawback of the previous datasets. The limited data and performance constraints in existing models necessitate the creation of a dataset with all-scale spatial context. We propose **SpaceVista-1M**, a diverse, real-world, all-scale reasoning dataset, as the **first** to the best of our knowledge. SpaceVista-1M primarily comprises diverse spatial reasoning question–answer pairs, with rich semantic (category, rationale), 2D (mask, box, point), and 3D (depth, camera parameters, point cloud) annotations, obtained either natively or through processing. The construction pipeline in Fig. 3 follows the step-by-step procedure of preparing, transforming, and generating to obtain an all-scale dataset by integrating specialized models.

Data Preparation. We begin by selecting widely used video datasets that provide 3D scene modeling (Ling et al., 2024; Xia et al., 2024; Park et al., 2020; Liu et al., 2025a; Dai et al., 2017; Yeshwanth et al., 2023) along with camera intrinsic and extrinsic parameters. Most of these sources are videos of static scenes without moving objects. Leveraging the known camera parameters, we estimate depth maps and normal maps using specialized metric depth models (Hu et al., 2024; Piccinelli et al., 2025) and video depth models (Chen et al., 2025). For semantic understanding, we extract per-frame semantics and bounding boxes using proprietary grounding specialists (Ren et al., 2024; Liu et al., 2023b). To establish cross-frame object consistency, by further integrating SAM 2 (Ravi et al., 2024) with the previously mentioned grounding experts, we enable robust object ID association and mask generation. This pipeline ensures both semantic and spatial consistency across frames. Detailed preparation can be found in Appendix. B.3.1

Task Construction. With the help of official camera parameters and the preparations mentioned above, we can obtain the positions and dimensions of target objects. As a common practice (Deng et al., 2025b), we adopt a canonical view space of the reference frame, defined as a 3D Cartesian coordinate system centered at the camera's optical center. We then design 19 tasks and their corresponding workflows, even including scale-specific tasks such as tabletop object manipulation and drone-view area estimation. Taking object counting as an example, which follows: detect objects, propagate masks across frames, track identities over time, filter out scenes with camera parameters and ambiguous objects, and derive temporally consistent counts. For each task, we obtain the data by similar carefully designed computational workflows. A detailed description of each task and its workflow can be found in Appendix B.3.

QA Construction. The pipeline for constructing the QA data is shown in Fig. 3. At the construction level of QA, we employ two strategies: GPT-based and template-based. For relatively fixed questions such as counting and object size, we adopt a template-based approach to obtain reasonable QA pairs. To ensure the diversity of the questions, we manually curate over 3,000 templates. However, for more flexible questions like planning, we use a GPT-based (OpenAI, 2025a) method to generate reasonable answers in naturally language. Additionally, through appropriate randomizing and prompting, we obtain multiple options to serve as rewards for RL. QA previews and quality control can be found in Appendix F.3 and Appendix B.4.7 respectively.

CoT Annotation. To facilitate an efficient cold start, we follow Feng et al. (2025) to leverage cognition-inspired few-shot prompting strategy with Qwen2.5-VL-72B-Instruct (Bai et al., 2025) to generate CoT rationales. After employing the filtering policy for low-quality or inconsistent rationale outputs, we obtain the CoT for SpaceVista-1M, with high-quality rationale for fundamental knowledge injection for SFT.

Table 1: Comparison of popular spatial reasoning datasets. Only spatial reasoning QA is included. Lower QA/Scene Ratio usually means more diverse language and visual scenes. “free”,“reg”, and “mc” mean free-form, regression, and multiple-choice, respectively. SpaceVista-1M does not differentiate QA pairs by the type; i.e., the semantically similar questions with reg/mc/free answers are counted only once.

Usage	Dataset	Type	QA Pairs↑	Video Scenes↑	QA/Scene Ratio↓
Train	SpaceR	reg/mc	191K	1.2K	159
	SPAR-7M	reg/mc/free	7M	4.5K	1,556
	Spatial-MLLM	reg/mc/free	120K	1.5K	83
	InternSpatial	free	2.5M	5.5K	455
	SpaceVista-1M (Ours)	free/reg/mc	1M	38K	25
Benchmark	TempCompass	mc	7.5K	0.4K	18
	VideoMME	mc	2.7K	0.9K	3
	All-Angles	mc	2.1K	90	23
	VSI-Bench	reg/mc	5.0K	0.3K	17
	MMSI-Bench	mc	1.0K	-	-
	SPAR-Bench	reg/mc	7.2K	-	-
	STI-Bench	mc	2.0K	0.3K	7
	SpaceVista-Bench (Ours)	reg/mc	3K	0.5K	6

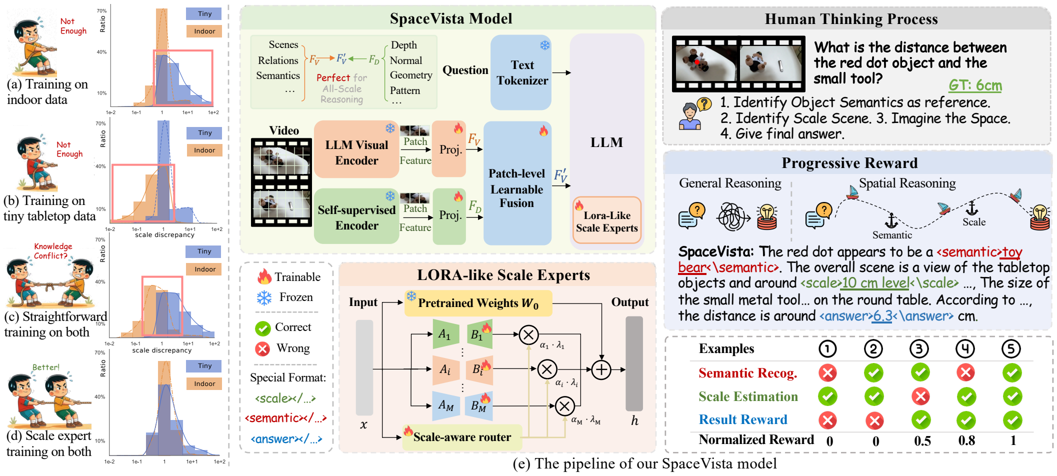
Input Extension. Usually, people refer to objects in videos using more than just language. To support this, we extend video-based QA with extra annotations from the video’s key frames. Besides plain visual input, we allow three extra inputs: point, bounding box, and mask, which may support future interactive usage. Each input type is designed to fit its own template and CoT rationales.

Quality Control & Evaluation. To ensure data quality, we conduct manual verification on a small portion training set for quality control in Appendix B.4.7. However, for measurement-related evaluation, human judgment is also susceptible to experiential bias. We choose a more reliable pathway based on measuring and recording real-world data, retrieving authoritative sources, and performing human annotation for both distance and non-distance problems, shown in the green block Fig.4(a). For tiny and tabletop scenes, we capture and annotate videos of over 50 objects of different sizes. For some indoor and outdoor scenes, we search for the landmarks and retrieve statistics from authoritative sources like Wikipedia. As for other tasks like camera moving, the experts is hired for checking and annotating. By aligning the answer with the physical world, SpaceVista-Bench comprises more than 3,000 QA pairs with 99% accuracy across 500 unique video scenes. Please refer to the details and analysis in Appendix B.2.7.

In summary, we propose SpaceVista-1M, an open-source, real-world, all-scale dataset with spatial video QA. SpaceVista-1M contains 1 million QA pairs spanning 19 tasks, 5 scale types, and over 50 subscene categories. Additionally, we encourage readers to consult the appendix, which presents meticulous source investigations (Sec. B.2), systematic processing procedures (Sec. B.3), in-depth distribution analyses (Sec. B.4), and also licensing (Sec. B.4.8).

4 METHOD

Overview. Our objective is to enhance spatial reasoning by elaborately designing and conditioning the model on explicit and detailed **all-scale information**. We first utilize a dense, expressive self-supervised encoder beyond semantics to strengthen the model’s overall spatial perception. However, mixing different types of knowledge without distinction hinders, rather than facilitates the model’s reasoning in Fig. 4(a-d), a problem known as **knowledge conflict**. In all-scale reasoning, this conflict appears when similar visual patterns are interpreted differently at different scales. To mitigate such conflict, we propose a LoRA-like scale expert architecture to maintain the independence of scale-level knowledge, while maintaining parameter efficiency, as shown in Fig 4(e). Finally, drawing on human reasoning about scale, we introduce reward-based progressive reasoning paths that employ essential anchors to constrain the reasoning process to a reliable CoT path.


Not Enough
(a) Training on indoor data
Not Enough
(b) Training on tiny tabletop data
Knowledge Conflict?
(c) Straightforward training on both
Better!
(d) Scale expert training on both
Ratio
70%
40%
10%
Tiny
Indoor
1e-2
1e-1
1
1e+1
1e+2
scale discrepancy
SpaceVista Model
Scenes
Relations
Semantics
…
F_V
F'_V
F_D
Depth
Normal
Geometry
Pattern
…
Perfect for All-Scale Reasoning
Question
Text Tokenizer
Video
LLM Visual Encoder
Self-supervised Encoder
Patch Feature
Proj.
F_V
F_D
Patch-level Learnable Fusion
F'_V
LLM
Lora-Like Scale Experts
Human Thinking Process
What is the distance between the red dot object and the small tool?
GT: 6cm
1. Identify Object Semantics as reference.
2. Identify Scale Scene. 3. Imagine the Space.
4. Give final answer.
Progressive Reward
General Reasoning
Spatial Reasoning
Semantic
Scale
SpaceVista: The red dot appears to be a <semantic>toy bear<\semantic>. The overall scene is a view of the tabletop objects and around <scale>10 cm level<\scale> …, The size of the small metal tool… on the round table. According to …, the distance is around <answer>6.3<\answer> cm.
Trainable
Frozen
Correct
Wrong
Special Format:
<scale></...>
<semantic></...>
<answer></...>
LORA-like Scale Experts
Input
Pretrained Weights W_0
Output
A_1
B_1
A_i
B_i
A_M
B_M
$\alpha_1 \cdot \lambda_1$
$\alpha_i \cdot \lambda_i$
$\alpha_M \cdot \lambda_M$
x
h
Scale-aware router
Examples ① ② ③ ④ ⑤
Semantic Recog. ✗ ✓ ✓ ✗ ✓
Scale Estimation ✓ ✓ ✗ ✓ ✓
Result Reward ✗ ✗ ✓ ✓ ✓
Normalized Reward 0 0 0.5 0.8 1

(e) The pipeline of our SpaceVista model

Figure 4: The left part (a-d) shows that the undifferentiated mixture of cross-scale knowledge hinders, rather than facilitates, the model's reasoning process. The horizontal axis represents the scale discrepancy, defined as $\frac{answer}{gt}$ (=1 for the ideal situation), and the vertical axis denotes the proportion of answers. Fig.(e) is our SpaceVista model, where "`<think>`" is omitted for clarity.

Preliminaries. The number of frames is first denoted as T with the temporal patch size τ. The visual representations from Qwen-2.5-VL visual encoder are denoted as $F_V \in \mathbb{R}^{t \times d_V \times H \times W}$, where $t = \frac{T}{\tau}$ is temporal dimension of the feature, d_V is the feature dimension per patch, and H and W are the numbers of patches p along the height and width of each frame, respectively. Then, each $i \in t \times d_V$ of F_V is directly converted to an image token T_V^i as input.

Beyond Semantics. Most open-sourced MLLM tokenizers including Qwen-2.5-VL visual encoder are pretrained on semantically rich text–image pairs via contrastive training, and thus often lack a well-formed understanding of information beyond semantics. Meanwhile, El Banani et al. (2024); Tong et al. (2024b;a) draw a valuable conclusion that self-supervised vision models, such as DINO series, learn rich depth, normal, and pattern representations. Therefore, leveraging popular DINOv3 (Siméoni et al., 2025)'s strong dense features seems to be a natural approach beyond simple semantics. The last layer of DINOv3 produces patch-level dense features $F_D \in \mathbb{R}^{T \times d_D \times H_D \times W_D}$. We pad and regularize the original image to align with the patch size p, enforcing $H_D = H$ and $W_D = W$. We then apply a simple MLP, $\mathbb{R}^{d_D} \to \mathbb{R}^{d_V}$, to map channel dimensions. For the temporal dimension, we use the same temporal pooling with the previously mentioned temporal patch size τ to aggregate across T, yielding features $F'_D \in \mathbb{R}^{t \times d_V \times H \times W}$. The fusion of the video feature F_V and dense feature F'_D is shown as:

$$F'_V = \mathrm{CA}(F_V, F'_D, F'_D) + F_V, \tag{1}$$

where $\mathrm{CA}(q, k, v)$ denotes multi-layer cross-attention over the query, key, and value inputs. Then, we convert F'_V into a fused image token T_V^i, and the remaining calculations proceed as before.

Scale Experts Design. During all-scale mixed training in Fig.4(a-d), potential cross-scale knowledge conflicts lead to suboptimal results. This underscores the importance of preserving knowledge independence between scales during training. Inspired by Wu et al. (2024a); Buehler & Buehler (2024); Chen et al. (2024b), we further introduce a LoRA-like module that adds scale experts by fine-tuning only 0.5% of the overall parameters for each expert. The original LoRA is using $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times d}$ with the rank $r \ll \min(d, k)$ to approximate orginal weights W_0. To construct scale LoRA experts, We attach M scale experts $\{(A_i, B_i)\}_{i=1}^{M}$ to mitigate potential scale-level knowledge interference. Each expert i has a base weight α_i and is dynamically scaled by a learned factor λ_i:

$$h = W_0 x + \sum_{i=1}^{M} \alpha_i^* \, B_i A_i x, \text{where } \alpha_i^* = \alpha_i \cdot \lambda_i, \tag{2}$$

where x, h are the input and output of the projection layer, and α_i^* is the scaled factor. The learned factor λ_i is obtained through a scale router-primarily an MLP and a softmax. We apply M scale experts to each layer of the foundation LLM. Therefore, different layers, according to their respective conditions, obtain appropriate λ_i to allocate the experts within the layer. Given that scenarios of scales can overlap (for example, an indoor scene may include some tabletop context), in the ideal case, the routers can select the suitable experts at different layers.

Process Reward Design. After basic SFT training, RL is used to align the model with human perception. Inspired by how humans approach spatial observation tasks, we model the reasoning process explicitly. Humans typically proceed by: 1) identifying the task-specified semantics (if they help), 2) perceiving the global scale by inspecting surrounding objects (if it helps), and 3) inferring the answer from spatial relations. Following this paradigm, we construct 3 different anchors for RL that enforce the reasoning path to traverse the resulting anchor states. While certain reasoning anchors are not helpful to some tasks, we provide the minimal, sufficient ground-truth anchors for each question to guide the model in selecting the appropriate ones. We design the following three reward components based on these anchor formats: `<semantics>`, `<scale>`, and `<answer>`. Semantic reward R_{semantic} is used to identify the referenced objects; Scale reward R_{scale} is used to estimate the scale of the overall scene; Correctness reward R_{answer} is used to ensure the answer is well derived. The updated correctness reward $\bar{R}_{\text{answer}}$ can be formed into

$$\bar{R}_{\text{answer}} = \sum_{k=1}^{3} \prod_{n=1}^{k} R_{j_n}, \text{with } (j_1, j_2, j_3) = (\text{answer}, \text{scale}, \text{semantic}), \tag{3}$$

$$\text{where} \quad R_{\text{scale}} = \max(0, 1 - \frac{|\log C_{\text{ans}} - \log C_{\text{gt}}|}{2}), \quad R_{\text{semantic}} = \frac{S_{\text{ans}} S_{\text{gt}}}{\|S_{\text{ans}}\| \|S_{\text{gt}}\|}. \tag{4}$$

$C_{\text{ans}}, C_{\text{gt}}$ is the estimated scene scale in the same measurement; $S_{\text{ans}}, S_{\text{gt}}$ is the calculated semantic embedding. C_{gt} and S_{gt} can be easily obtained from Sec.3. It is crucial to note that the order of $(j_1, ..., j_n)$ matters; rewards at the beginning are stricter and more important. Also, because tasks differ, for example in the camera rotation task, R_{semantic} and R_{scale} are not needed. Thus, $\bar{R}_{\text{answer}}$ under such circumstances collapses to a standard R_{answer}. The calculation of format reward R_{format} and answer reward R_{answer} remains the same as common practice (Feng et al., 2025; Guo et al., 2025a) to encourage the generation of valid and executable answers. Therefore, our reward design forms the accurate reward signals to ensure all-scale spatial compliance and encourage human-like thinking. It is worth noting that the evaluation does not involve these anchors besides the actual answer.

RL Training Objective. For each question i, we define the reward R_i to include both the updated correctness reward $\bar{R}_{\text{answer}}$ and R_{format} following Guo et al. (2025a), and use this overall reward R_i to compute groupwise normalized advantages $A_i = \frac{R_i - \text{mean}(\{R_j\})}{\text{std}(\{R_j\})}$. $\{R_j\}$ is the response group related to R_i. The final policy π_θ is updated by maximizing

$$\mathbb{J}(\theta) = \mathbb{E}_{q,\{o_i\}} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\min\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} A_i, \, \text{clip}\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, \, 1-\epsilon, \, 1+\epsilon \right) A_i \right) \right) - \beta \, \mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \right], \tag{5}$$

where $\pi_{\theta_{\text{old}}}$ and π_θ are the old and new policy model respectively. $\mathbb{D}_{\text{KL}}$ represents KL divergence.

Training Strategy. We start with a cold-start phase on SpaceVista-1M, optimizing the input projection, feature-fusion modules, and scale experts. Next, we introduce the scale router to further train each scale-specific expert on the appropriate inputs, encouraging specialization. Finally, building on the SFT model, we apply RL training to obtain the final SpaceVista-7B reasoning model.

5 EXPERIMENT

Datasets. We use SpaceVista-1M in Sec. 3 for SFT and RL; its sources are detailed in Appendix B.2.

Model Configurations. Our model is built on Qwen2.5-VL-7B for main experiments and Qwen2.5-VL-3B for ablation. Our model is trained on up to 16 NVIDIA A800 (80GB) GPUs. We process a maximum of 32 frames during training, each with a resolution of $128 \times 28 \times 28$ pixels. During inference, we increase the resolution ($256 \times 28 \times 28$ pixels) to enhance performance. During the expert training phase, we employ 4 experts, each tailored to a distinct scenario. We set the group size of GRPO to 8. We first perform SFT on CoT data of SpaceVista-1M for two epochs to obtain the SFT model. This is followed by RL training for 2.5k steps on multi-choice and regression data to produce the final SpaceVista-7B. Additional details are provided in Appendix C.1.

Benchmarks. We evaluate our model on 5 benchmarks, VSI-Bench (Yang et al., 2025a), STI-Bench (Li et al., 2025e), SpaceVista-Bench (Ours), MMSI-Bench (Yang et al., 2025b) and SPAR-Bench (Zhang et al., 2025d). Among the benchmarks, the former three are video-based, while the

Table 2: Performance comparison across five spatial reasoning benchmarks. Among them, SpaceVista-Bench is our proposed all-scale benchmark. Open-sourced general models are evaluated with a comparable size. The highest performance of the open-sourced model is marked **bold**.

	Multi-Image		Video		
Model	MMSI-Bench	SPAR-Bench	VSI-Bench	STI-Bench	SpaceVista-Bench
Human	97.2	67.3	79.2	-	81.3
Closed-sourced Commercial Model & 70B-class model					
GPT-5(OpenAI, 2025)	40.7	37.4	44.2	39.3	33.7
Gemini-2.5-pro(DeepMind, 2025)	36.9	36.3	45.0	41.4	33.8
InternVL3.5-38B (Wang et al., 2025c)	36.9	31.0	66.3	39.2	30.7
Qwen2.5-VL-72B (Bai et al., 2025)	30.7	32.4	30.7	40.7	31.1
Open-sourced General Model					
LLAVA-Onevision-7B (Li et al., 2024a)	24.5	30.6	32.4	29.0	13.6
LLaVA-NeXT-Video-7B (Liu et al., 2024a)	26.8	31.3	35.6	29.9	23.7
InternVL3.5-8B (Wang et al., 2025c)	30.9	36.0	38.2	33.2	24.5
Qwen2.5-VL-7B (Bai et al., 2025)	31.7	33.1	32.7	32.1	28.9
Open-sourced Specialized Model					
SpaceR-7B (Ouyang et al., 2025)	26.1	37.6	46.9	37.0	21.2
SpatialMLLM-4B (Wu et al., 2025a)	27.0	31.5	48.4	30.5	24.2
VILASR-7B (Wu et al., 2025b)	30.2	37.6	45.4	31.5	23.6
VG LLM-4B (Zheng et al., 2025)	-	-	46.1	29.3	28.8
Qwen2.5-VL-7B $w/.$ SpaceVista-1M	27.3	36.9	42.0	35.0	29.5
SpaceVista-7B (Ours)	29.1	38.1	46.3	35.9	34.5
SpaceVista-7B (Ours) $w/.$ RL	**32.3**	**41.6**	**48.6**	**38.2**	**36.7**

Table 3: Module ablation study using Qwen-2.5-VL-3B on SpaceVista after RL.

Module	VSI-Bench	SpaceVista-Bench
Vanilla	44.4	31.0
$w/.$ Scale	46.3 (+1.9)	34.8 (+3.8)
$w/.$ Scale &Semantic	46.8 (+2.4)	35.4 (+4.4)
$w/.$ Expert Finetuning	45.8 (+1.4)	34.8 (+3.8)

Table 4: Modality ablation study of the extra input types beyond semantic information.

Input	VSI-Bench	SpaceVista-Bench
Vanilla	44.4	31.0
$w/.$ VGGT	44.3 (-0.1)	31.4 (+0.4)
$w/.$ DINOv3	46.4 (+2.0)	32.1 (+1.1)
$w/.$ VGGT + DINOv3	45.3 (+0.9)	31.7 (+0.7)

latter two are multi-image benchmarks. We argue that video and multi-image tasks share rather strong similarities and collectively serve as important benchmarks for cross-frame spatial understanding. For all evaluations, we follow the configuration used in the official Qwen2.5-VL demo, with top_p = 0.001 and temperature = 0.01.

Comparison on Spatial Reasoning Datasets. Our method attains competitive performance across all spatial reasoning benchmarks in Tab. 2. On VSI-Bench, we achieve comparable results approaching the state of the art. More importantly, our approach delivers substantially superior performance in our all-scale benchmark SpaceVista-Bench, markedly exceeding 3% compared with proprietary and open-source models. Thus, SpaceVista-7B represents a robust baseline for both indoor and all-scale scenes, where the full comparison table of each benchmark is shown in Appendix. D.4 for reference.

Comparison on Subsets of SpaceVista-Bench. In Tab.5, we analyze the performance of popular models on each subset of our SpaceVista bench. In general, the small-scale subsets challenge both commercial and general models, likely due to biases in the pre-training corpus. Limited by device constraints, close-range shots constitute only a small fraction of the data, while the abundance of indoor and outdoor scenes yields relatively higher performance. We also observe that most models perform at a relatively low level on SpaceVista-Bench, indicating that it has the expected discriminative power for all-scale reasoning and can serve as a foundational benchmark to help the community enrich the overall evaluation ecosystem. Our SpaceVista-7B, although exhibiting minor improvements on indoor scenes, attains comparatively high comprehensive scores across other scenarios and in overall evaluations. The results indicate a clear boost

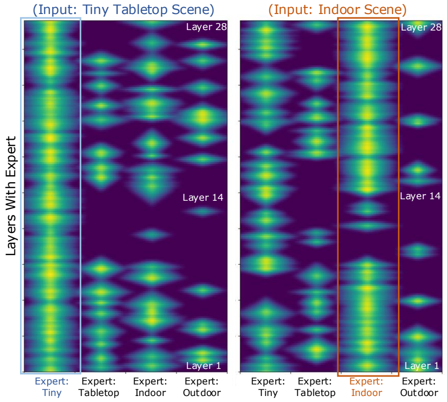


Figure 5: Visualization of scale-expert activations on salient tokens with an appropriate threshold. This shows the router selects experts based on the input during inference.

of around 6% compared with any size of the open-source models in comprehensive all-scale spatial reasoning.

Table 5: The SpaceVista-Bench leaderborad. We utilize green (1st), blue (2nd), and yellow (3rd) backgrounds to distinguish the top three results within each scene. We employ **bold** and underlined text to denote the bests and second-best results across all open-source models. All the baselines are instruction-tuned and are evaluated on the same resolution and fps.

Models	**SpaceVista-Bench** **Tiny Tabletop**	 **Tabletop**	 **Indoor**	 **Outdoor**	 **Overall**
Closed-sourced Commercial Model					
🥉 GPT-5(OpenAI, 2025)	32.3	20.3	39.0	43.0	33.7
GPT-4o(Hurst et al., 2024)	21.7	13.3	34.3	38.3	26.9
🥈 Gemini-2.5-pro(DeepMind, 2025)	33.0	38.7	34.5	29.0	33.8
Gemini-2.5-flash(DeepMind, 2025)	20.7	30.0	19.9	26.9	24.4
Claude-Sonnet-4(Anthropic, 2025b)	27.3	19.3	38.1	34.1	29.7
Claude-Opus-4.1(Anthropic, 2025c)	21.7	29.5	24.3	30.0	26.4
Open-Source General Model					
Internvl3.5-38B (Wang et al., 2025c)	29.3	25.2	41.2	27.0	30.7
Internvl3.5-14B (Wang et al., 2025c)	27.7	22.3	31.3	24.3	26.4
Internvl3-78B (Zhu et al., 2025)	**38.3**	23.3	42.2	30.3	33.5
Internvl3-38B (Zhu et al., 2025)	18.7	14.3	34.8	**38.0**	26.5
GLM-4.5V (Team et al., 2025)	23.0	17.8	27.3	25.2	23.3
GLM-4.1V-Thinking (GLM et al., 2024)	30.7	19.3	29.0	13.3	23.1
Qwen2.5VL-72B (Bai et al., 2025)	27.7	20.3	29.6	28.0	26.4
Qwen2.5VL-32B (Bai et al., 2025)	25.3	19.3	38.1	30.7	28.4
LLAVA-Onevision-72B (Li et al., 2024a)	25.0	12.0	15.3	11.7	16.0
LLAVA-Onevision-7B (Li et al., 2024a)	17.5	8.0	13.3	11.6	12.6
Open-Source Specialized Model					
SpaceR (Ouyang et al., 2025)	12.9	17.3	34.9	19.8	21.2
Spatial-MLLM (Wu et al., 2025a)	17.3	20.3	36.1	23.1	24.2
VLM-3R (Wu et al., 2025a)	15.1	24.6	**45.1**	26.9	27.9
🥇 SpaceVista-7B (Ours)	33.4	**37.1**	42.2	34.1	**36.7**

Ablation on Each Component. 1) Scale Expert: We examine how potential information conflicts during cross-scale training are mitigated. As shown in Tab.3, the experts yield substantial gains. As the number of experts increases, the performance also improves accordingly in Tab. 6. Furthermore, visualizing the activation distributions of different LoRA experts across scenes (Fig.5) indicates that scale-specific knowledge is somehow disentangled. 2) Reward: In Tab. 3, the progressive reward achieves higher performance than the unconstrained reasoning path. These optional anchors indeed serve as a valuable halfway point in the all-scale reasoning process. This highlights the importance of specifying thinking anchors when designing all-scale reasoning.

Ablation on Each Modality. As shown in Tab. 4, incorporating DINO v3 yields greater gains than VGGT with its obvious advantage of self-supervised dense cues. In contrast, VGGT's raw geometry features are harder for a simple fusion model to use without the strong decoder. Also, VGGT can be easily influenced by the blur or occlusion in the video. We further provide performance of the rendered 2.5D in Appendix. D.6 as interesting explorations.

Table 6: Ablation of the number of experts based on the same training settings.

Num of Expert(s) (M)	Training Data (Each Expert)	VSI-Bench	SpaceVista -Bench (Ours)
None	All	44.4	31.0
1	All	44.2 (-0.2)	31.0 (0)
2	1/2	45.6 (+1.2)	32.7 (+1.7)
4	1/4	45.7 (+1.3)	32.9 (+1.9)
6	1/6	43.1 (-1.1)	26.7 (-4.3)

More Experiments. To facilitate a deeper understanding, we provide more previews, statistics, experiments, user studies, and discussion in the appendix, especially Appendix D,E for more insights.

6 DISCUSSION AND CONCLUSION

Discussion. It is believed that SpaceVista can facilitate widespread application in various areas on all scales, such as 1) spatial captioning, 2) spatial guided visual generation, 3) interactive world models. Although our all-scale model shows strong performance in various spatial reasoning tasks, there is still potential for improvement, for example, μm level for precision manufacturing, mm-level for medical surgery, km-level coverage for remote sensing, and $10km$-scale for cartography.

Conclusion. In this work, we introduce a novel task for all-scale reasoning from visual spatial context, which requires the machine to understand multimodal information and respond with the correct answer and rationale. To advance this field, we develop the first open-source, all-scale, spatial reasoning dataset, SpaceVista-1M, for cold start and reinforcement learning. Additionally, we handcraft SpaceVista-Bench, an accurate, multi-scale, video-based benchmark that strictly adheres to physical world measurements and perceptions. Our proposed SpaceVista-7B model further establishes a robust baseline with enhanced cross-scale perception. During experiments, we compare our SpaceVista-7B model with several existing models and demonstrate our proposed model's promising performance in all-scale reasoning. Additionally, our task and dataset have great potential in applications such as industrial manufacturing, embedded systems, and autonomous driving to understand complicated spatial environments in the wild.

7 NECESSARY STATEMENT

7.1 REPRODUCIBILITY STATEMENT

We will open-source the dataset, code, and models on our demo page. Appendix F presents comprehensive visual previews and documentation of the dataset, and the release will follow the Creative Commons Attribution (CC BY) license and Apache License 2.0 specified in Appendix B.4.8. Appendix C details hyperparameter settings, training and evaluation protocols, and extended analyses. To facilitate reproducibility, we will provide configuration files and scripts aligned with the main results. Please refer to the mm2km website for the most recent releases and updates.

7.2 ETHICS STATEMENT

This work focuses on technical advances in multi-scale spatial reasoning and dataset construction, without conducting interventional studies on human subjects or collecting sensitive personal data. Data are curated via expert-driven automated annotation with a small, carefully manual benchmark, following de-identification and compliant release practices. The model and datasets aim to improve machine understanding and generalization across scales and scenarios and do not provide actionable guidance for misuse. No undisclosed conflicts of interest or improper military usage are involved. Potential bias and fairness risks are acknowledged and mitigated through diverse, multi-scale evaluations. Privacy, copyright, legal compliance, and research ethics (including appropriate documentation and review) are carefully observed. Accordingly, the topic presents no ethical conflicts.

7.3 THE USE OF LARGE LANGUAGE MODELS

The LLM was indeed used only for language polishing (e.g., grammar, spelling, clarity, tone) on text whose content and structure were created by the authors. No substantive changes to claims, data interpretation, or conclusions were introduced by the LLM.

REFERENCES

Anthropic. Claude 3.7 sonnet and claude code. `https://www.anthropic.com/news/claude-3-7-sonnet`, 2025a.

Anthropic. Introducing claude 4. https://www.anthropic.com/news/claude-4, 2025b. Accessed: 2025-07-24.

Anthropic. Claude opus 4.1. https://www.anthropic.com/news/claude-opus-4-1, 2025c. Version VIII, released in 2025.

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

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Eric L Buehler and Markus J Buehler. X-lora: Mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and molecular design. *APL Machine Learning*, 2(2), 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*, 2024b.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22831–22840, 2025.

Yihang Chen, Qianyi Wu, Mengyao Li, Weiyao Lin, Mehrtash Harandi, and Jianfei Cai. Fast feedforward 3d gaussian splatting compression. *arXiv preprint arXiv:2410.08017*, 2024c.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024d.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024f.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Google Deepmind. Gemini 2.0: our new ai model for the agentic era. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message`, 2024.

Google DeepMind. Gemini 2.5: Our most intelligent ai model. `https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/`, 2025.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. *arXiv preprint arXiv:2507.16443*, 2025a.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Xingguang Wei, Tianshuo Yang, Min Dou, et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *arXiv preprint arXiv:2506.18385*, 2025b.

Zihao Dongfang, Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Danda Pani Paudel, Luc Van Gool, Kailun Yang, and Xuming Hu. Are multimodal large language models ready for omnidirectional spatial reasoning? *arXiv preprint arXiv:2505.11907*, 2025.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

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. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning. *arXiv preprint arXiv:2506.00318*, 2025.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2485–2494, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025b.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. Lvis: A dataset for large vocabulary instance segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5351–5359, 2019.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26181–26191, 2025.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Steven LaValle. Rapidly-exploring random trees: A new tool for path planning. *Research Report 9811*, 1998.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025a.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025b.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*, 2025c.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024c.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025d.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*, 2025e.

Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. Improved visual-spatial reasoning via r1-zero-like training. *arXiv preprint arXiv:2504.00883*, 2025.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Fei Liu, Zihao Lu, and Xianke Lin. Vision-based environmental perception for autonomous driving. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 239:39 – 69, 2022.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023b.

Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y. Zhang, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotny. Uncommon objects in 3d. In *arXiv*, 2025a.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024b.

Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025b.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2024c. URL `https://arxiv.org/abs/2412.04468`.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025c.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *arXiv preprint arXiv:2504.20024*, 2025.

Mahya Nikouei, Bita Baroutian, Shahabedin Nabavi, Fateme Taraghi, Atefe Aghaei, Ayoob Sajedi, and Mohsen Ebrahimi Moghaddam. Small object detection: A comprehensive survey on challenges, techniques and real-world applications. *ArXiv*, abs/2503.20516, 2025.

OpenAI. Introducing gpt-4.5. `https://openai.com/index/introducing-gpt-4-5/`, 2025a.

OpenAI. Introducing o3 and o4 mini. `https://openai.com/index/introducing-o3-and-o4-mini/`, 2025b.

OpenAI. GPT-5 System Card. Technical report, OpenAI, August 2025. Accessed: 2025-08-10.

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

Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17359–17369, 2025.

Kiru Park, Timothy Patten, and Markus Vincze. Neural object learning for 6d pose estimation using a few cluttered images. In *The European Conference on Computer Vision (ECCV)*, 2020.

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

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025.

Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, et al. We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. *arXiv preprint arXiv:2508.10433*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

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

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stanley T. Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. *ArXiv*, abs/2411.16537, 2024.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21686–21697, 2024a.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models, 2025d. URL `https://arxiv.org/abs/2502.08636`.

Tao Wen, Jiepeng Wang, Yabo Chen, Shugong Xu, Chi Zhang, and Xuelong Li. Metric-solver: Sliding anchored metric depth estimation from a single image. *arXiv preprint arXiv:2504.12103*, 2025.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025a.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing, 2025b. URL `https://arxiv.org/abs/2506.09965`.

Peiran Wu, Yunze Liu, Miao Liu, and Junxiao Shen. St-think: How multimodal large language models reason about 4d worlds from ego-centric videos. *arXiv preprint arXiv:2503.12542*, 2025c.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024a.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024b. URL `https://arxiv.org/abs/2412.10302`.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jiaping Xiao, Rangya Zhang, Yuhang Zhang, and Mir Feroskhan. Vision-based learning for drones: A survey. *IEEE transactions on neural networks and learning systems*, PP, 2023.

Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025.

Tong Xu. Recent advances in rapidly-exploring random tree: A review. *Heliyon*, 10(11):e32451, 2024. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2024.e32451. URL `https://www.sciencedirect.com/science/article/pii/S2405844024084822`.

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 Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025b.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025c.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a. URL `https://arxiv.org/abs/2501.13106`.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025b.

Haoyu Zhang, Meng Liu, Zaijing Li, Haokun Wen, Weili Guan, Yaowei Wang, and Liqiang Nie. Spatial understanding from videos: Structured prompts meet simulation data. *arXiv preprint arXiv:2506.03642*, 2025c.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025d.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yujie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv preprint arXiv:2503.22976*, 2025e.

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

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture. *arXiv preprint arXiv:2509.02359*, 2025f.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024b. URL `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024c. URL `https://arxiv.org/abs/2410.02713`.

Zixuan Zhao. Advances and challenges in small object detection: A comparative analysis of state-of-the-art models and future directions. *Theoretical and Natural Science*, 79:145–153, 01 2025.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *arXiv preprint arXiv:2505.24625*, 2025.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. *arXiv preprint arXiv:2508.02095*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

APPENDICES CONTENTS

A Important Information **21**

A.1 Task Distribution 21

A.2 Performance Radar 21

B Data Construction **21**

B.1 Data Comparison 22

B.2 Data Source 22

B.2.1 Tiny Tabletop Scene 23

B.2.2 Tabletop Scene 23

B.2.3 Indoor Scene 23

B.2.4 Wild Indoor Scene 23

B.2.5 Outdoor Scene 24

B.2.6 Drone Scene 24

B.2.7 Our Own Collected Data 24

B.3 Task Construction 25

B.3.1 Data Preparation 25

B.3.2 Type: Distance 27

B.3.3 Type: Counting 27

B.3.4 Type: Planning 28

B.3.5 Type: Relation 28

B.3.6 Data Post-Processing 29

B.3.7 Benchmark Construction 29

B.4 Data Statistics 29

B.4.1 Target Category Distribution 29

B.4.2 Scale Distribution 31

B.4.3 Subscene Type Distribution 31

B.4.4 Object Size Distribution 31

B.4.5 Camera To Object Distribution 31

B.4.6 QA Statistics Across Scenes 32

B.4.7 Data Quality Control 32

B.4.8 License 33

B.5 Supplementary citation 33

C Model Detail **34**

C.1 Parameter Setting 34

C.2 Patch Level Encoder Ablation 34

C.3 LoRA Like Expert Ablation 34

D Observation Results **35**
D.1 GRPO Reward Observation 35
D.2 Expert Observation 35
D.3 Reasoning Vs Memorizing (Out-of-Distribution Problem) 35
D.4 Detailed Analysis on Each Benchmark 37
D.5 The Hardest Scene 40
D.6 Why 2.5D>3D 41
D.7 Scaling-Up Analysis 41
D.8 Leaderboard Detail 42

E FAQ **42**
E.1 Error Accumulation 42
E.2 All Scale Possibilities 43
E.3 Discussion of Dataset 43

F Preview **44**
F.1 Scene Preview 44
F.2 Template Preview 45
F.3 QA Preview 45

A IMPORTANT INFORMATION

A.1 TASK DISTRIBUTION

Our SpaceVista-1M consists of a wide range of tasks, including both general tasks and scale-specific tasks. Fig. A6 illustrates the data composition for each scene task, where bubble sizes indicate the relative data volume.

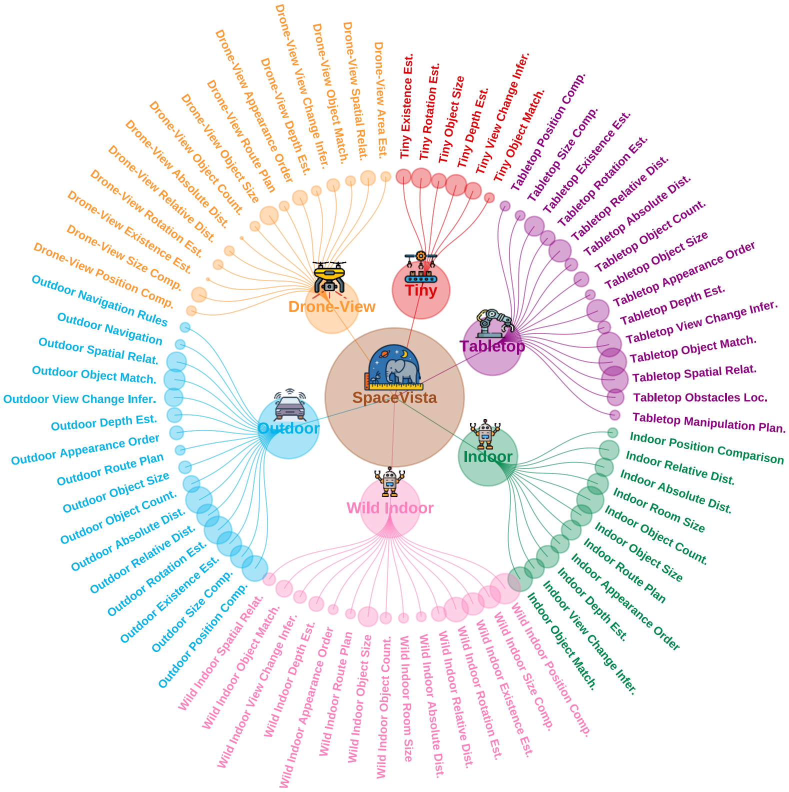


Figure A6: Statistical chart of QA types. The spatial reasoning tasks for various scenes include abbreviations, for example, "Est." for Estimation, "Dist." for Distance, "Loc." for Location, and "Com." for Comparison.

A.2 PERFORMANCE RADAR

The comparison across models is carried out on multiple spatial reasoning benchmarks. We evaluate eight multimodal large models on five distinct benchmarks, with the results visualized in the radar chart in Fig. A7.

SpaceVista-7B achieves significant improvement across the benchmarks, highlighting its superiority in spatial reasoning tasks. While models, including LLAVA-Onevision-7B (Li et al., 2024a), demonstrate competitive performance, SpaceVista-7B consistently exhibits superior robustness and adaptability across a range of tasks, thereby solidifying its position as a robust model in spatial reasoning.

B DATA CONSTRUCTION

Our SpaceVista-1M dataset spans 19 spatial reasoning task types, including scale-specific tasks, comprising 1 million QA pairs and 38 thousand videos collected across diverse scenes. This scale and variety enable large-scale training of perceptual understanding and spatial reasoning, and support comparative analysis across tasks and environments.

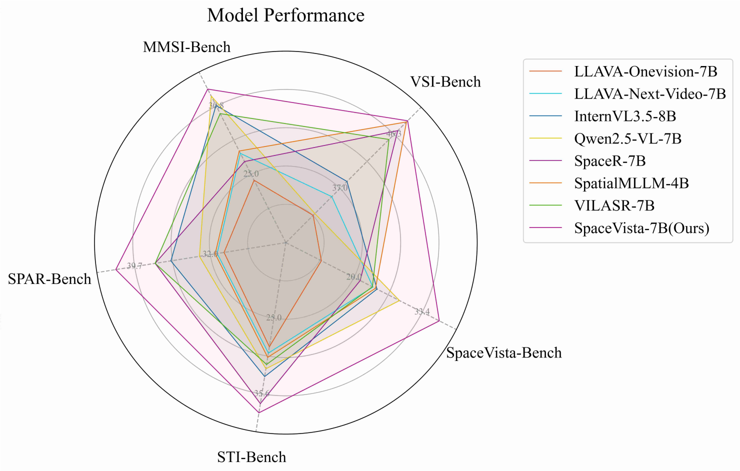


Figure A7: Comparison across popular spatial reasoning benchmarks. Our SpaceVista-7B model achieves certain performance boosts across all benchmarks.

This chapter details the data sources for each scene category (Sec. B.2), the end-to-end task construction pipeline (Sec. B.3.1), and key dataset statistics (Sec. B.4).

B.1 DATA COMPARISON

Table B7: The datasets we used to build SpaceVista-1M and SpaceVista-Bench. "†" means the datasets are only used for evaluation in SpaceVista-Bench. "‡" means data collected by us and used for accurate evaluation. The definition of scenes is the number of unique spaces, and one scene can be transformed into multiple questions.

Dataset	**Type**	**Scenes**
uCO3D(Liu et al., 2025a)	Tiny, Tabletop	10,000
WildRGB-D(Xia et al., 2024)	Tabletop	11,300
SMOT(Park et al., 2020)	Tabletop	13
SpaceR(Ouyang et al., 2025)	Indoor	1,500
Spar-Bench(Zhang et al., 2025e)	Indoor	4,500
Scannet Series(Dai et al., 2017; Yeshwanth et al., 2023)	Indoor	460
VSI-Bench†(Yang et al., 2025a)	Indoor	288
MMSI-Bench†(Yang et al., 2025b)	Indoor	231
DL3DV(Ling et al., 2024)	Drone, Indoor, Outdoor	10,510
STI-bench†(Li et al., 2025e)	Indoor, Outdoor, Tabletop	372
Our own collected data ‡	Tiny, Tabletop, Outdoor	500

Our current dataset encompasses a broad diversity of scene categories, as summarized in Tab. B7. The data sources span a wide range of scenarios, including tiny, tabletop, indoor, outdoor, and drone-view.

To ensure evaluation quality and robustness, we apply multiple rounds of processing and rigorous filtering to all collected data. We remove redundant or inconsistent samples across datasets. Because scenes may overlap across sources, which can compromise the independence of the training and test splits, we removed from the training set any scene that appears in all the benchmarks. This strict separation prevents leakage and enables a fair assessment of generalization. Consequently, the SpaceVista-1M provides broad scene diversity, with a clean, reliable benchmark SpaceVista-Bench.

B.2 DATA SOURCE

Sec. B.2 presents data sources that form our dataset, and systematically describes the provenance and acquisition of seven scene sources. These sources combine multiple public datasets and our own collected data, as detailed in Sec. B.2.1- B.2.7. These scenes span object-centric through scene-level contexts and exhibit substantial variation in scale, shape, pattern, and illumination.

When building the dataset, our foundational data construction process must adhere to the following key criteria:

- **Video Data with 3D Modeling**: The data must consist of video sequences accompanied by either official or third-party 3D modeling. This enables effective use of camera parameters for robust data processing.
- **Multi-Frame & Multi-Scale**: The dataset should support meaningful spatial reasoning across multiple frames and scales. Its complexity must be sufficient to prevent trivial single-frame assessments from representing the full sequence.
- **Comprehensive Annotations & Metadata**: Each sample must include the following: (a) camera intrinsics and extrinsics, (b) detection and segmentation labels, and (c) dense depth maps. These elements support a broad range of downstream tasks.

B.2.1 TINY TABLETOP SCENE

We curate small-scale, small-object videos from uCO3D (Liu et al., 2025a), selecting sequences where the object size falls below a predefined threshold to instantiate the tiny tabletop scenario. uCO3D comprises approximately 170,000 high-resolution, object-centric 360-degree videos captured via crowdsourcing, covering more than 1,000 LVIS (Gupta et al., 2019) categories grouped into 50 categories. For each video, uCO3D applies VGGSfM (Wang et al., 2024a) for motion analysis and 3D Gaussian Splatting to generate accurate camera poses, depth maps, sparse and dense point clouds, and semantic captions. The resulting subset contains everyday small objects, such as stationery, food, and decorative items, placed on flat surfaces such as tables, counters, and shelves. These scenes provide complete viewpoint coverage, precise geometry, and rich semantic labels, which make them well-suited for fine-grained 3D object modeling and spatial video reasoning. Here, we only select a small part of uCO3D for around 10,000 videos for tiny objects after filtering.

B.2.2 TABLETOP SCENE

For tabletop scene modeling, we select two datasets: WildRGB-D (Xia et al., 2024) and SMOT (Park et al., 2020). WildRGB-D consists of approximately 8,500 objects across 46 categories, recorded in around 20,000 RGB-D videos, with iPhones rotating 360 degrees around objects to replicate real-world interactions. It includes single-object, multi-object, and hand-occlusion videos, all automatically annotated via SLAM-generated camera poses and reconstructed point clouds, making it suitable for spatial reasoning tasks. To select samples for spatial reasoning, we specifically choose around 10,000 videos with multiple objects in a scene. SMOT (Park et al., 2020) is a challenging small dataset collected by a mobile robot, comprising 13 video sequences.

The tabletop, commonly referred to as the “table” scene, encompasses not only the planar surface of a table but also extends to various other surfaces, including sand, beds, wardrobes, floors, and similar environments. In combination, these datasets offer richly varied planar scenes, providing a robust foundation for challenging spatial video reasoning benchmarks.

B.2.3 INDOOR SCENE

Indoor scenes are among the earliest domains studied in spatial video reasoning. Key datasets, including ScanNet (Dai et al., 2017) and ScanNet++ (Yeshwanth et al., 2023), collect RGB-D scans using handheld cameras, yielding aligned RGB images, depth maps, and 3D reconstructions. ScanNet contains more than 1,500 scenes and 2.5 million frames spanning common indoor spaces, such as offices and bedrooms, with annotations for over twenty object categories. ScanNet++ extends this setting with higher geometric fidelity and more complex layouts. The combination of focused object classes, structured environments, and rich annotations makes these datasets central benchmarks for spatial reasoning.

B.2.4 WILD INDOOR SCENE

Beyond scan-based indoor modeling, DL3DV (Ling et al., 2024) adopts a video-based pipeline that replaces active scanning with video capture and camera parameter estimation. Building on this framework, and further compressed using 3D Gaussian Splatting (Chen et al., 2024c), DL3DV enables high-precision 3D reconstruction of wild indoor scenes. The dataset covers a broad range of object categories, including challenging reflective and transparent instances. Compared with conventional scan-based datasets, these scenes exhibit greater geometric and appearance variability, providing a more realistic and demanding benchmark for spatial video reasoning.

B.2.5 OUTDOOR SCENE

In addition to tabletop and indoor scene modeling, DL3DV (Ling et al., 2024) collects extensive in-the-wild outdoor videos encompassing landmarks, street corners, private courtyards, and urban parks. Camera parameters are calibrated using COLMAP (Schönberger et al., 2016; Schönberger & Frahm, 2016). The DL3DV-10K dataset includes 10,510 videos in 4K resolution, totaling about 51.2 million frames, covering 65 types of locations. Each video is annotated for whether it is indoors or outdoors as well as for levels of reflection, transparency, and lighting conditions. Compared to conventional scan-based indoor datasets, these outdoor scenes exhibit richer geometric complexity, greater diversity of materials, and wider environmental variation, offering more challenging benchmarks for spatial video reasoning.

B.2.6 DRONE SCENE

DL3DV (Ling et al., 2024) extends outdoor scene modeling by incorporating drone-captured videos that provide aerial perspectives to complement ground level views. Videos are recorded using unmanned aerial vehicles (UAVs), and camera parameters are calibrated through COLMAP (Schönberger et al., 2016; Schönberger & Frahm, 2016), following the same reconstruction pipeline applied to handheld footage. The DL3DV Drone subset consists of more than 100 videos covering a variety of scenes, including open plazas, tree-lined pathways, rooftop platforms, and landmark facades. DL3DV enhances spatial video reasoning by introducing unique geometric structures and varied viewpoints.

Although the data scale is not as large as tabletop or indoor, the drone-view scenes establish a more rigorous benchmark for aerial mapping and spatial video reasoning by expanding scene diversity and viewpoint range.

B.2.7 OUR OWN COLLECTED DATA

The data collection methods described above rely on advanced specialized models and fully automated pipelines. While we incorporate limited manual filtering, whether the resulting data can be used as an accurate evaluation of real-world perception is still a question. This limitation motivates our collection of higher-fidelity data to better align with physical world perception.

Our dataset consists of two types: 1) measured, recorded, and manually annotated data, and 2) existing video data enhanced by retrieving and verifying publicly available information. The former is suitable for tiny objects, tabletop objects, whereas the latter is designed for indoor and outdoor scenarios.

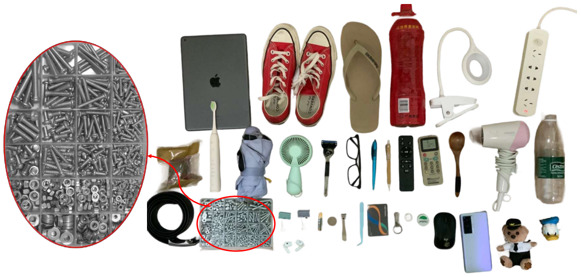


Figure B8: Our self-collected data features various categories of objects, with tabletops and tiny tabletops ranging from 0.4m to 3mm, even including transparent and reflective objects.

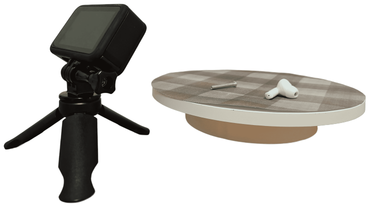


Figure B9: A photo of the real scene for the collection of tiny tabletop.

Data from self-recording and measurement. Precise spatial annotations (e.g., location and dimensions) are scarce in existing datasets such as uCO3D and WildRGB-D. To address this, we captured length and positional data for nearly 50 object categories across diverse scenarios. Using GoPro 11, iPhone 15, and Vivo X70, we systematically varied object arrangements, distances, lighting conditions, and backgrounds into over 200 videos and 1,000 QA pairs. As illustrated in Fig. B8 and Fig. B9, they show the objects used for self-collected data and a real scene of tiny tabletop data collection. Although we collected the raw high-resolution videos up to 2.7K/60fps, it is still necessary to resize and resample it for better comparison. The resulting measurements are consolidated into a unified perceptual space that closely approximates physical world geometry.

Data Retrieved from authoritative sources. Adopting a similar rationale, it is apparent that spatial information derived solely from wild videos lacks the precision required for robust evaluation. Consequently, alternative methodologies must be explored. To address this, we propose a systematic approach that first identifies landmark objects within existing datasets and then manually retrieves images of these objects from authoritative sources, such as Wikipedia[1], architectural drawings, and official design documents, to obtain accurate spatial information, as shown in Fig. B10. This method ensures that the evaluation data is not only more precise but also more consistent with human perceptual judgments and preferences.

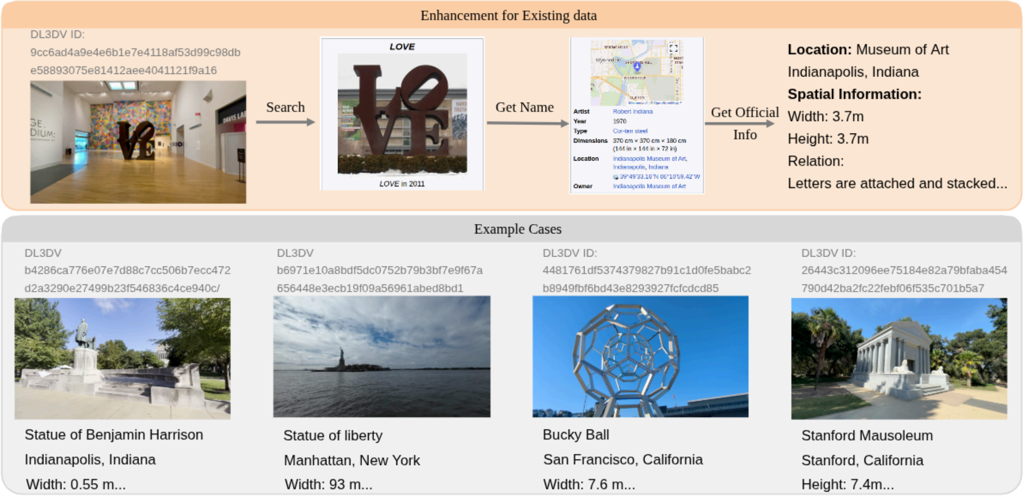

Enhancement for Existing data
DL3DV ID:
9cc6ad4a9e4e6b1e7e4118af53d99c98db
e58893075e81412aee4041121f9a16
Search
LOVE
LOVE in 2011
Get Name
Artist Robert Indiana
Year 1970
Type Cor-ten steel
Dimensions 370 cm × 370 cm × 180 cm (144 in × 144 in × 72 in)
Location Indianapolis Museum of Art, Indianapolis, Indiana
39°49′33.16″N 86°10′59.42″W
Owner Indianapolis Museum of Art
Get Official Info
Location: Museum of Art
Indianapolis, Indiana
Spatial Information:
Width: 3.7m
Height: 3.7m
Relation:
Letters are attached and stacked...
Example Cases
DL3DV
b4286ca776e07e7d88c7cc506b7ecc472
d2a3290e27499b23f546836c4ce940c/
Statue of Benjamin Harrison
Indianapolis, Indiana
Width: 0.55 m...
DL3DV
b6971e10a8bdf5dc0752b79b3bf7e9f67a
656448e3ecb19f09a56961abed8bd1
Statue of liberty
Manhattan, New York
Width: 93 m...
DL3DV ID:
4481761df5374379827b91c1d0fe5babc2
b8949fbf6bd43e8293927fcfcdcd85
Bucky Ball
San Francisco, California
Width: 7.6 m...
DL3DV ID:
26443c312096ee75184e82a79bfaba454
790d42ba2fc22febf06f535c701b5a7
Stanford Mausoleum
Stanford, California
Height: 7.4m...

Figure B10: Examples of identifying outdoor landmark objects from existing datasets and retrieving their scale-related ground truth data.

B.3 TASK CONSTRUCTION

Upon acquiring the appropriate dataset, we initially perform necessary data preparation and processing in Sec. B.3.1. Subsequently, we carefully design workflow for each task (Sec. B.3.3-B.3.5), and we present detailed task explanations in Tab. B8. The final output consists of high-quality QA pairs, facilitating the cold-start and reinforcement learning processes of MLLMs.

B.3.1 DATA PREPARATION

Previous popular approaches, such as InternSpatial (Deng et al., 2025b), required estimating camera intrinsic and extrinsic parameters, which introduced cumulative errors that propagated through subsequent tasks. However, since we exclusively utilize datasets with known camera parameters (as detailed in Sec B.2), our framework operates under conditions close to ground truth.

We first employ Metric3Dv2 (Hu et al., 2024) and UniDepthV2 (Piccinelli et al., 2025) to obtain accurate metric depth maps and normal maps. The metric depth maps provide precise distance measurements between the camera and scene objects, while the normal maps facilitate robust plane estimation. There are two challenges during construction. **1) Video consistency**: According to observation, the metric depth model may not have that level of consistency across frames. So, we use Video-Depth-Anything (Chen et al., 2025) to ensure consistency by minimizing the energy function,

$$D^* = \underset{D}{\operatorname{argmin}} \left\{ \|D - M\|_F^2 + \lambda \left\| \nabla_t(D) - \nabla_t(N) \right\|_F^2 \right\}, \tag{6}$$

---

[1] https://www.wikipedia.org/

Table B8: Detailed explanation of 19 tasks included in SpaceVista-1M.

Task	Description
	General Indoor Scenes
Position Comparison	Compare the positions of two objects within or across frames, assessing their spatial relationships in terms of left/right, above/below, and near/far.
Size Comparison	Compare the positions of two objects within or across frames, involving three pairs of size relationships: wider/thinner, taller/shorter, larger/smaller.
Existence Estimation	Determine whether there are objects across frames whose positional/size relationships with the specified object meet the constraint conditions.
Object Counting	Estimate how many objects meet the constraint conditions across frames.
Rotation Estimation	Estimate the rotation angle of an object across multiple frames.
Absolute Distance	Estimate the closest distance between two objects within or across the frames.
Object Size	Estimate the longest dimension of an object within or across the frames.
Route Planning	Choose what action should be performed between a sequence of actions within or across the frames in order to route from a start point to a target.
Appearance Order	Given a video, determine the N-th appearance order of several objects.
Depth Estimation	Estimate the relative or absolute distance of objects from the camera viewpoint in a single image or across multiple images.
View Change Inference	Infer how the camera viewpoint has changed (position and orientation) across the video frames.
Object Matching	Determine whether two objects in the beginning and end frames of a video are the same physical object instance or different instances of the same object type.
Spatial Relation	Analyze and describe the spatial relationships (e.g., support, hanging, adhesion, stacking, encircling, plug-in) between multiple objects or cameras across the frames.
	Indoor Scenes
Every Type in General	All task types from Indoor Scenes can be applied to drone-view perspectives.
Room Size	Estimate the volume of the room(s) across the frames.
	Outdoor Scenes
Every Type in General	All task types from Indoor Scenes apply to Outdoor Scenes except for Room Size estimation.
Navigation	Determine the optimal path or movement strategy to navigate from one location to another across different views (similar to the Route Planning mentioned in Indoor Scenes).
	Drone-View Scenes
Every Type in General	All task types from Indoor Scenes can be applied to drone-view perspectives.
Route Plan	Given a series of aerial images, choose what action should be performed between a sequence of actions in order to route from a start point to a target (similar to the Route Planning mentioned in Indoor Scenes).
Area Estimation	Estimate the size or area of regions or objects from an aerial perspective.
	Tabletop Scenes
Every Type in General	All task types from Indoor Scenes can be applied to drone-view perspectives.
Object Location	Determine the precise position of objects on a table surface, typically corresponding to other objects.
Destination Location	Identify target positions related to single objects (i.e. left, right, front ...) as part of manipulation planning.
Obstacles Location	Identify and locate objects with the AABB box that may interfere with manipulation as part of manipulation planning.
Manipulation Planning	Determine the sequence of actions needed to rearrange objects or achieve a specific configuration on the table.

where M,N represent metric depth model maps and Video-Depth-Anything map . **2) Extreme Scale**: Although the metric depth model is trained on the datasets as DDAD (Guizilini et al., 2020) and NYUv2 (Silberman et al., 2012), it may have a certain level of adaptation to the extreme situations. For extreme situations, including drone-view and tiny objects, it is still necessary to provide a prerequisite to adjust the depth normalization accordingly.

For fine-grained semantic understanding at the pixel level, we leverage the advanced proprietary model DINO-X (Ren et al., 2024) to extract semantic information and bounding boxes for complex scenes, while relying on Grounding DINO (Liu et al., 2023b) for simpler samples. To address cross-frame consistency challenges in video data, we integrate the aforementioned grounding models with SAM2's (Ravi et al., 2024) advanced tracking capabilities, generating temporally consistent masks and unique object IDs across frames based on Grounded-SAM2[2].

By this stage, we obtain a comprehensive understanding of each frame, including bounding boxes, masks, categories, and object IDs, laying a solid foundation for downstream task formulation.

B.3.2 TYPE: DISTANCE

The distance-related tasks, including object size, room size, object distance, and relative distance, rely on depth maps and computer vision techniques to measure object and spatial dimensions from monocular images. The method converts 2D depth keypoints into 3D point clouds using camera calibration parameters and applies Principal Component Analysis (PCA) to extract dimensional information, focusing on objects larger than 20×20 pixels. For object size estimation, the system segments visible objects using instance masks and projects the masked depth values into 3D space. PCA determines the principal axes of the point cloud, with height measured along the vertical axis and width derived from the convex hull of points projected onto the dominant plane. Relative distances are calculated by comparing 3D centroids in world coordinates, and room dimensions are estimated by analyzing the spatial distribution of depth points and identifying major planar surfaces corresponding to walls.

The method uses camera intrinsics and extrinsics to express all measurements in a consistent world coordinate system, addressing the scale ambiguity of monocular systems. Multiple frames are processed to improve robustness, with temporal averaging reducing noise in the estimates. The technique assumes piecewise rigid scenes, operates on standard RGB images, and produces metric-scale measurements. Accuracy depends on the quality of depth estimation and segmentation. Overall, it demonstrates how 2D computer vision pipelines can be extended to 3D measurement tasks through precise geometric reasoning.

B.3.3 TYPE: COUNTING

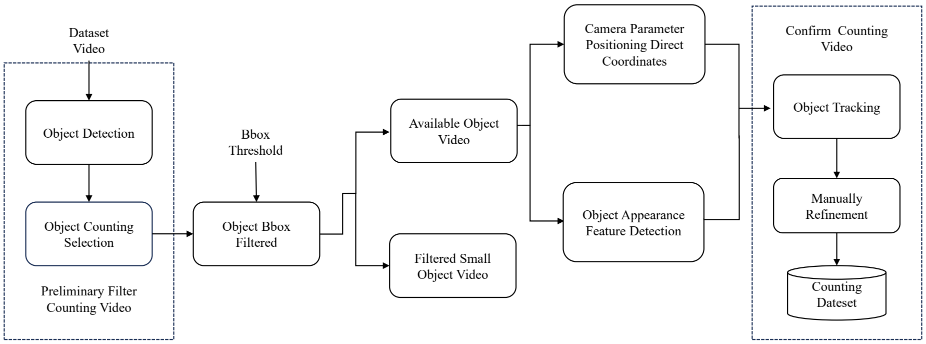

Dataset Video
Object Detection
Object Counting Selection
Preliminary Filter Counting Video
Bbox Threshold
Object Bbox Filtered
Available Object Video
Filtered Small Object Video
Camera Parameter Positioning Direct Coordinates
Object Appearance Feature Detection
Confirm Counting Video
Object Tracking
Manually Refinement
Counting Dateset

Figure B11: Automatic Processing Pipeline for Counting Task Scenes. Through data filtering, object tracking, and counting, the final counting video is obtained after data confirmation.

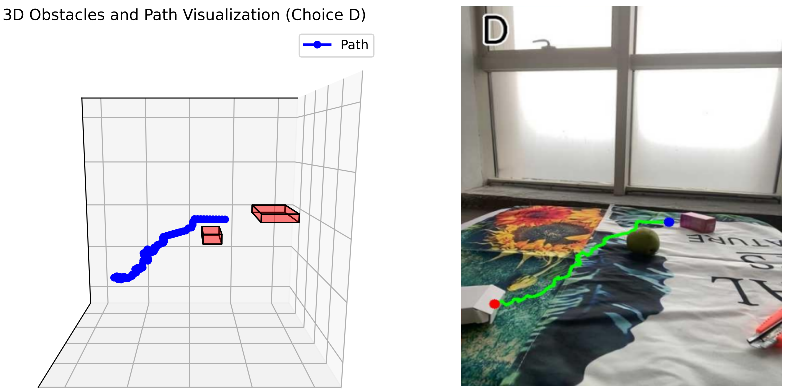
Object counting across real-world scenes faces diverse visual conditions and a high cost of manual labeling, which motivates an automatic pipeline that adapts to scene type. The automatic pipeline addresses object counting through two methodologies tailored to specific scenarios, and Fig. B11

---

[2] https://github.com/IDEA-Research/Grounded-SAM-2

illustrates the workflow that maintains high accuracy while reducing manual effort across indoor, outdoor, and tabletop scenes. For outdoor video sequences, the open-vocabulary detection model (Ren et al., 2024; Cheng et al., 2024) uses text prompts with a confidence threshold of 0.3 for zero-shot detection, projects 2D observations into 3D world coordinates to enforce spatial consistency, and tracks objects via motion prediction with confirmation after at least ten consistent detections. Given the difficulty of reliably detecting very small objects in outdoor scenes and to mitigate ID switching and trajectory fragmentation under severe occlusions, scenes are prefiltered to those containing 2 to 10 objects with a minimum bounding-box size of 32 pixels. For tabletop scenarios, grounding model (Ren et al., 2024; Liu et al., 2023b) and SAM2 (Ravi et al., 2024) are employed, where open-vocabulary detection uses text and bounding box thresholds of 0.4, and mask propagation applies IoU and center distance thresholds of 0.4 and 32 pixels, respectively, to distinguish instances. Both methodologies output object categories and their corresponding counts for each video.

B.3.4 TYPE: PLANNING



(a)3D Path Planning Visualization

(b)Trajectory Execution Process

Figure B12: Visualization of robotic manipulation planning. Fig.(a) visualizes the option for moving the red box to the left of the upper box. Fig.(b) represents the key frame to carry out the manipulation.

In robotic manipulation tasks, effective route planning is essential to ensuring smooth and accurate object movement. The route planning pipeline proceeds as follows. First, depth information and object detection are utilized to identify the category, position, shape, and size of all objects within the image. Subsequently, an arbitrary object is selected as the manipulation source and another as the target position, with the objective being to relocate the source object to a designated position (e.g., front, back, left, right, or above) relative to the target object. Based on this configuration, an LLM generates corresponding manipulation instructions, such as *"What is the correct route of placing the apple on the box"*. Next, the actual spatial positions of the objects are computed using both intrinsic and extrinsic camera parameters. The Rapidly-exploring Random Tree (RRT) (LaValle, 1998; Xu, 2024) algorithm is then employed to plan a collision-free path, where the bounding boxes of objects serve as obstacle constraints during path computation. Finally, two types of data are generated from the planned path: 1) multiple paths are projected onto the camera plane, with the correct trajectory serving as the ground truth answer, and 2) the coordinate variations along the path are translated into natural language instructions via the LLM. For instance, when the x-coordinate of the object decreases while the y-coordinate remains constant in the camera space, the LLM produces the instruction *"move the object to the left."* Fig. B12 demonstrates the visualization of robotic manipulation under the option, showing the planned movement of the red box to the left of the upper box. This figure highlights the spatial relationship and intended positioning within the manipulation task.

B.3.5 TYPE: RELATION

In spatial relation analysis, we combine semantic information with 3D positional data through an automatic reasoning process to ensure consistency in both semantic and spatial aspects. Our analysis operates primarily at the semantic level. We first identify and extract common candidate relations, such as support, attach, insert, and surround. Based on the consistent 3D keypoint semantics established earlier, we generate potential relation pairs that may exhibit these spatial relationships. These candidate pairs are then evaluated for spatial plausibility by integrating 3D positional data with the few-shot prompt through Chain-of-Thought (CoT) reasoning using the foundation model. Finally, the validated pairs are processed by GPT for transformation and answer generation, ensuring semantically and spatially consistent outputs.

B.3.6 DATA POST-PROCESSING

To address the cold-start challenge in SFT, we prioritize the acquisition of explicit "thinking process" rationales—step-by-step explanations that clarify how answers are derived. For example, in object counting, the model is prompted to articulate intermediate reasoning (e.g., *"there are 2 cups on the table and 3 on the chair, totaling 5"*), enriching task understanding and facilitating more robust generalization.

Following common practice (Feng et al., 2025), we acquire high-quality rationales by distilling from advanced open-source and proprietary large models. Specifically, we use Qwen2.5-VL-72B and Gemini-2.5-Pro for complex tasks, and Qwen2.5-VL-32B for simpler ones, balancing reasoning depth with efficiency. We then compare these generated rationales and their corresponding answers with previously collected cases. When GPT answers are different from the answers from previous workflows, we apply a confidence-based filtering strategy to curate the training set, retaining only instances with consistent, well-supported reasoning. This pipeline generates a cleaner, rationale-augmented dataset, mitigating SFT cold-start effects and enhancing downstream performance.

B.3.7 BENCHMARK CONSTRUCTION

Our benchmark comprises two components: **1) Measurement-Related.** For the scale-related portion requiring precise scale annotations, we collect approximately 500 videos across diverse scenes using the two methods described in Appendix B.2.7 and human annotation for other spatial tasks, covering tiny, tabletop, and outdoor settings. For the indoor evaluation set, we instead selected suitable data from ScanNet-based datasets (e.g., VSI-bench and SPAR-bench) and constructed a series of scale-focused questions on top of these bases. **2) Non-Measurement.** For the non-measurement questions, we manually annotate the data collected in the previous step to produce additional spatial reasoning QA pairs. In total, we curate 3,000 fully human-annotated QA pairs for model evaluation.

B.4 DATA STATISTICS

From a visual perspective, our dataset comprises wild scenes spanning scales from millimeters to kilometers. Although the raw dataset contains over 100 million frames, we calculate unsupervised annotations as intermediate information at both the pixel and semantic levels for a curated subset of 10 million frames. These frames vary in resolution from 480p to 2.7K, with frame rates ranging from 24 to 30 fps. During data processing, we preserve the original resolution whenever possible and apply uniform sampling during training as needed.

In terms of the QA component, we employ a combination of templated generation and GPT-based methods to produce 1 million QA pairs with a theoretical duplication rate of only 0.0005%. These pairs are structured into diverse answer formats, including free-form, multiple-choice, and regression-based responses, catering to different analytical needs. Rigorous quality control measures are implemented, with detailed analyses provided in Sec. B.4.7.

In this section, we first conduct a diversity analysis of the visual scenes, examining their composition, categories, and object size distributions (Sec. B.4.1-Sec. B.4.5). We then present a statistical overview of the QA pairs, along with an evaluation of quality control mechanisms (Sec. F.3-Sec. B.4.7). And also, at the beginning of the appendix, Fig. A6 illustrates the data composition for each scene task, where bubble sizes indicate the relative data volume.

B.4.1 TARGET CATEGORY DISTRIBUTION

The introduction of diverse scenarios, such as tabletop, indoor, and outdoor, aims to establish a more inclusive object composition system. Due to the limited drone data, we incorporate drone-view data into the outdoor analysis. By approximating complex object distribution patterns to the real world, this approach enhances the scene adaptation capabilities of visual reasoning models. To quantitatively assess the impact of scene diversity on model generalization, we use the word cloud to compare object distribution characteristics across different scenarios, as shown in Figs. B13–B18. The results reveal that indoor scenes are predominantly composed of rigid objects such as furniture and electronics, exhibiting a highly structured spatial layout. In contrast, outdoor scenes feature more scale-varying objects like vehicles and natural landscapes, demonstrating spatial openness. Meanwhile, tabletop scenes focus on manipulable items such as tools and daily necessities, reflecting precise spatial arrangements. These cross-scene differences provide complementary training samples, effectively mitigating the risk of overfitting to specific scenarios. Thus, the necessity of a multi-scenario strategy to enhance cross-domain generalization is validated.

Overall, each subset scenario differs significantly from the previous indoor-dominated setting, highlighting the diversity of our scenes.

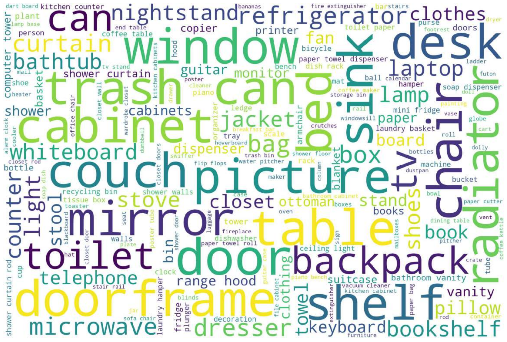


Figure B13: The word cloud of the previous indoor spatial reasoning datasets.

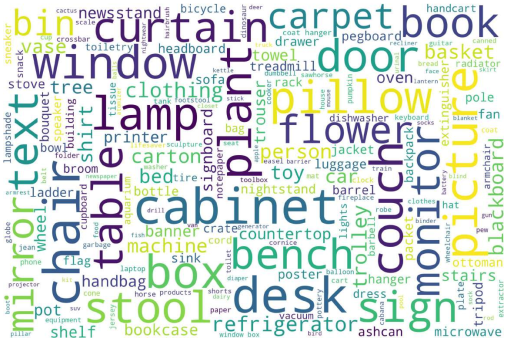


Figure B14: The word cloud of our indoor subset.

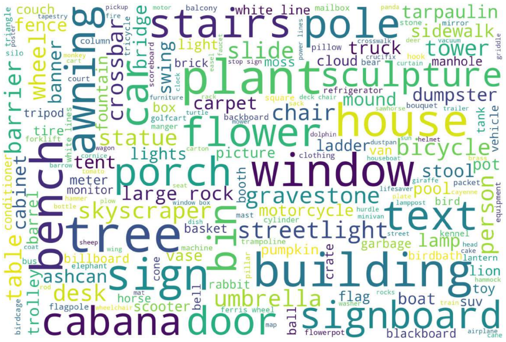


Figure B15: The word cloud of our outdoor subset.

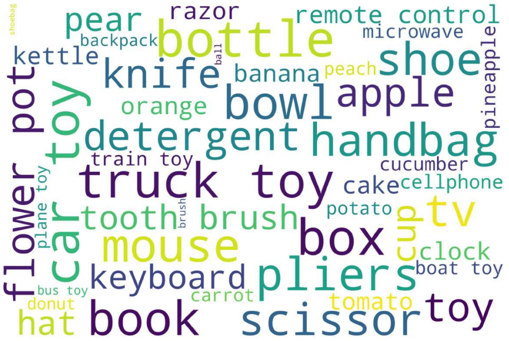


Figure B16: The word cloud of our tabletop subset.

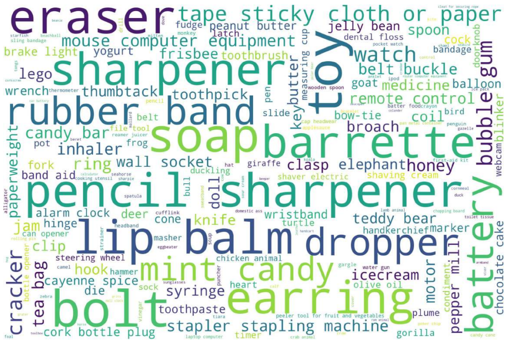


Figure B17: The word cloud of our tiny tabletop subset.

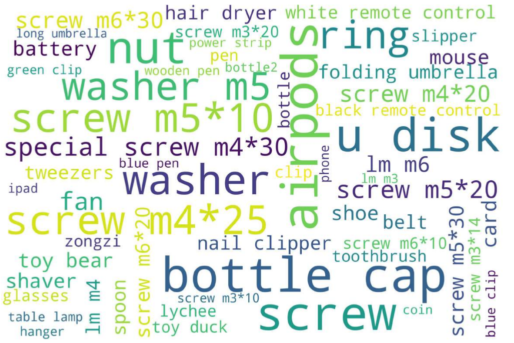


Figure B18: The word cloud of the self-collected subset. Note: We use standard ISO 7046 to denote the models of the screw, which looks like *“m4*10”*.

B.4.2 SCALE DISTRIBUTION

To evaluate the dynamic range of depth across different scenes, we statistically analyze the distributions of the maximum and minimum depth values in each scenario, with the results visualized in Fig. B19. This analysis reveals the variation in extreme depth ranges. Notably, the farthest depth point progressively decreases from the drone scene to the tiny tabletop scene, indicating a consistent reduction in the overall scene. While we can see some extreme values for tiny object scenes, it might be the small object around the window, and extreme depth represents the outside of the window view. It is not unavoidable for data construction and will not affect the overall quality.

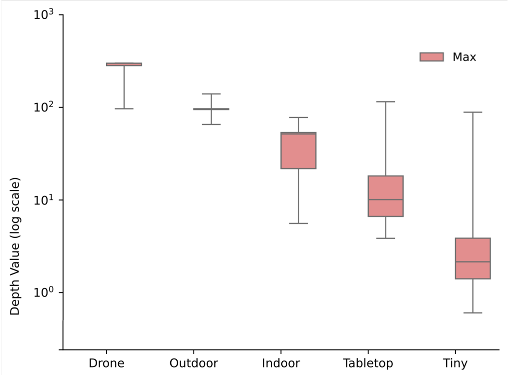


Figure B19: The distribution of the maximum depth value of our dataset. The maximum distance denotes the farthest point observed.

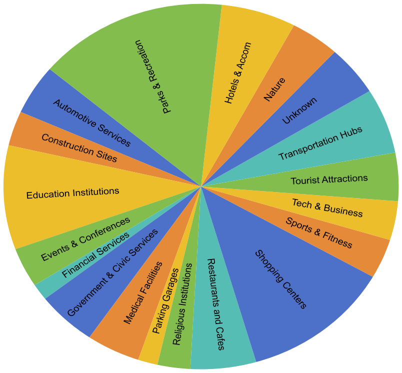


Figure B20: The distribution of the specific sceneries. Note: this chart is just for basic knowledge. Due to the latter filtering policy, there might be some vague or inaccurate analysis.

B.4.3 SUBSCENE TYPE DISTRIBUTION

While our dataset is largely derived from multiple existing sources, we perform a thorough analysis of its scene type diversity. As shown in Fig. B20, the dataset covers a broad range of real-world scenarios, enhancing its complexity and generalizability. To quantify this diversity, we utilize LLM for scene understanding, leveraging object-level annotations from the video data. However, certain subsets, such as partial tabletop scenes and most of the tiny tabletop data, are excluded from the analysis due to limited visual cues. As a result, these statistics primarily illustrate the dataset's variety rather than providing an exact distribution for downstream tasks.

B.4.4 OBJECT SIZE DISTRIBUTION

To enhance spatial understanding at design scales, we analyze the distribution of object sizes in the dataset. The results, shown in Fig. B21, reveal a relatively uniform distribution for objects smaller than 50m, while those exceeding 100m exhibit a certain tail distribution. This trend likely reflects real-world bias in object sizing, with high-rise buildings, common in urban environments, dominating the larger size categories. Consequently, the observed minor long-tail distribution aligns with real-world phenomena and is considered an acceptable characteristic of the dataset.

B.4.5 CAMERA TO OBJECT DISTRIBUTION

To examine biases regarding camera positioning relative to the subject, we analyze the distance (depth) between the camera and the primary object, with the statistical results shown in Fig. B22. The distribution of object-camera distances follows a spindle-shaped pattern, with few instances where the object is positioned closer than 10 cm or farther than 500 m from the camera. This trend is largely influenced by the focusing limitations of most hardware, like lenses, which exhibit reduced sensitivity to objects at extreme distances. Notably, this distribution mirrors that of conventional optical devices in real-world settings and should not be interpreted as a dataset bias.

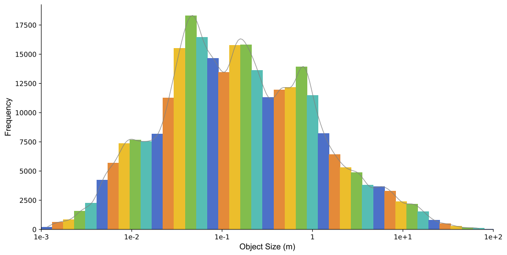


Figure B21: The distribution of the size of the existing objects.

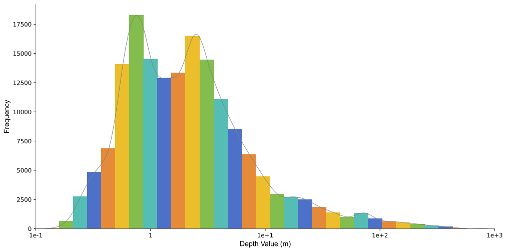


Figure B22: The distribution of the distance between the target object and the camera.

B.4.6 QA STATISTICS ACROSS SCENES

We also provide the overall statistics of SpaceVista-1M dataset in Tab. B9. The SpaceVista-1M dataset consists of approximately 1 million QA pairs, covering a wide range of tasks and scene types across all scales, from tiny tabletop objects to large-scale outdoor and drone-view scenarios, with scales ranging from 1 millimeter to 0.7 kilometers. Its diversity offers extensive challenges for model training and evaluation, enhancing the model's adaptability and reasoning capabilities across different environments.

Table B9: Statistics of QA Pairs for different tasks in SpaceVista-1M.

Task Category	**Total**	**Tiny Tabletop**	**Tabletop**	**Indoor**	**Wild Indoor**	**Outdoor**	**Drone-View**
Scale Distribution	1mm-0.7km	2mm - 5cm	5cm-2m	0.5m-20m	0.3m-50m	0.5m-500m	10m-0.7km
All Scenes	1,014K	79K	242K	162.5K	213.3K	284.3K	33.1K
General Scenes Tasks							
Position Comparison	70.5K	–	10K	22K	18K	20K	0.5K
Size Comparison	88K	–	8K	–	30K	40K	10K
Existence Estimation	82K	15K	25K	–	20K	20K	2K
Rotation Estimation	85.5K	18K	20K	–	22K	25K	0.5K
Relative Distance	81K	–	24K	11K	15K	30K	1K
Absolute Distance	99K	–	25K	26K	13K	34K	1K
Object Counting	21.3K	–	1K	11K	3.5K	5.5K	0.3K
Object Size	157K	15K	30K	33K	38K	34K	7K
Route Plan	2.5K	–	–	-	1K	1K	0.5K
Appearance Order	27.3K	–	4K	15K	3K	4.5K	0.8K
Depth Estimation	102K	19K	32K	10K	15K	23K	3K
View Change Inference	51.7K	6K	27K	8K	4K	6.5K	0.2K
Object Matching	102K	3K	24K	12K	26K	32K	5K
Spatial Relation	19K	-	6K	–	4K	8K	1K
Indoor Scenes Tasks							
Room Size	15.3K	–	–	14.5K	0.8K	–	–
Outdoor Scenes Tasks							
Navigation	0.8K	–	–	–	–	0.8K	–
Drone-View Scenes Tasks							
Area Estimation	0.3K	–	–	–	–	–	0.3K
Tabletop Scenes Tasks							
Obstacles Location	3K	–	3K	–	–	–	–
Manipulation Planning	6K	3K	3K	–	–	–	–

B.4.7 DATA QUALITY CONTROL

During construction of our dataset, we distinguish between two notions of answer correctness: **1) strict correctness**, which requires that an answer conform to objective physical reality, and **2) perceptual correctness**, which requires that an answer align with typical human judgments. Since strict correctness is difficult to 1 for training data derived from in-the-wild videos (due to issues like missing calibration, occlusions, and limited metadata), we adopt the perceptual criterion. Specifically, during validation, we present annotators with both the question and a candidate answer and ask them to judge its acceptability. Consequently, the reported accuracy should be interpreted as agreement with human perception rather than strict fidelity to physical-world quantities or metric scale. For these statistics and the user study, we use MTurk[3] for these statistics and the user study. Dataset tasks and corresponding human checking accuracies are shown in Fig. B10. It is important that perceptual

---

[3] https://www.mturk.com/

correctness is only used in training data quality control, while model evaluation still follows strict correctness.

Table B10: Human checking accuracy over each task category. "∼" means we observe unusual variation for different annotators.

	Task Categories				
Task	Position Comp.	Size Comp.	Existence Est.	Rotation Est.	Relative Dist.
Accuracy	95%	84%	94%	95%	82%
Task	Room Size	Object Count	Object Size	Route Plan	Appear. Order
Accuracy	84%	87%	81%	∼65%	80%
Task	View Change	Object Match	Spatial Rel.	Navigation	Area Est.
Accuracy	96%	93%	95%	∼63%	78%
Task	Manip. Plan	Absolute Dist.	Depth Est.	Obstacles	
Accuracy	73%	84%	95%	67%	

B.4.8 LICENSE

We conduct a systematic review of the open-source licenses for the datasets we use, with the results summarized in Tab. B11. The analysis indicates that CC BY 4.0 and Apache License 2.0 are the most widely adopted. After comprehensive consideration, our SpaceVista-1M dataset adopts the **Creative Commons Attribution (CC BY) 4.0** or **Apache License 2.0** for different sources of data, which is already used by most of the source data.

Table B11: The licenses for the dataset and benchmark included in this paper.

Dataset	**Type**	**License**
	Benchmarks	
VSI-Bench(Yang et al., 2025a)	Indoor	Apache License 2.0
STI-bench(Li et al., 2025e)	Indoor	Apache License 2.0
MMSI-Bench(Yang et al., 2025b)	Indoor	CC BY 4.0
STI-Bench(Li et al., 2025e)	Outdoor, Tabletop	Apache License 2.0
Spar-Bench(Zhang et al., 2025e)	Indoor	Apache License 2.0
SpaceVista-Bench (Ours)	Tiny, Tabletop, Indoor, Outdoor	Apache License License 2.0 & CC BY 4.0
	Training Datasets	
uCO3D(Liu et al., 2025a)	Tiny, Tabletop	CC BY 4.0
SMOT(Park et al., 2020)	Tabletop	Unknown
WildRGBD(Xia et al., 2024)	Tabletop	None
SpaceR(Ouyang et al., 2025)	Indoor	CC BY-NC 4.0
Scannet Series(Yeshwanth et al., 2023)	Indoor	ScanNet Terms of Use
DL3DV(Ling et al., 2024)	Indoor, Outdoor, Drone	DL3DV-10K Terms of Use
SpaceVista-1M (Ours)	Tiny, Tabletop, Outdoor	Apache License License 2.0 & CC BY 4.0

B.5 SUPPLEMENTARY CITATION

Due to the page limit, we have omitted some citations in Tab. 1. Here, we provide a supplementary table of citations.

Table B12: Supplementary citation of Tab. 1

Dataset	**Citation**	**Dataset**	**Citation**
SpaceR	Ouyang et al. (2025)	All-Angles	Yeh et al. (2025)
SPAR-7M	Zhang et al. (2025e)	MVBench	Li et al. (2024b)
Spatial-MLLM	Wu et al. (2025a)	VSI-Bench	Yang et al. (2025a)
InternSpatial	Deng et al. (2025b)	MMSI-Bench	Yang et al. (2025c)
Video-MME	Fu et al. (2024)	SPAR-Bench	Zhang et al. (2025e)
TempCompass	Liu et al. (2024b)	STI-Bench	Li et al. (2025e)

C MODEL DETAIL

C.1 PARAMETER SETTING

SFT. The model architecture is based on Qwen2.5-VL-7B-Instruct, a 7-billion parameter vision-language model capable of processing both images (resized to 100,352 pixels) and videos (16,384 pixels at 16/32 frames). In the ablation study, we use the 3B model for efficiency. For fine-tuning, we employ a selective freezing strategy: while the vision tower and multi-modal projector remain frozen to preserve pretrained visual representations, the language model is fully trainable. Training utilizes full parameter fine-tuning with a DeepSpeed[4] ZeRO-2 configuration for memory optimization. The model is trained on our proposed dataset for spatial understanding in indoor environments, with samples truncated at 32,768 tokens. We implement a cosine learning rate schedule (initial LR=5e-7) with 10% warmup over 2 epochs. We maintain computational efficiency through mixed-precision bfloat16 training.

RL. We conduct our experiments using the Qwen2.5-VL (Bai et al., 2025) on a custom spatial dataset. The training utilizes 7 GPUs with DeepSpeed acceleration and mixed-precision bf16 training with flash attention. Key hyperparameters include a batch size of 1 per device, gradient accumulation steps of 1, an initial learning rate of 1e-6 with cosine scheduling, and weight decay of 0.01. The model processes input sequences up to 16,384 tokens long while generating outputs up to 1,024 tokens. Training runs for 2 epochs with evaluation performed every 200 steps. For inference, we use vLLM on a separate GPU with temperature 1.0 and generate 8 samples per input.

Other Setting. We set the number of experts M to 4 in most cases. We also add LoRA with the same default behavior as PEFT. Additionally, we apply expert scaling factors on a layer-wise basis rather than globally.

Ablation Setting. Unless otherwise noted, we conduct all ablation experiments using the Qwen2.5-VL-3B model because of resource constraints; all other settings are identical to those described above.

C.2 PATCH LEVEL ENCODER ABLATION

We evaluate several visual encoders with dense feature or geometry-aware representations, including VGGT-1B (Deng et al., 2025a)(the only publicly available model) and the generalDINOv3 ViT-Base, and perform ablations on the patch encoder. Tab. C13 reports the performance gains and computational costs associated with each model. Across encoders, DINOv3 achieves more favorable efficiency–accuracy trade-offs with a smaller parameter budget. We attribute this to its self-supervised pretraining, which is not constrained by labeled data and thus confers stronger generalization. In contrast, VGGT exhibits strong reconstruction capabilities but depends on annotations that lack rich semantic content and further relies on a large decoder to recover geometry. Consequently, compared to VGGT, DINOv3 features are more readily consumed by the fusion module, facilitating more effective mapping.

Table C13: Ablation of the patch-level encoder across different sizes of models on the indoor set VSI-Bench based on the same SFT training settings.

Model&Parameter	**Video-Only**	**+VGGT**	**+DINO v3**	**+VGGT +DINO v3**
SpaceVista-3B (Ours)	41.9	43.3	43.5	43 .3
SpaceVista-3B (Ours) *w/o.* fusion module	-	42.0	44.8	44.7
SpaceVista-7B (Ours)	45.0	45.7	46.3	46.0
Extra Parameter	0	909M	303M	1,320M

C.3 LORA LIKE EXPERT ABLATION

On top of the same 3B pretrained base model, we compare three training strategies: **1) Full-parameter Fine-tuning**, **2) Vanilla LoRA**, and **3) LoRA-like Expert**, with the results shown

---

[4] https://github.com/deepspeedai/DeepSpeed

Table C14: Ablation of the LoRA-like expert in the SFT training stage.

Model	**Benchmark**	**w/. Full-parameter Fine-tuning**	**w/. Vanilla LoRA Fine-tuning**	**w/. LoRA-like Expert (model-wise)**	**w/. LoRA-like Expert (layer-wise)**
SpaceVista-3B	VSI-Bench	43.5	42.9	43.9	45.3
	SpaceVista-Bench	29.5	29.4	32.5	33.0
Trainable Parameters		3B	20M	80M+30M	80M+34M

in Fig. C14. We observe that vanilla SFT-based fine-tuning still suffers from latent cross-scale information conflicts. The difference between model-wise and layer-wise is that, for each input, the router is calculated and implemented to the whole model or to separate layers, respectively. In contrast, the model-wise LoRA-like Expert yields clear gains over both full-parameter fine-tuning and vanilla LoRA. Furthermore, scaling to a higher-capacity, layer-wise LoRA-like Expert delivers additional improvements.

D OBSERVATION RESULTS

D.1 GRPO REWARD OBSERVATION

During reinforcement learning training, we observe a relatively stable increase in reward without evidence of reward hacking, as shown in Fig. D23. In most settings, the reward reliably converges within a few thousand environment steps, after which further training yields minimal additional improvements. This suggests that the learning dynamics are well-behaved under our setup and that extending training beyond the convergence point offers limited marginal benefit. Additionally, this may be also treated as the curve of data amount and its performance during post training.

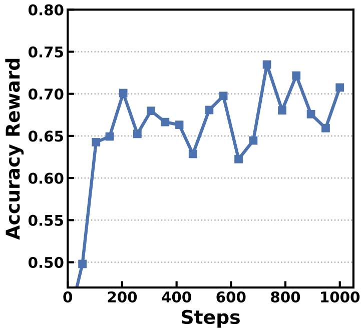


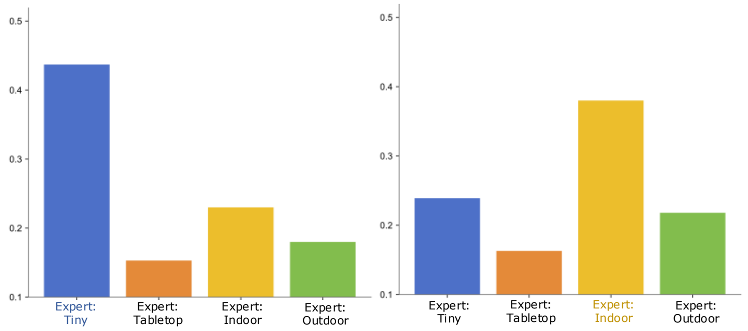
Figure D23: Visualization of GRPO updated and normalized correctness reward chart. This figure visualizes how the reward grows during the RL training stage.

D.2 EXPERT OBSERVATION

We select 10 samples from tiny and indoor scenes and visualize the expert scale distribution in Fig. D24. As shown, inputs from each scene type tend to activate the expert specialized for that scene. This demonstrates the model's ability to distinguish scene-specific characteristics and allocate resources accordingly. By activating the most relevant expert, the model ensures efficient processing and enhanced performance in scene-specific tasks, highlighting its ability to focus on distinct features and patterns within each scene.

D.3 REASONING VS MEMORIZING (OUT-OF-DISTRIBUTION PROBLEM)

In our experiments, we observe that models often exhibit a strong bias toward memorizing fixed sizes for certain objects—for instance, chairs are typically assumed to be 50-70 cm tall. Consequently, the network tends to rely on memorized size priors rather than reasoning about object scale. However, this phenomenon presents a dual nature. On one hand, human perception of size and scale also



(a) Normalized distribution of the selected tiny scale samples

(b) Normalized distribution of the selected indoor scale samples

Figure D24: Visualization of the normalized scale of each expert with different selected samples. It reflects the model's capacity to allocate resources according to the inherent properties of each scene.

depends on reference objects and familiar benchmarks, which are essential for intuitive understanding. On the other hand, since real-world spatial relationships can vary significantly, such biases may lead to erroneous judgments in atypical cases.

We argue there is two types of Out-of-Distribution (OOD) that should be discussed separately. 1) **OOD category with normal size** 2) **normal category with OOD size**.

For **OOD category with normal size**, to systematically evaluate the impact of this bias and its potential implications for advancing the field, we design three specialized subsets at the same scale:

- **Seen Set:** Common object categories from the training distribution (i.e., bicycle, table, chair).
- **Seen Set with Various Scales:** bjects of the same category (i.e., different sizes and shapes of screw).
- **Unseen Set:** Rare or culturally specific objects requiring contextual size reasoning (i.e., ethnic items with regional characteristics, such as a traditional food).

The Seen Set provides baseline performance metrics for familiar objects but may overlook biases due to training conformity. The Seen Set with scale variety directly probes size generalization for known categories, but it is limited to variations within seen objects. The Unseen Set evaluates robustness to novel, culturally diverse scenarios but risks introducing confounders beyond scale bias. Collectively, these subsets balance ecological validity with experimental control, offering a comprehensive framework to diagnose size-related biases. This structured approach enables us to analyze how size biases manifest under different conditions, combining ecological validity with controlled experimentation. As shown in Fig. D15, all-scale training benefits the overall reasoning model; however, the general models still tend to memorize the regular size of the target object.

Table D15: Reasoning VS memorizing analysis of different subsets.

Model	Seen Set (Normal)	Seen Set (Various Scales)	Unseen Set
Qwen2.5-VL-3B-Instruct	35.7	34.7	23.1
Qwen2.5-VL-7B-Instruct	37.0	38.9	28.0
SpaceVista-7B (Ours)	**37.3**	**41.0**	**32.8**

For **normal category with OOD size**, we need to develop a dataset with precise annotation. The Guinness World Records (GWR) is a globally recognized organization[5] that catalogs uncommon objects and forms. We obtain precise size measurements along with the corresponding images/videos, and construct a series of QA pairs about object sizes as shown in Fig. D25. The GWR data comprises diverse scenes, including outdoor, indoor, and drone, with over 50 images and over 50 questions.

---

[5] https://www.guinnessworldrecords.com/records/showcase

Table D16: Performance comparison across GWR dataset.

Size-Related QA	Qwen2.5VL-7B	Qwen2.5VL-3B	SpaceVista-7B	SpaceVista-3B
SpaceVista-Bench	49.9	44.0	58.3	49.3
GWR set	27.8	23.1	31.1	27.3

Because only a small portion of the records is documented on the website, we used nearly all available website content to construct this GWR test set. All questions were created through human annotation to ensure dataset quality. This data is used solely for insight and analysis, not for official purposes. The licensing status of GWR content is unclear. If the license permits, we will release this GWR set on Hugging Face.

Figure D25: The Guinness World Records (GWR) is a globally recognized organization that catalogs uncommon objects and forms. We scraped precise size measurements along with the corresponding images/videos, and constructed a series of QA pairs about object sizes.

As shown in Table D16, we evaluate the popular Qwen2.5-VL model and our SpaceVista-7B model. Because the GWR data contain only size-related questions, we select the size-related subset of SpaceVista-Bench to ensure a fair comparison. We find that these OOD data are challenging for both the general-purpose model and our specialist model. However, the OOD challenge does not produce a clear performance gap between Qwen2.5-VL and SpaceVista. Although our model is not designed for purely image-based tasks, this potential bias suggests a promising direction for future work in VLLMs.

Our analysis of potential bias has two parts:

1. **Depth Knowledge.** Current metric depth models estimate distance primarily based on accurate camera parameters, such as focal length. These parameters vary across different scales, which is why our model performs slightly better than a general model.
2. **Scale Prior.** Human distance estimation also strongly relies on reference objects (i.e., scale priors in question). When these references are unusual, humans also unavoidably exhibit bias. Thus, scale priors are a double-edged sword and cannot be simply described as good or bad.

D.4 DETAILED ANALYSIS ON EACH BENCHMARK

We conduct a comprehensive evaluation of SpaceVista-7B across multiple benchmarks, including STI-Bench (Li et al., 2025e), SPAR-Bench (Zhang et al., 2025e), MMSI-Bench (Yang et al., 2025b) and VSI-Bench (Yang et al., 2025a). In this section, we analyze SpaceVista-7B's performance on each benchmark and compare it to other state-of-the-art models. The results from these benchmarks provide a thorough assessment of SpaceVista-7B's spatial reasoning capabilities, highlighting its versatility and adaptability across diverse tasks.

Table D17: Performance comparison of our SpaceVista-7B and other baselines on STI-Bench.We use **bold** and underlined text for the top two within open-source categories, while ranks are computed across all model categories. In Static Understanding, "Dim. Meas." refers to Dimensional Measurement. In Dynamic Understanding, "Disp. & P.L.", "Speed & Acc.", "Ego Orient.", "Traj. Desc.", and "Pose Est." represent Displacement and Path Length, Speed and Acceleration, Ego-Centric Orientation, Trajectory Description, and Pose Estimation, respectively. This table includes only the popular model for which the detailed scores are available. For average-score comparisons, see Table 2.

Model/Method	**Rank**	**Avg.**	**Static Understanding**			**Dynamic Understanding**				
			Dim. Meas.	Spatial Relation	3D Video Grounding	Disp. & P.L.	Speed & Acc.	Ego Orient.	Traj. Desc.	Pose Est.
Closed-source Models										
GPT-4o (Hurst et al., 2024)	8	34.8	27.1	51.8	29.0	23.2	35.4	33.7	32.0	53.6
Gemini-2.0-Flash (Deepmind, 2024)	3	38.7	31.9	50.0	31.8	27.7	32.1	10.8	38.5	61.3
Claude-3.7-Sonnet (Anthropic, 2025a)	2	40.5	29.8	45.5	35.7	28.9	38.8	40.0	47.4	62.6
Gemini-2.5-Pro (DeepMind, 2025)	1	41.4	38.7	53.8	36.9	33.9	33.1	52.5	47.4	50.4
Open-source Models										
VideoLLaMA3-7B (Zhang et al., 2025a)	7	35.2	29.4	48.6	36.1	21.5	36.7	23.2	**54.6**	48.1
MiniCPM-V-2.6 (Yao et al., 2024)	10	26.9	27.7	44.5	29.0	19.0	25.7	7.0	30.8	35.6
VideoChat-R1 (Li et al., 2025d)	9	32.8	23.2	47.3	31.5	22.4	31.1	26.0	47.9	48.3
InternVL2.5-78B (Chen et al., 2024e)	4	38.5	29.9	**52.8**	31.6	24.9	37.2	**49.2**	43.6	**53.6**
VideoChat-Flash (Li et al., 2024c)	6	36.3	**33.6**	51.4	33.1	**27.1**	32.3	22.2	54.2	51.4
SpaceVista-7B (Ours)	5	38.2	33.1	47.2	**37.6**	23.6	**37.3**	39.6	43.1	51.2

Table D18: Performance comparison of our SpaceVista-7B and other baselines on SPAR-Bench.We use **bold** and underlined text for the top two within open-source categories, while ranks are computed across all model categories. OO, OC, and MV refer to object-object, object-camera, and multi-view, respectively. This table includes only the popular model for which the detailed scores are available. For average-score comparisons, see Table 2.

Model/Method	**Rank**	**Avg.**	**Low**	*Depth-OC*	*Depth-OC-MV*	*Depth-OO*	*Depth-OO-MV*	*Dist-OC*	*Dist-OC-MV*	*Dist-OO*	*Dist-OO-MV*	**Medium**	PosMatch	CamMotion	ViewChgI	**High**	DistI-OO	DistI-OO-MV	ObjRel-OC-MV	ObjRel-OO	ObjRel-OO-MV	SpImag-OC	SpImag-OC-MV	SpImag-OO	SpImag-OO-MV
Baseline																									
Chance Level (Random)	-	-	-	-	-	-	-	-	-	-	-	-	22.65	24.50	-	25.09	23.82	22.02	31.25	25.27	22.16	25.81	24.42	24.17	26.89
Chance Level (Frequency)	5	32.74	31.19	43.09	43.51	17.38	13.05	41.90	30.99	27.40	32.17	38.25	29.01	26.75	59.00	32.29	52.94	50.60	28.25	26.92	26.59	26.34	26.74	26.49	25.77
SPAR-Bench(full)																									
InternVL2-2B (Chen et al., 2024f)	12	28.06	21.74	18.06	24.81	23.20	**20.97**	19.47	19.95	26.83	20.61	22.83	39.69	23.00	5.81	35.42	51.18	55.95	46.00	31.59	23.82	36.02	34.30	17.55	22.41
InternVL2-4B (Chen et al., 2024f)	6	32.01	28.94	23.94	27.22	20.00	18.12	42.57	40.16	31.29	28.18	29.16	49.87	21.00	16.62	35.70	56.76	55.36	40.25	36.81	25.21	28.76	32.27	21.19	24.65
InternVL2-8B (Chen et al., 2024f)	4	33.02	26.83	25.75	30.88	20.76	20.78	39.03	36.19	19.15	22.19	**36.49**	**63.36**	28.00	18.11	37.37	64.71	54.46	42.75	37.36	26.32	34.14	31.10	20.86	24.65
InternVL2.5-2B (Chen et al., 2024e)	10	30.14	25.79	39.67	39.72	12.12	15.03	30.94	29.59	20.22	19.02	22.93	37.91	24.25	6.64	36.41	51.47	56.85	50.25	33.79	24.10	27.15	35.17	26.49	22.41
InternVL2.5-4B (Chen et al., 2024e)	9	30.55	25.66	29.06	32.97	21.77	16.83	20.84	26.85	28.13	28.79	29.75	47.07	33.25	8.92	35.16	54.12	58.93	35.50	29.67	34.63	24.73	31.39	19.21	28.29
InternVL2.5-8B (Chen et al., 2024e)	2	36.28	29.46	25.78	29.31	23.79	18.76	46.82	42.68	22.62	25.89	31.88	61.32	28.00	6.32	43.80	59.71	56.85	51.75	44.23	**41.55**	36.56	**41.57**	22.52	39.50
LLaVA-Onevision-0.5B (Li et al., 2024a)	11	29.48	30.14	49.22	42.72	18.04	14.92	31.48	25.67	28.98	30.10	15.89	24.43	21.75	1.50	33.42	50.88	50.00	32.00	27.75	26.04	30.91	34.01	24.50	24.65
LLaVA-Onevision-7B (Li et al., 2024a)	7	31.20	21.79	30.33	26.94	18.58	13.87	10.43	13.64	31.24	29.29	26.13	38.68	30.25	9.47	40.14	56.47	55.06	37.25	48.63	38.23	30.38	33.72	26.49	35.01
Qwen2-VL-2B (Wang et al., 2024b)	13	24.60	19.43	38.03	40.63	18.84	14.09	7.81	7.07	17.82	11.14	27.55	26.21	25.25	**31.20**	28.22	54.12	49.11	21.75	25.27	12.47	23.92	27.62	24.83	14.85
Qwen2-VL-7B (Wang et al., 2024b)	8	30.74	27.52	35.97	35.22	20.83	12.88	28.68	29.95	28.21	28.45	20.44	35.37	20.25	5.69	37.03	59.71	52.38	30.25	38.46	41.00	22.04	28.49	22.52	38.38
Qwen2.5-VL-7b (Bai et al., 2025)	3	33.07	28.75	31.33	33.66	21.99	14.97	42.88	37.73	23.83	23.64	22.97	33.33	28.75	6.83	40.27	58.24	51.49	44.75	**50.00**	32.13	33.87	32.85	**27.15**	31.93
LLaVA-v1.5-7b (Liu et al., 2023a)	14	23.65	10.85	5.17	12.53	17.37	11.34	7.25	5.26	18.73	9.12	26.50	24.43	26.75	28.31	34.09	51.18	52.38	34.25	24.18	26.87	34.68	29.94	22.52	30.81
LLaVA-v1.6-7b (Liu et al., 2023a)	15	13.21	8.53	12.14	0.00	20.35	0.27	10.76	0.41	24.27	0.00	4.79	6.62	7.75	0.00	20.18	51.76	7.74	6.25	32.14	6.37	**39.52**	10.47	21.52	5.88
SpaceVista-7B (Ours)	1	41.68	**42.51**	**57.78**	**51.94**	**24.44**	20.22	**57.02**	**51.12**	**42.62**	**34.98**	36.02	31.04	**41.00**	0.00	**46.82**	**66.76**	**63.10**	**56.5**	**50.00**	**41.55**	37.10	37.21	**27.15**	**42.02**

On **STI-Bench**, SpaceVista-7B ranks fifth overall and exhibits strong performance on 3D video grounding as well as speed and acceleration estimation. It achieves 37.6% on 3D video grounding and 37.3% on speed-related tasks. Gemini-2.5-Pro (DeepMind, 2025) attains the highest average score of 41.4%, followed by Claude-3.7-Sonnet (Anthropic, 2025a). In contrast, Ego-Centric Orientation, Trajectory Description, and Displacement and Path Length remain highly challenging, as they require accurate modeling of egocentric camera motion, long-range temporal integration, and stable 3D reasoning under viewpoint changes and occlusions. Dynamic, long-term spatiotemporal reasoning remains a challenge for current vision-language models. The evaluation results are presented in Tab. D17.

SpaceVista-7B attains the highest overall performance among all compared models on **SPAR-Bench**, with an average accuracy of 41.68% and rank 1. SPAR-Bench evaluates spatial compositional reasoning over object–object(OO), object–camera(OC), and multi-view(MV) relations under low, medium, and high difficulty settings. Across all difficulty levels, SpaceVista-7B consistently ranks within the top two, and on the most challenging OC and MV subsets it reaches up to 66.76%, indicating robust modeling of complex object–camera relations under large viewpoint changes, as summarized in Tab. D18. Meanwhile, most OO subsets remain highly challenging for all models, and reasoning about fine-grained multi-object spatial relations in heavily occluded scenes with subtle depth and ordering differences is still problematic.

On **MMSI-Bench**, SpaceVista-7B achieves an average accuracy of 30.7% and ranks fifth overall, representing the strongest performance among all open-source models. It performs particularly well on positional-relationship tasks, such as camera–object reasoning with 45.3%, and maintains competitive results on attribute and motion categories, as summarized in Tab. D19, indicating a reasonably balanced multi-dimensional spatial understanding. Nevertheless, all models, including SpaceVista-7B, remain far below the human upper bound of 97.0%, and sub-tasks involving camera motion and the composite MSR metric are still notably difficult.

Finally, in the **VSI-Bench** evaluation, SpaceVista-7B outperforms all other models, excelling in object counting, appearance sequencing, and absolute distance tasks, achieving 62.9% in object counting and 36.0% in absolute distance, surpassing several open-source models, including LLaVA-Video-72B (Zhang et al., 2024c) and LLaVA-OneVision-72B (Li et al., 2024a). The results of this evaluation are shown in Tab. D20.

Table D19: Performance Comparison of our SpaceVista-7M and other baselines on MMSI-Bench.We use **bold** and <u>underlined text</u> for the top two within open-source categories, while ranks are computed across all model categories. Cam., Obj., Reg., Meas., and Appr. denote Camera, Object, Region, Measurement, and Appearance, respectively. This table includes only the popular model for which the detailed scores are available. For average-score comparisons, see Table 2.

Model/Method	**Rank**	**Avg.**	**Positional Relationship**						**Attribute**		**Motion**		**MSR**
			Cam.-Cam.	Obj.-Obj.	Reg.-Reg.	Cam.-Obj.	Obj.-Reg.	Cam.-Reg.	Meas.	Appr.	Cam.	Obj.	–
Baseline													
Blind GPT-4o	32	22.7	20.2	17.0	29.6	13.9	29.4	19.2	21.8	12.1	20.2	29.0	20.2
Random Guessing	29	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Human Level	1	97.2	95.7	98.9	97.5	94.2	98.8	96.4	95.3	98.5	98.6	98.7	97.0
Closed-source Models													
o3 (OpenAI, 2025b)	2	41.0	45.2	39.4	37.0	44.2	47.1	62.6	54.7	28.8	31.1	32.9	34.9
GPT-4.5 (OpenAI, 2025a)	3	40.3	34.4	29.8	39.5	51.2	47.1	55.4	39.1	33.3	41.9	40.8	36.4
GPT-4o (Hurst et al., 2024)	7	30.3	34.4	24.5	23.5	19.8	37.6	27.7	32.8	31.8	35.1	36.8	30.8
Gemini-2.5-Pro (DeepMind, 2025)	4	36.9	39.7	31.9	39.5	45.3	35.2	43.3	51.5	21.2	36.4	30.2	34.3
Claude-3.7-Sonnet (Anthropic, 2025a)	10	28.7	32.3	26.6	22.2	34.9	37.6	42.2	25.0	22.7	21.6	32.9	22.7
Seed1.5-VL (Guo et al., 2025b)	8	29.7	32.2	30.8	25.9	23.2	38.8	32.5	39.0	21.2	36.4	25.0	26.2
Open-source Models													
InternVL3-78B (Zhu et al., 2025)	12	28.5	<u>34.4</u>	23.4	32.1	12.8	<u>37.6</u>	26.5	37.5	19.7	**28.4**	31.6	29.3
InternVL2.5-78B (Chen et al., 2024e)	12	28.5	23.7	22.3	**39.5**	29.1	31.8	**42.2**	35.9	19.7	17.6	26.3	27.3
Qwen2.5-VL-72B (Bai et al., 2025)	5	30.7	25.8	34.0	34.6	23.3	34.1	36.1	**45.3**	27.3	<u>27.0</u>	30.3	27.3
LLaVA-OneVision-72B (Li et al., 2024a)	13	28.4	**43.0**	31.9	33.3	30.2	<u>37.6</u>	<u>38.6</u>	28.1	19.7	13.5	32.9	15.7
InternVL3-38B (Zhu et al., 2025)	23	26.3	21.5	20.2	33.3	23.3	35.3	25.3	39.1	21.2	16.2	31.6	25.8
InternVL2.5-38B (Chen et al., 2024e)	16	27.9	18.3	22.3	35.8	22.1	**38.8**	34.9	37.5	25.8	14.9	<u>38.2</u>	25.3
Qwen2.5-VL-32B (Bai et al., 2025)	17	27.7	24.7	26.6	29.6	22.1	32.9	31.3	31.2	24.2	18.9	35.5	27.8
InternVL2.5-26B (Chen et al., 2024e)	15	28.0	24.7	19.1	29.6	33.7	31.8	37.3	35.9	30.3	10.8	31.6	26.8
NVILA-15B (Liu et al., 2024c)	6	30.5	30.1	**39.4**	28.4	36.0	**38.8**	20.5	29.7	<u>31.8</u>	18.9	35.5	27.8
InternVL3-14B (Zhu et al., 2025)	20	26.8	19.4	24.5	24.7	23.3	<u>37.6</u>	24.1	31.2	22.7	24.3	31.6	29.3
Llama-3.2-11B-Vision (Grattafiori et al., 2024)	27	25.4	25.8	30.8	32.0	25.6	21.2	25.9	20.3	19.7	25.6	28.9	19.2
InternVL3-9B (Zhu et al., 2025)	21	26.7	18.3	25.5	32.1	29.1	31.8	22.9	29.7	24.2	16.2	<u>38.2</u>	26.8
InternVL3-8B (Zhu et al., 2025)	26	25.7	25.8	31.9	<u>37.0</u>	25.6	35.3	28.9	23.4	24.2	16.2	32.9	14.6
InternVL2.5-8B (Chen et al., 2024e)	10	28.7	32.3	27.7	29.6	32.6	24.7	32.5	26.6	27.3	16.2	31.6	<u>30.3</u>
NVILA-8B (Liu et al., 2024c)	14	28.1	17.2	29.8	24.7	30.2	22.4	34.9	34.4	25.8	25.7	34.2	29.8
Qwen2.5-VL-7B (Bai et al., 2025)	25	25.9	24.7	24.5	24.7	25.6	29.4	26.5	25.0	18.2	20.3	**39.5**	25.8
LLaVA-OneVision-7B (Li et al., 2024a)	30	24.5	20.4	33.0	29.6	29.1	25.9	30.1	29.7	25.8	18.9	34.2	11.6
InternVL2.5-4B (Chen et al., 2024e)	23	26.3	31.2	23.4	21.0	31.4	34.1	25.3	23.4	24.2	13.5	31.6	**36.8**
Qwen2.5-VL-3B (Bai et al., 2025)	22	26.5	26.9	27.7	30.9	29.1	28.2	34.9	31.2	16.7	17.6	27.6	23.2
InternVL3-2B (Zhu et al., 2025)	28	25.3	26.9	25.5	29.6	31.4	28.2	27.7	26.6	22.7	12.2	23.7	23.7
InternVL2.5-2B (Chen et al., 2024e)	9	29.0	28.0	27.7	24.7	<u>37.2</u>	29.4	36.1	<u>43.8</u>	15.2	21.6	31.6	26.8
InternVL3-1B (Zhu et al., 2025)	19	27.0	24.7	<u>35.1</u>	22.2	30.2	29.4	30.1	32.8	28.8	17.6	19.7	26.3
InternVL2.5-1B (Chen et al., 2024e)	24	26.1	23.7	26.6	24.7	25.6	31.8	25.3	31.2	30.3	17.6	25.0	26.3
DeepSeek-VL2 (Wu et al., 2024b)	18	27.1	23.7	31.9	22.2	36.0	30.6	22.9	28.1	15.2	**28.4**	26.3	28.3
DeepSeek-VL2-Small (Wu et al., 2024b)	11	28.6	24.7	28.7	18.5	33.7	**38.8**	27.7	28.1	**33.3**	24.3	25.0	29.8
DeepSeek-VL2-Tiny (Wu et al., 2024b)	31	24.0	29.0	27.7	21.0	23.3	17.6	31.3	14.1	24.2	14.9	25.0	27.3
SpaceVista-7B (Ours)	5	30.7	26.9	23.2	30.9	**45.3**	27.1	36.1	34.4	26.7	23.3	35.5	25.8

In general, breakthroughs in specialized domains tend to lead to a decline in general VLM capabilities. This phenomenon has been widely explored in mathematical reasoning, code reasoning, and spatial reasoning. To analyze general ability, we evaluate the performance of SpaceVista-7B on the widely accepted video benchmark Video-MME (Fu et al., 2025). Video-MME is a full-spectrum, multi-modal benchmark of MLLMs in general video analysis. The comparison is shown as Tab. D21.

Table D20: Performance comparison of our SpaceVista-7B and other baselines on VSI-Bench.We use **bold** and underlined text for the top two within open-source categories, while ranks are computed across all model categories. This table includes only the popular model for which the detailed scores are available. For average-score comparisons, see Table 2.

Model / Method	Rank	Avg.	Obj Appear ance Order	Object Abs Distance	Object Counting	Object Rel Distance	Object Size Estimation	Room Size Estimation	Route Planning	Object Rel Direction
Proprietary Models(API)										
GPT-4o(Hurst et al., 2024)	10	34.0	28.5	5.3	46.2	37.0	43.8	38.2	31.5	41.3
Gemini-1.5 Flash (API)(Team et al., 2024)	3	42.1	37.8	30.8	49.8	37.7	53.5	54.4	31.5	41.0
Gemini-1.5 Pro (API)Team et al. (2024)	2	45.4	34.6	30.9	56.2	51.3	64.1	43.6	36.0	46.3
Open-source Models										
InternVL2-2B(Chen et al., 2024f)	16	26.5	6.3	24.0	25.7	32.1	20.0	29.2	30.4	44.1
InternVL2-8B(Chen et al., 2024f)	6	37.5	46.4	29.0	31.3	38.0	48.9	**44.2**	28.9	33.4
InternVL2-40B(Chen et al., 2024f)	7	37.0	44.7	26.2	41.3	**47.6**	48.2	27.5	27.8	32.7
LongVILA-8B(Chen et al., 2024d)	17	21.6	25.5	9.1	29.1	29.6	16.7	0.0	32.5	30.7
VILA-1.5-8B(Lin et al., 2023)	14	28.9	24.8	21.8	17.4	32.1	50.3	18.8	31.0	34.8
VILA-1.5-40B(Lin et al., 2023)	12	31.2	32.9	24.8	22.4	40.5	48.7	22.7	31.5	25.7
LongVA-7B(Zhang et al., 2024a)	13	29.2	15.7	16.6	38.0	33.1	38.9	22.2	25.4	43.3
LLaVA-Video-7B(Zhang et al., 2024c)	8	35.6	30.6	14.0	48.5	43.5	47.8	24.2	34.0	42.4
LLaVA-Video-72B(Zhang et al., 2024c)	4	40.9	48.6	22.8	48.9	42.4	57.4	35.3	35.0	36.7
LLaVA-NeXT-Video-7B(Zhang et al., 2024b)	8	35.6	30.6	14.0	48.5	43.5	47.8	24.2	34.0	42.4
LLaVA-NeXT-Video-72B(Zhang et al., 2024b)	4	40.9	48.6	22.8	48.9	42.4	57.4	35.3	35.0	36.7
LLaVA-OneVision-0.5B(Li et al., 2024a)	15	28.0	5.8	28.4	46.1	28.3	15.4	28.3	34.5	36.9
LLaVA-OneVision-7B(Li et al., 2024a)	11	32.4	24.4	20.2	47.7	42.5	47.4	12.3	29.4	35.2
LLaVA-OneVision-72B(Li et al., 2024a)	5	40.2	44.6	23.9	43.5	42.5	57.6	37.5	32.5	39.9
Qwen2.5-VL-7B (Bai et al., 2025)	9	34.4	32.7	17.5	34.0	35.8	51.9	36.6	29.4	37.7
SpaceVista-7B (Ours)	1	48.6	**56.3**	**36.0**	**62.9**	44.2	**58.1**	42.0	**38.9**	**49.7**

Table D21: General ability on popular video benchmark Video-MME

Model	Video-MME
VideoLLaMA2	47.9
LLaVA-OneVision-7B	58.2
Qwen2.5VL-7B	63.8
InternVL3-8B	65.3
VG-LLM-8B (Spatial Model)	59.3
Qwen2.5VL-7B ($w/.$ 1/5 SpaceVista-1M)	59.1
SpaceVista-7B (Spatial Model)	59.6

Therefore, we consider our SpaceVista general ability comparable, and also don't believe it has "lost" general ability or merely follows a preset spatial template. It is still undeniable that specialist models are inspiring for future explorations of general MLLMs.

D.5 THE HARDEST SCENE

Table D22: Results analysis of different scenes. The model mentioned below is trained in a balanced subset of SpaceVista-1M for better control of experiment conditions.

Model	SpaceVista-Bench (Ours)			
	Indoor	Outdoor	Tabletop	Tabletop
Qwen2.5-VL-7B	30.34	18.31	23.79	19.37
$w/.$ balance training	38.77	24.90	30.17	20.86

When testing scenes at varying scales, several critical questions arise: Which scenarios pose greater challenges, and to what extent is data complexity the primary bottleneck? To systematically investigate these issues, we design a controlled observational experiment.

We identify tasks that exhibit consistent properties across different scales, including object size, object comparison, absolute and relative distance, and depth estimation. For fairness in comparison, we train models using videos from diverse scenes while maintaining similar quantities of QA pairs and video samples. Under these controlled conditions, we evaluate and compared performance across different scale-dependent scenarios. In Tab.D22, it seems indoor data is the easiest task. We hypothesize that a human-scale estimation bias—arising because both humans and GPT focus on objects expressible in basic units like meters in pretraining corpora—leads to this preference.

D.6 WHY 2.5D>3D

Table D23: Comparison of the robustness of the model training of 3D and 2.5D. All the models are trained on 3D or 2.5D data along with the video. However, we vary the evaluation input of these models to see the robustness. "–" denotes experiments we consider unnecessary. "low" means using low resolution visual for 3D reconstruction. This table includes only the popular model for which a detailed score is available. For average-score comparisons, see Table 2. "($n\%$)" means the relative decrease compared to the original input.

Settings	**Eval Input**	**VSI-bench**	**SpaceVista-Bench**
Training with $w/.$ 3D	visual $w/.$ 3D	44.3	31.4
	visual $w/.$ 3D (low)	38.1 (-14%)	–
	visual $w/o.$ 3D	34.0 (-23%)	–
Training with $w/.$ 2.5D	visual $w/.$ 2.5D	45.6	33.0
	visual $w/.$ 2.5D (low)	43.9 (-4%)	32.3 (-2%)
	visual $w/o.$ 2.5D	40.7 (-10%)	29.1(-12%)

In addition to introducing VGGT(Wang et al., 2025a) and DINO v3(Siméoni et al., 2025) as extra signals, we conduct a series of targeted ablation studies. This suggests that representation formats like VGGT, when used in their native encoder output, are wonderful for capturing geometry information, but suboptimal for capturing semantic information or overall scenes, especially for low resolution and uncommon scenarios. In Tab.D23, we use "3D" to denote the pure geometric features from VGGT, and "2.5D" to denote the additional 12 viewing angles of the overall scene rendered by the decoder and the renderer. We use the special prompt and the image token to provide

As shown in Tab.D23, 2.5D is usually more robust in spatial reasoning. Rendering to 2.5D enables effective exploitation of pretrained image tokenizers, which in turn provides more reliable semantic information.

Below is the special prompt for 2.5D finetuning.

"*Please think about this question as if you were a human pondering deeply. Consider detailed information from the video frames and coarse spatial information from the 3D point cloud image. Provide the model's thought process and reasoning between the <think> </think> tags, and give your final answer between the <answer> </answer> tags. <video> The images below are obtained from the 3D point clouds based on the video frames above. The following point cloud images are randomly selected viewpoints; some may be completely unhelpful, while others may contain important information. Please discern carefully. <image> Provide your reasoning between the <think> </think> tags and your final answer between the <answer> </answer> tags.*"

D.7 SCALING-UP ANALYSIS

We investigate prospective scaling behavior across three model sizes—3B, 7B, and 32B—to inform future model development. Our analysis is conducted using the same SpaceVista-1M dataset while holding all model settings nearly constant. However, there is a minor difference between different scale models. We use LoRA rather than full scale to finetune 32B model. Since using more experts will inevitably increase the inference time, we use fewer experts as the scale increases. However, we still hold the strong belief that it does not affect the overall scaling exploration of our SpaceVista-1M data and model.

Table D24: Scaling model with SpaceVista-1M. "Qwen2.5-VL-*B" indicates that the SFT model used for evaluation is trained on the corresponding base model.

Foundation Model	**Qwen2.5-VL-3B**	**Qwen2.5-VL-7B**	**Qwen2.5-VL-32B**
VSI-Bench	43.5	46.3	49.0
SpaceVista-Bench	29.5	34.5	36.3

Table D25: The release time and model source of LLMs used

Model	**Release Time**	**Source**
GPT-5(OpenAI, 2025)	2025-08	`https://openai.com/gpt-5/`
GPT-4o(Hurst et al., 2024)	2024-05	`https://gpt4o.ai/`
Claude-Opus-4.1(Anthropic, 2025c)	2025-08	`https://www.anthropic.com/news/claude-opus-4-1`
Claude-Sonnet-4(Anthropic, 2025b)	2025-05	`https://www.anthropic.com/claude/sonnet`
Gemini-2.5-Pro(DeepMind, 2025)	2025-06	`https://deepmind.google/technologies/gemini/pro/`
Gemini-2.5-Flash(DeepMind, 2025)	2025-06	`https://deepmind.google/models/gemini/flash/`
Internvl3.5-38B (Wang et al., 2025c)	2025-08	`https://huggingface.co/OpenGVLab/InternVL3_5-38B-Instruct`
Internvl3.5-14B (Wang et al., 2025c)	2025-08	`https://huggingface.co/OpenGVLab/InternVL3_5-14B-Instruct`
Internvl3-78B (Zhu et al., 2025)	2025-04	`https://huggingface.co/OpenGVLab/InternVL3-78B`
Internvl3-38B (Zhu et al., 2025)	2025-04	`https://huggingface.co/OpenGVLab/InternVL3-38B`
GLM-4.5V (Team et al., 2025)	2025-08	`https://www.glm45.com/glm45v`
GLM-4.1V-Thinking (GLM et al., 2024)	2025-07	`https://huggingface.co/zai-org/GLM-4.1V-9B-Thinking`
Qwen2.5VL-72B (Bai et al., 2025)	2025-01	`https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct`
Qwen2.5VL-32B (Bai et al., 2025)	2025-01	`https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct`
LLAVA-Onevision-72B (Li et al., 2024a)	2024-08	`https://huggingface.co/llava-hf/llava-onevision-qwen2-72b-ov-hf`
LLAVA-Onevision-7B (Li et al., 2024a)	2024-08	`https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov`

As summarized in Tab. D24, the dataset affords a certain degree of support for the 32B model's capabilities. Nevertheless, beyond this observation, the main results are achieved by the 7B configuration, whereas ablation studies are primarily conducted with the 3B model.

D.8 LEADERBOARD DETAIL

To assess the spatial reasoning ability of both closed-source and open-source models, we evaluate the latest available versions. Tab. 5 presents their performance across the Tiny Tabletop, Tabletop, Indoor, and Outdoor scenarios, whereas Tab. D25 provides an overview of their release dates and sources. For closed-source models accessed via API and open-source models, the generation configurations are summarized in Tab. D26 and D27, respectively.

E FAQ

E.1 ERROR ACCUMULATION

Our data construction pipeline is primarily based on metric depth estimation and the corresponding transformation to canonical view space. It should be noted that this approach may introduce potential error accumulation, especially considering that current metric depth estimation models have not yet achieved high performance at full scale.

To address concerns regarding error accumulation, we justify our methodology from the following perspectives: **1) data quality assurance:** To ensure alignment with human perception, we implement a multi-tiered validation process. Specifically, we conduct manual verification on a subset of the training set, perform full human annotation on the entire test set, and additionally collect real-world measured data to construct a dedicated test subset. These measures effectively ensure that the automatically generated data remains suitable for learning human perceptual models. We argue that even if minor error accumulation exists, it does not compromise the overall quality and contribution of the dataset. **2) forward-looking methodological contribution:** The proposed data construction framework and model architecture will have a significant impact on the field of all-scale spatial reasoning. Importantly, as more accurate all-scale inference methods emerge in the future, we will continuously integrate higher-quality data to refine this work. This dynamic updating mechanism ensures the long-term relevance and value of our research.

Table D26: Generating parameters for Closed-Source LLMs.

Model	Generation Setup
GPT-5	"model" : "gpt-5", "temperature" : 0, "max_tokens" : 1024
GPT-4o	"model" : "gpt-4o", "temperature" : 0, "max_tokens" : 1024
Claude-Opus-4.1	"model" : "claude-opus-4.1", "temperature" : 0, "max_tokens" : 1024
Claude-Sonnet-4	"model" : "claude-sonnet-4", "temperature" : 0, "max_tokens" : 1024
Gemini-2.5-Pro	"model" : "gemini-2.5-pro", "temperature" : 0, "max_tokens" : 1024
Gemini-2.5-Flash	"model" : "gemini-2.5-flash", "temperature" : 0, "max_tokens" : 1024

E.2 ALL SCALE POSSIBILITIES

Currently, our data coverage remains limited in addressing the full spectrum of spatial scales, despite the equal importance of spatial understanding across these domains. At fine scales, domains such as minimally invasive surgery call for millimeter-level models, while precision manufacturing—especially semiconductor production—pushes into the nanometer range. These capabilities underpin progress in healthcare and technology. In contrast, large-scale applications, including satellite remote sensing and cartography, typically work with resolutions of 10 kilometers or greater.

While spatial understanding is equally essential across these extremes, the imaging and 3D modeling techniques involved extend well beyond conventional real-world sensing methods. As a result, our current work does not fully address these diverse scales. Nevertheless, we aim to expand our capabilities in the future by integrating modeling across a broader range of dimensions, thereby bridging these gaps and enabling more unified spatial analysis.

E.3 DISCUSSION OF DATASET

We use the free-form subset of SPAR-7M(Zhang et al., 2025e), which consists of approximately 100K samples, about 1% of the original dataset. This part of the data is later processed and filtered with original Scannet (Dai et al., 2017), Scannet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021) to fit the requirements of our dataset. However, we do not consider our model to be trained on SPAR-7M, nor do we compare it against models trained on SPAR-7M in SparBench. We observe that SPAR-7M's data design leads to over 200 QA pairs per scene on average, which can cause overfitting in indoor scenarios. Instead, we leverage SPAR-7M's scan-based characteristics to construct our own CoT for cold-start purposes. It is important to note that neither SpaceR nor SPAR-7M includes CoT reasoning. We generate CoT following the method described in Sec. 3 and apply filtering and screening to ensure quality. These processed data sources, along with the wild video dataset, are integrated into SpaceVista-1M, while acknowledging the additional labeling and

Table D27: Generating parameters for Open-Source LLMs.

Model	Generation Setup
Internvl3.5-38B	do_sample = False, temperature = 0, max_new_tokens = 512
Internvl3.5-14B	do_sample = False, temperature = 0, max_new_tokens = 512
Internvl3-38B	do_sample = False, temperature = 0, max_new_tokens = 512
Internvl3-78B	do_sample = False, temperature = 0, max_new_tokens = 512
GLM-4.5V	do_sample = False, temperature = 0, max_new_tokens = 1024
GLM-4.1V-Thinking	do_sample = False, temperature = 0, max_new_tokens = 1024
Qwen2.5VL-32B	do_sample = False, max_new_tokens = 1024
Qwen2.5VL-72B	do_sample = False, max_new_tokens = 1024
LLAVA-Onevision-7B	do_sample = False, temperature = 0, max_new_tokens = 1024
LLAVA-Onevision-72B	do_sample = False, temperature = 0, max_new_tokens = 1024

filtering steps involved in our pipeline. Overall, these decisions support our position that our data retains a meaningful degree of independence from SPAR-7M and SpaceR.

F PREVIEW

F.1 SCENE PREVIEW

Indoor Scenes. Our indoor dataset consists of simple and clean room-scale environments such as living rooms, meeting rooms, and classrooms. An overview of the data is provided in Fig. F26, highlighting the simplicity and cleanliness of our indoor scenes compared to more complex wild indoor environments. Living rooms feature sofas, coffee tables, and shelves arranged along walls with open floor space. Meeting rooms include evenly spaced chairs around a central table, while classrooms have rows of desks facing a blackboard or screen. These scenes show limited object variety and limited scene complexity.

Wild Indoor Scenes. Representative wild indoor scenes, captured via multi-view smartphone recordings in complex and unconstrained environments such as shopping malls, banquet halls, and art galleries, are illustrated in Fig. F27. These scenes exhibit diverse architectural layouts and high object density. Like in shopping malls, elements such as escalators, display shelves, and glass facades create multi-layered structures with frequent reflections and occlusions. Compared to previous indoor scenes, wild indoor scenes have irregular layouts, dense furniture, diverse objects, and uneven lighting, leading to more complex spatial arrangements. This contrast underscores the structured and clear nature of our data, which supports controlled spatial reasoning evaluation.

Outdoor Scenes. Our outdoor scenes include various environments such as parks, tourist landmarks, and others, captured from both ground and aerial views, as shown in Fig. F28. Parks contain irregularly shaped walking paths winding through dense clusters of trees, shrubs, and open lawns, creating a mix of natural textures and spatial variations. These areas often include water features, benches, and varied terrain elevations. Therefore, outdoor scene layouts usually involve plazas, staircases, and structured open spaces that introduce rich geometric complexity.

Drone Scenes. Fig. F29 shows examples from a drone's perspective. Aerial, low-angle, and oblique views offer detailed spatial structures that are not easily visible from the ground. Playgrounds exhibit clear arrangements of play equipment and open spaces, while parking lots display orderly rows of vehicles and marked boundaries. Parks show clusters of trees, pathways, and water bodies, revealing a layered combination of natural and built elements. These diverse viewpoints provide a more complete understanding of scene layout and environmental features, supporting improved spatial reasoning.

Tabletop Scenes. Examples of tabletop scenes are illustrated in Fig. F30. These scenes capture everyday objects such as keyboards, boxes, and fruits arranged on tabletops, characterized by natural occlusions, varying object placements, and diverse background textures. The dataset employs dynamic multi-view acquisition using mobile devices, enabling richer structural coverage compared to traditional static indoor datasets. This approach captures subtle interactions between objects and background elements, as well as changes in viewpoint and lighting conditions.

Tiny Tabletop Scenes. The Fig. F31 shows the tiny tabletop scenes from our dataset. These data are 360-degree turntable videos to capture objects from every angle, solving occlusion issues and improving scene completeness.

Our Collected Scenes. We use mobile devices to capture and collect data for some Tabletop and Tiny Tabletop scenes. Our collected data, shown in Fig. F32, features diverse objects and detailed multi-view coverage, enabling fine-grained spatial analysis. The data is similar to the previously mentioned tabletop and tiny tabletop. Tabletop scenes have relatively large objects and rich and diverse backgrounds, which are suitable for capturing diverse objects and natural environments in daily life; while Tiny Tabletop scenes focus on smaller objects, emphasizing detail integrity and multi-view coverage, which facilitates in-depth research on the subtle structure and morphology of these scenes.

F.2 TEMPLATE PREVIEW

As shown in Tab. F28, we present three exemplar applications: point input for Object Counting, bounding box input for Object Distance, and original input for Spatial Relation. Other scenes and tasks are similar to the example template.

F.3 QA PREVIEW

We provide a comprehensive set of SpaceVista-1M QA pairs here for preview in Tab.F29-Tab.F47. Note that the RL-oriented multiple-choice and regression formats omit anchors like `<semantic>` and `<scale>`, since they can be easily injected during training from the meta information. Since if objects are referred to by a bounding box, the only changes needed are to change the object name into the corresponding object point/bbox/mask. Each question takes only one video with one form of referring. For example, *"Where is the toothbrush relative to the keyboard from the view of the start frame?"* $\rightarrow$ *"Where is the red mask referred object relative to the keyboard from the view of the start frame?"*. So, in this preview, we only provide the natural language questions for clarity.

Overall, these previews highlight the diversity of our all-scale reasoning SpaceVista-1M dataset.

Table F28: Multi-type template preview. Examples using the point input for Object Counting, the bounding-box input for Object Distance, and the original input for Spatial Relation.

Point Input Template

- Refer to the red point in the starting frame and count how many objects are of that type.
- Count the number of objects whose class is referred to by the red point in the first frame throughout the video.
- Using the red point in the first frame as reference, count how many objects of that class appear in the entire video.
- Count every object like the one highlighted by the red point in the video’s first frame.
- Find all video objects that are of the same kind as the one identified by the red point.
- Identify the class from the red point in frame one and tally all instances of that class in the video.
- How many objects in the video resemble the one tagged with the red point in the first frame?
- Search for all items that belong to the same class as the one shown by the red point in frame one.
- Track all objects of the same category as the red-point one from the first frame and count them.
- Count the total number of objects in the video that correspond to the class defined by the red point in the first frame.
- Use the red point to find a class and count how many such instances are there in the video.
- Using the initial frame’s red point as a guide, total up all objects of that class.
- From the first frame’s red point, find that class and count its appearances across the video.
- Match the object under the red point to others in the video and count them.
- Take the red-pointed object as example and count all others like it in the video.

Bounding Box Input Template

- How far apart do the objects enclosed by the red bounding box and blue bounding box appear in these frames?
- What space lies between the red bounding box and the blue bounding box in these frames?
- What is the distance measurement between the red bounding box and blue bounding box in the video?
- What is the distance between the red-bounded object and the blue-bounded object in the video?
- Measure the distance separating the red bounding box and blue bounding box in the video frames.
- What is the estimated distance between the red bounding box and the blue bounding box in the video?
- What is the measured distance between the red bounding box and blue bounding box in the footage?
- Calculate the ground distance from the red bounding box to the blue bounding box based on the frames.
- Find the ground distance between the red bounding box and the blue bounding box in these images.
- How wide is the space between the red bounding box and the blue bounding box in the video?
- Based on the frames, what is the distance from the red bounding box to the blue bounding box?
- Please estimate the ground distance between the red bounding box and the blue bounding box in these images.
- What is the approximate distance between the red bounding box and blue bounding box in these images?
- Provide an estimate for the distance between the red bounding box and blue bounding box seen in the footage.
- How far is the red bounding box from the blue bounding box in the frames?

Original Input Template

- Describe how desk and chair are spatially positioned relative to each other.
- What is the spatial relation type between desk and chair in the video?
- What type of spatial relationship exists between desk and chair in these frames?
- Estimate the spatial relation (such as support, stacking, adhesion, hanging, plug-in) between desk and chair in these frames.
- What is the most likely spatial relationship (support, stacking, adhesion, hanging, plug-in) between cabinet and book?
- Can you describe the spatial relationship type of awning and awning?
- Identify how picture and ceiling are spatially related in the video sequence.
- Between desk and chair, what spatial link exists?
- What spatial relation links tag to hat in the given frames?
- What spatial relation best fits cable and computer mouse in the video frames?
- Identify how cable and socket are spatially related in the video sequence.
- Describe the spatial relation (e.g., support, stacking, adhesion, hanging, plug-in) between fork and spoon.
- Explain the spatial relation between toy camera and building blocks in the video.
- How would you classify the spatial relation between sticky note and tumbler?
- What type of spatial relationship exists between toy block and toy train in these frames?.

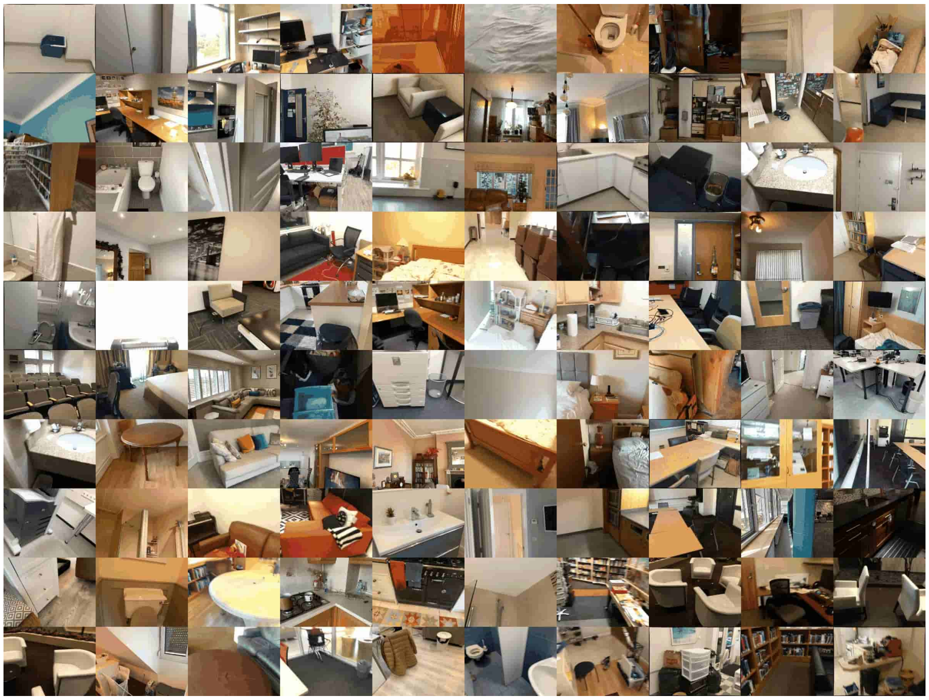


Figure F26: Indoor data are rather simple and clean scenes inside a room. The overall scene is not as complex as the wild indoor scene.

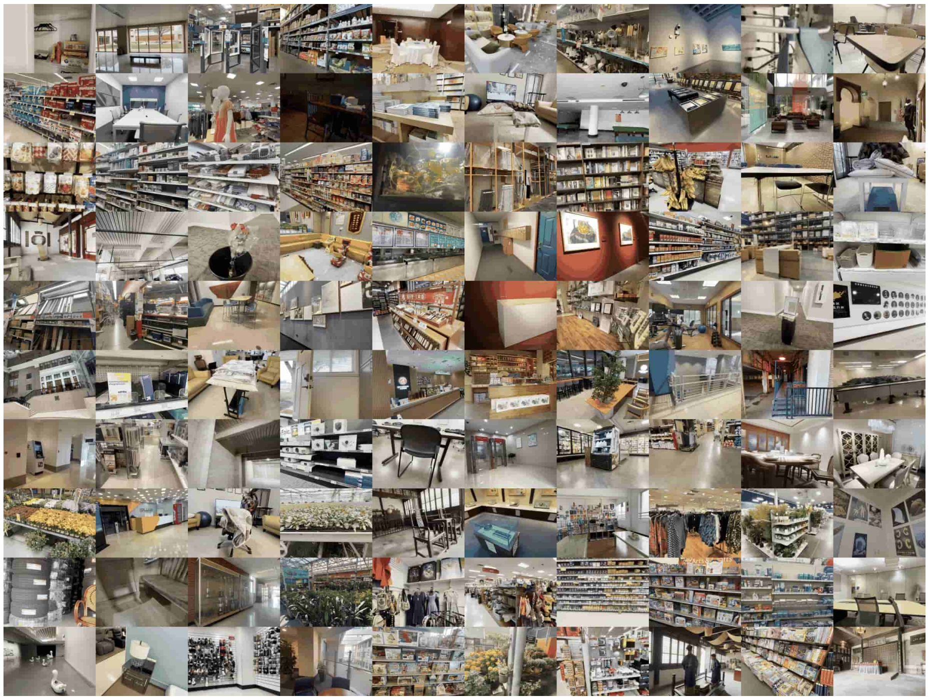


Figure F27: Wild indoor data includes more light changes, reflections, and transparency. The objects included are more diverse.

Figure F28: Outdoor data is jointly collected from ground views, incorporating street, park, building and so on.

Figure F29: Drone data captures ground objects from above at oblique angles, providing more complete structural coverage than traditional ground-based capture methods.

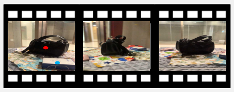
Figure F30: In this tabletop scene, videos capture tabletop objects exhibiting rich background variation and natural occlusions, delivering clearer structural coverage of the objects than traditional static indoor datasets.

Figure F31: Tiny tabletop objects captured with rich details for small objects, focusing on fine-scale scenes, unlike typical large or complex indoor or outdoor datasets.

Figure F32: These samples are collected by us. As small-scale, Tabletop, and Tiny Tabletop datasets offer rich details with accurate annotation.

Table F29: The spatial relation task QA preview.

Spatial Relation Task



Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
What is the most likely spatial relationship between the red point and the blue point?
Options: A.Stacking B. Dhesion C. Support D. Adjacent E. Plug-in

Bounding Box Input

```
[{"x1":91, "y1":181,
"x2":439, "y2":458},
{"x1":241, "y1":471,
"x2":479, "y2":627}]
```

Point Input

```
[[328, 287], [548,
360]]
```

Mask Input

SFT GT
<think>Based on the spatial relationship in the image, the red point appears to be positioned below or supporting the blue point. Given their relative arrangement and visual positioning, the red point is most likely the foundation or support for the blue point.</think> <answer>The red point likely serves as the foundation or support for the blue point.</answer>
RL GT
<think>*</think><answer>D</answer>

Table F30: The camera moving task QA preview.

Camera Moving Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
What directional path did the camera follow in the video, ignoring rotation?
Options: A.Front B. Back C. Left D. Right E: Down

Bounding Box Input	**Point Input**	**Mask Input**
Not applicable.	Not applicable.	Not applicable.

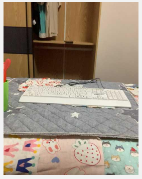
SFT GT
<think>The images show a sequence of shots from the same scene, with each frame zooming in slightly more. This indicates that the camera is moving forward towards the structure, capturing more detailed parts of the scene as it progresses. The camera keeps moving forward.</think> <answer>Front</answer>
RL GT
<think>*</think><answer>A</answer>

Table F31: The position comparison task QA preview.

Position Comparison Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
Where is the toothbrush relative to the keyboard from the view of the start frame?
Options: A.Left B. Right C. Front D. Back

Bounding Box Input	**Point Input**	**Mask Input**
Not applicable.	`[237, 9]`	

SFT GT
<think>The video starts with the toothbrush on the left side in the green cup. As the view rotates, its position shifts from left to right and then back to left. Before disappearing, it remains on the left side. So the toothbrush is on the left from the view of the start frame.</think> <answer>Left</answer>
RL GT
<think>*</think><answer>A</answer>

Table F32: The size comparison task QA preview.

Size Comparison Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
Based on the height of the referred object shown in the video, which is taller: the red point object or the blue point object?
Options: A.Streetlight B. Sign

Bounding Box Input	**Point Input**	**Mask Input**
`[{"x1":17, "y1":110, "x2":94, "y2":248}, {"x1":126, "y1":74, "x2":163, "y2":220}]`	`[[185, 52], [101, 144]]`	

SFT GT
<think>By analyzing the position of the light source in the image, it's clear that the streetlight is situated above the sign. Additionally, due to its height and the relatively distant position, the streetlight has a broader light coverage, further suggesting it is taller than the sign.</think> <answer>Streetlight.</answer>
RL GT
<think>*</think><answer>A</answer>

Table F33: The existence estimation task QA preview.

Existence Estimation Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
Determine which objects in the cabinet, printer, dishwasher, and trash bin are found in the footage.
Options: A.Printer B. Cabinet C. Dishwasher D. Trash bin

Bounding Box Input	**Point Input**	**Mask Input**
Not applicable.	Not applicable.	Not applicable.

SFT GT
<think>Based on the content in the images, I can see a clearly visible trash bin placed on the ground. Its presence is obvious in the scene, as it is clearly shown, and other objects in the image provide a reference for its position and shape. Therefore, it is reasonable to infer the existence of the trash bin.</think> <answer>Trash bin</answer>
RL GT
<think>*</think><answer>D</answer>

Table F34: The rotation estimation task QA preview.

Rotation Estimation Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.

Text Input
Calculate the rotation angle of the object in the red bounding box between the first and last frame.

Bounding Box Input	**Point Input**	**Mask Input**
`[{"x1":911, "y1":605, "x2":1201, "y2":853}]`	`[[480, 989]]`	

SFT GT
<think>By observing the change in the object's position between the frames, I estimate its rotation angle. Based on the relative rotation in the two images and using reference points, I approximate the object has rotated by about 136 degrees. The exact angle might vary slightly, but it's within this range.</think> <answer>136 degrees</answer>
RL GT
<think>*</think><answer>136</answer>

Table F35: The relative distance task QA preview.

Relative Distance Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.

Text Input
How close does the red mask get to the blue mask in the frames (in centimeters)?

Bounding Box Input	**Point Input**	**Mask Input**
`[{"x1":194, "y1":172, "x2":299, "y2":511}, {"x1":393, "y1":377, "x2":479, "y2":483}]`	`[[351, 245], [428, 442]]`	

SFT GT
<think>Based on the content of frames, the relative position of the red and blue objects changes minimally. The bottle seems 20 centimeters tall. Therefore, the relative distance between them in the frames is estimated to be approximately 7.8 centimeters. The contrast between their shapes and background positions helps in making this estimation.</think> <answer>8 centimeters</answer>
RL GT
<think>*</think><answer>7.8</answer>

Table F36: The absolute distance task QA preview.

Absolute Distance Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as ’let me think’, ’wait’, ’hmm’, ’oh, I see’, ’let’s break it down’, etc, or other natural language thought expressions. It’s encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.

Text Input
What is the estimated distance between the red mask and the blue mask in the video (in meters)?

Bounding Box Input	**Point Input**	**Mask Input**
`[{"x1":250, "y1":228, "x2":307, "y2":314}, {"x1":402, "y1":229, "x2":479, "y2":300}]`	`[[270, 278], [264, 442]]`	

SFT GT
<think>The red and blue objects are two regular boxes of around 0.1m in size. Based on the position of the red and blue objects in the image, the estimated absolute distance between them is approximately 0.31 meters. This estimation is made based on the size and arrangement of the objects in the frame.</think> <answer>0.31 meters.</answer>
RL GT
<think>*</think><answer>0.31</answer>

Table F37: The room size task QA preview.

Room Size Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as ’let me think’, ’wait’, ’hmm’, ’oh, I see’, ’let’s break it down’, etc, or other natural language thought expressions. It’s encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.

Text Input
Give an estimate of the room size based on the video (in square meters).

Bounding Box Input	**Point Input**	**Mask Input**
Not applicable.	Not applicable.	Not applicable.

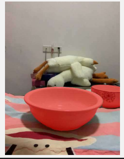
SFT GT
<think>By observing the spatial layout in the image, I estimate the room’s size. Based on the relative proportions and perspective of the table and sofa within the room, and considering typical room size proportions with normal table and sofa size, the area of the room is approximately 24.33 square meters. This estimation is derived from the visible space in the image.</think> <answer>24.33 square meters</answer>
RL GT
<think>*</think><answer>24.33</answer>

Table F38: The object counting task QA preview.

Object Counting Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as ’let me think’, ’wait’, ’hmm’, ’oh, I see’, ’let’s break it down’, etc, or other natural language thought expressions. It’s encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.

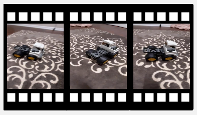
Text Input
How many bowls can be seen throughout the video?

Bounding Box Input	**Point Input**	**Mask Input**
`[{"x1":81, "y1":357, "x2":408, "y2":539}]`	`[[439, 243]]`	

SFT GT
<think>Based on the objects in the image, two bowls are clearly visible in different positions. Their number is easily identifiable in the frame. Therefore, it can be estimated that there are 2 bowls visible in the video.</think> <answer>2 bowls</answer>
RL GT
<think>*</think><answer>2</answer>

Table F39: The object size task QA preview.

Object Size Task



Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as ’let me think’, ’wait’, ’hmm’, ’oh, I see’, ’let’s break it down’, etc, or other natural language thought expressions. It’s encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.

Text Input
How wide is the toy in the first frame (in centimeters)?

Bounding Box Input	**Point Input**	**Mask Input**
`[{"x1":112, "y1":547, "x2":717, "y2":1089}]`	`[[820, 437]]`	

SFT GT
<think>By observing the size of the toy in the image and its relative proportion to surrounding objects, the width of the toy can be estimated. Based on the proportional reference in the frame, the toy’s width is approximately 7.4 centimeters.</think> <answer>7.4 centimeters</answer>
RL GT
<think>*</think><answer>7.4</answer>

Table F40: The route plan task QA preview for evaluation.

Route Plan/Navigation Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
What actions should you take to navigate from the toilet, facing the bathtub, to the door, given the following steps: 1. [please fill in] 2. Go forward until the door. You have reached the final destination?
A.Turn left B.Turn right C.Turn back

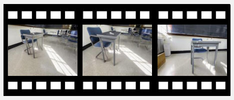
Bounding Box Input	**Point Input**	**Mask Input**
Not applicable.	Not applicable.	Not applicable.

SFT GT
<think>Based on the layout in the image, assuming you start at the toilet, facing the bathtub, the first step is to turn right, then continue forward until you reach the door. This path is aligned with the relative positioning of the objects in the room.</think> <answer>Turn right</answer>

RL GT
<think>*</think><answer>B</answer>

Table F41: The appearance order task QA preview.

Appearance Order Task



Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
Please provide the sequence in which these objects appear in the video: table, chair, blackboard. Options: A.["table", "blackboard", "chair"] B. ["chair", "table", "blackboard"] C. ["blackboard", "chair", "table"] D. ["table", "chair", "blackboard"]

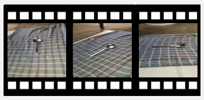
Bounding Box Input	**Point Input**	**Mask Input**
Not applicable.	Not applicable.	Not applicable.

SFT GT
<think>Based on the objects' sequence in the image, the table appears first, followed by the chair, and finally the blackboard. This order is determined by the relative positions of the objects in the frame.</think> <answer>Table, chair, blackboard.</answer>

RL GT
<think>*</think><answer>D</answer>

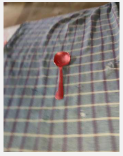
Table F42: The depth estimation task QA preview.

Depth Estimation Task



Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.

Text Input
Please provide the depth value of the object labeled by the red mask in the initial frame (in meters).

Bounding Box Input	**Point Input**	**Mask Input**
`[{"x1":466, "y1":528, "x2":620, "y2":1043}]`	`[[733, 535]]`	

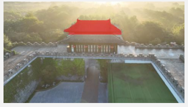
SFT GT
<think>Based on the position of the object labeled by the red mask in the image and considering the perspective and size of the object, the estimated depth of the object is 0.72 meters with the reference spoon's length of 0.17 meters.</think> <answer>0.72</answer>
RL GT
<think>*</think><answer>0.72</answer>

Table F43: The view change inference task QA preview.

Area Estimation Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide the numerical value (e.g., 42 or 3.14) within the <answer> </answer> tags.

Text Input
What is the area of the roof shown in the video (in square meters)?

Bounding Box Input	**Point Input**	**Mask Input**
`[{"x1":161, "y1":48, "x2":312, "y2":85}]`	Not applicable.	

SFT GT
<think>Based on the perspective in the image and the relative size of the objects, the area of the roof can be estimated. By considering the proportions of the objects and the background perspective, the area of the roof is approximately 146.0 square meters.</think> <answer>The area of the roof in the video is 146.0 square meters.</answer>
RL GT
<think>*</think><answer>146.0</answer>

Table F44: The object matching task QA preview.

Object Matching Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as ’let me think’, ’wait’, ’hmm’, ’oh, I see’, ’let’s break it down’, etc, or other natural language thought expressions. It’s encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
Which bounding box (green, blue, or yellow) in the last frame is the same object as the red bounding box in the first frame?
Options: A.Yellow B. Green

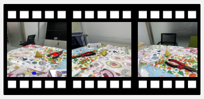
Bounding Box Input	**Point Input**	**Mask Input**

`[{"x1":87, "y1":382, "x2":398, "y2":493}]`
`[{"x1":122, "y1":515, "x2":237, "y2":617}, {"x1":62, "y1":376, "x2":381, "y2":534}]`
	Not applicable.	Not applicable.

SFT GT
<think>By analyzing the object positions and color contrasts in the image, the green bounding box matches the object of the red bounding box in the first frame. This conclusion is based on the relative positions and sizes of the objects in both frames.</think> <answer>The green box.</answer>
RL GT
<think>*</think><answer>B</answer>

Table F45: The obstacles location task QA preview.

Obstacles Location Task



Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as ’let me think’, ’wait’, ’hmm’, ’oh, I see’, ’let’s break it down’, etc, or other natural language thought expressions. It’s encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

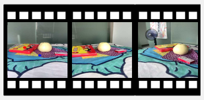
Text Input
Which object will cause obstacles when moving the pliers to the left of the box? (Red point denotes the pliers, Blue point denotes the target position.)
Options: A.Pineapple B. Charging plug C. None

Bounding Box Input	**Point Input**	**Mask Input**
Not applicable.	Not applicable.	Not applicable.

SFT GT
<think>The red-handled pliers are on the right, and the target is on the left side of the box at the lower right. If the pliers are moved to the box’s left side, the path crosses the pineapple in the center of the table, making it the main obstacle.</think> <answer>Pineapple.</answer>
RL GT
<think>*</think><answer>A</answer>

Table F46: The manipulation planning task QA preview.

Manipulation Planning Task



Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
How to move the box to the front of the apple from the viewpoint of the first frame? (Red point denotes the box, blue point denotes the target position.)
Options: A.Moving backward 43.6cm B. Moving left 10.2cm C. Moving up 45.7cm D. Moving backward 28.1cm

Bounding Box Input	**Point Input**	**Mask Input**
Not applicable.	Not applicable.	Not applicable.

SFT GT
<think>The red point denotes the current position of the box, and the blue point denotes the target. The task is to move the box in front of the apple by shifting it along the red-to-blue direction about 28.1cm.</think> <answer>Move the book backwards 28.1cm to put the box in front of the apple.</answer>
RL GT
<think>*</think><answer>D</answer>

Table F47: The area estimation task QA preview.

View Change Inference Task

Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection.
Video: `<video>`
Question: `<text>`
During RL: Please provide the thinking process within the <think> </think> tags. Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags.

Text Input
What is the view change between the input frames?
Options: A.Back B. Down C. Right D. Left E. Front

Bounding Box Input	**Point Input**	**Mask Input**
Not applicable.	Not applicable.	Not applicable.

SFT GT
<think>By analyzing the angle change between the frames, it's clear that the view shifts downward. This conclusion is drawn from comparing the position and angle of objects in the beginning frames.</think> <answer>Downward</answer>
RL GT
<think>*</think><answer>B</answer>