Spatial Understanding from Videos: Structured Prompts Meet Simulation Data

Spotlight Paper

Haoyu Zhang[1,2], Meng Liu[3,4†], Zaijing Li[1,2], Haokun Wen[1], Weili Guan[1], Yaowei Wang[1,2], Liqiang Nie[1†]

[1]Harbin Institute of Technology (Shenzhen) [2]Peng Cheng Laboratory
[3]Shandong Jianzhu University [4]Zhongguancun Academy

➢ Motivation

- **Spatial Uncertainty.** In the absence of explicit depth information, models must infer 3D structure from inherently limited 2D observations. This process is further complicated by occlusions, perspective distortions, and texture ambiguities, all of which introduce significant spatial uncertainty

➢ Motivation

- **Spatial Uncertainty.** In the absence of explicit depth information, models must infer 3D structure from inherently limited 2D observations. This process is further complicated by occlusions, perspective distortions, and texture ambiguities, all of which introduce significant spatial uncertainty

↓

SpatialMind Prompting Strategy
for multi-step logical reasoning

➢ SpatialMind Prompting Strategy

① Scene Decomposition

$\overbrace{\text{Local Modeling} \qquad \text{Coordinate Mapping} \qquad \text{Cognition Generation}}$

② Question Decomposition

➢ SpatialMind Prompting Strategy

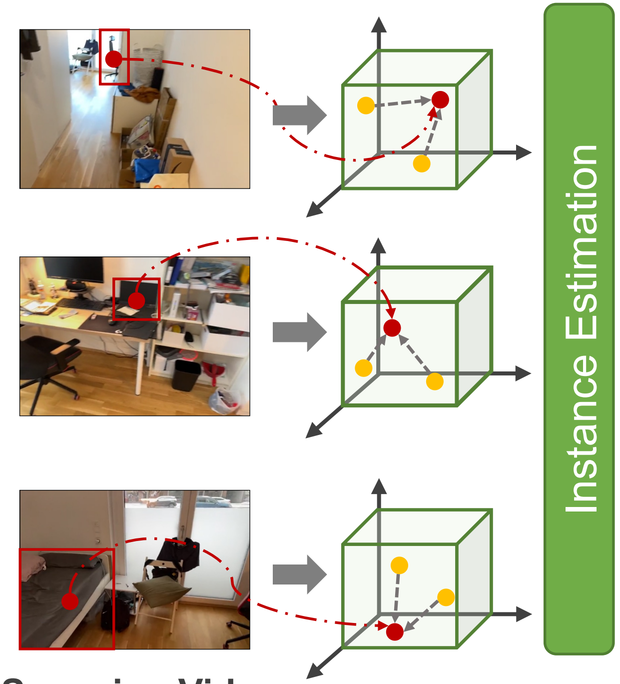
① Scene Decomposition

② Question Decomposition

Local Modeling

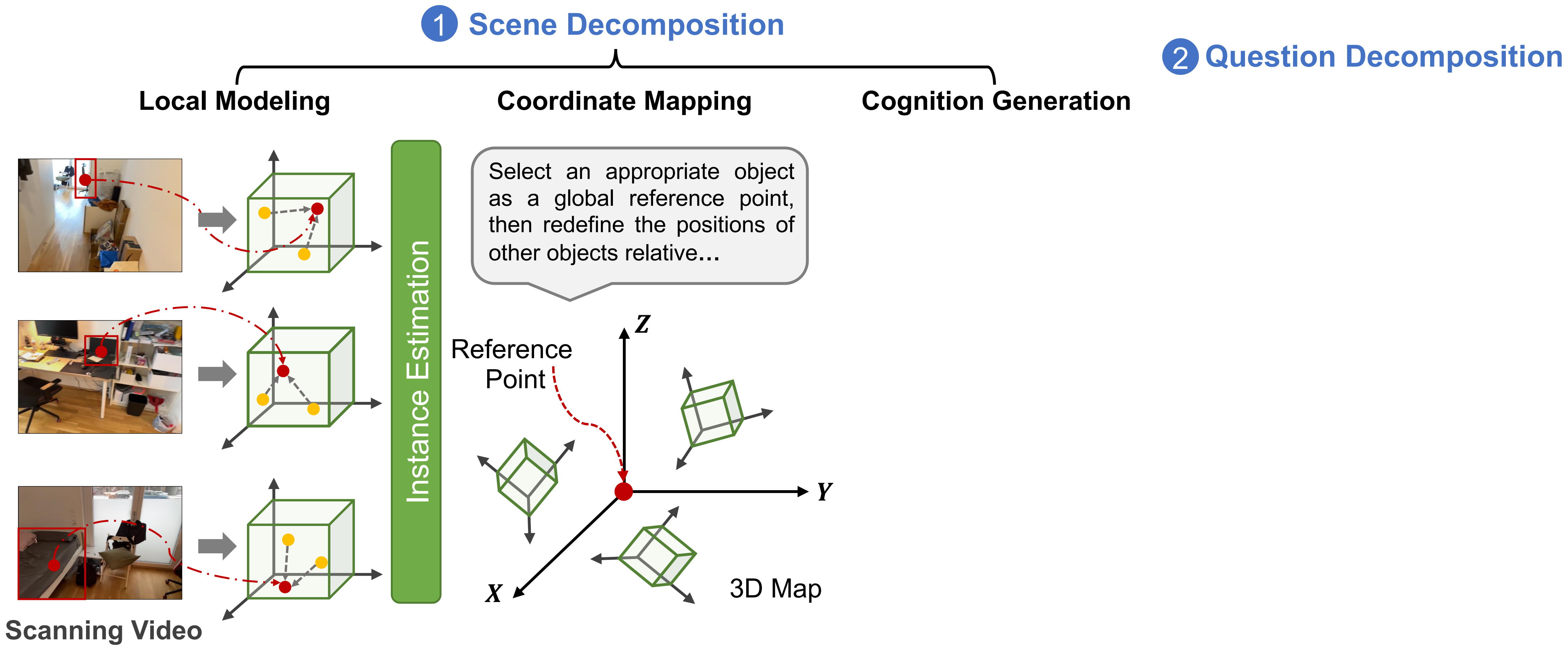
Coordinate Mapping

Cognition Generation

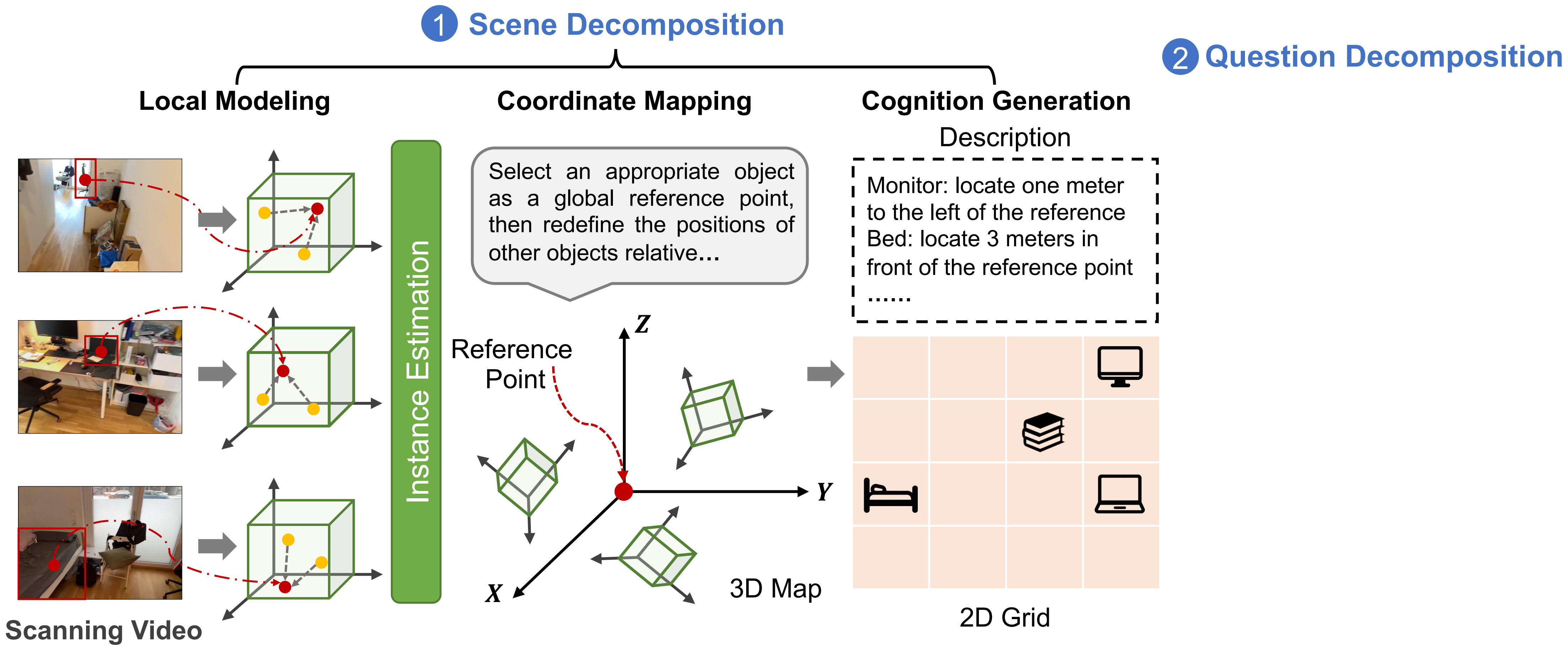


Scanning Video

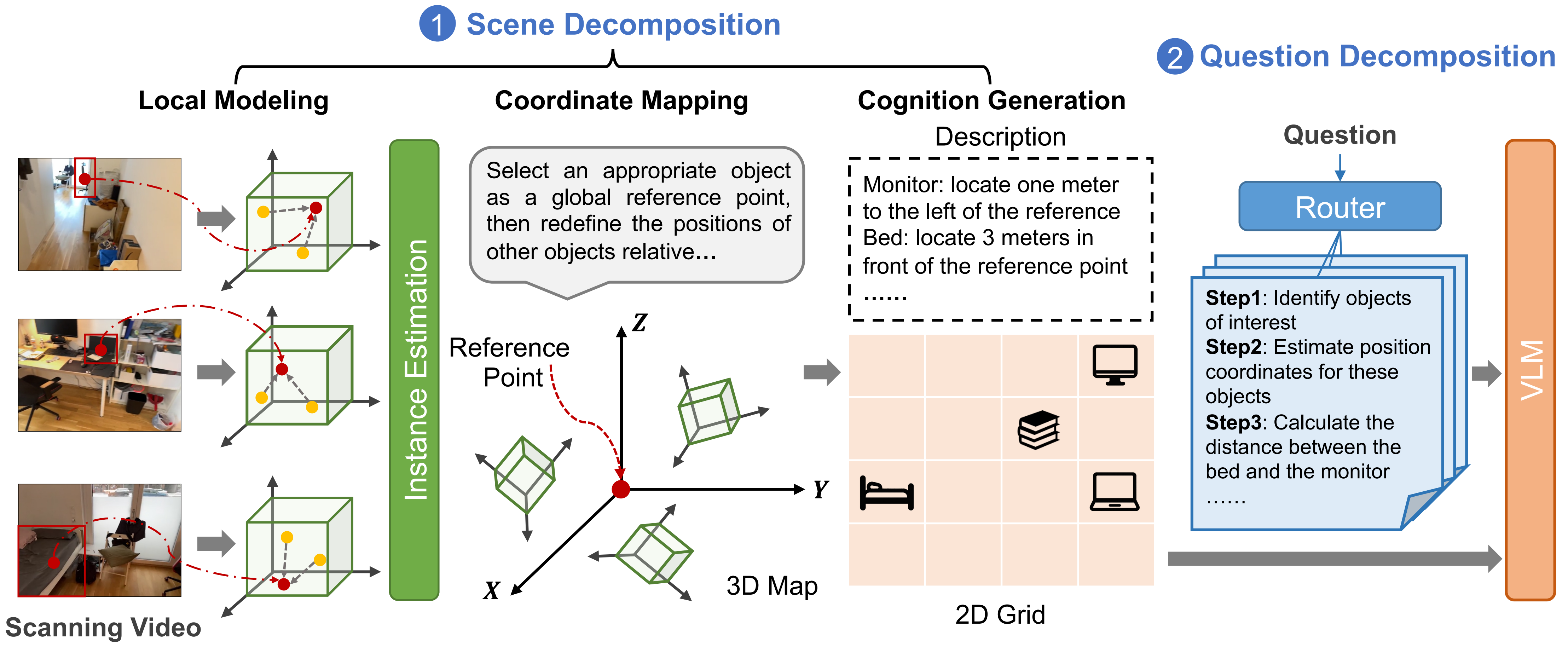
➢ SpatialMind Prompting Strategy



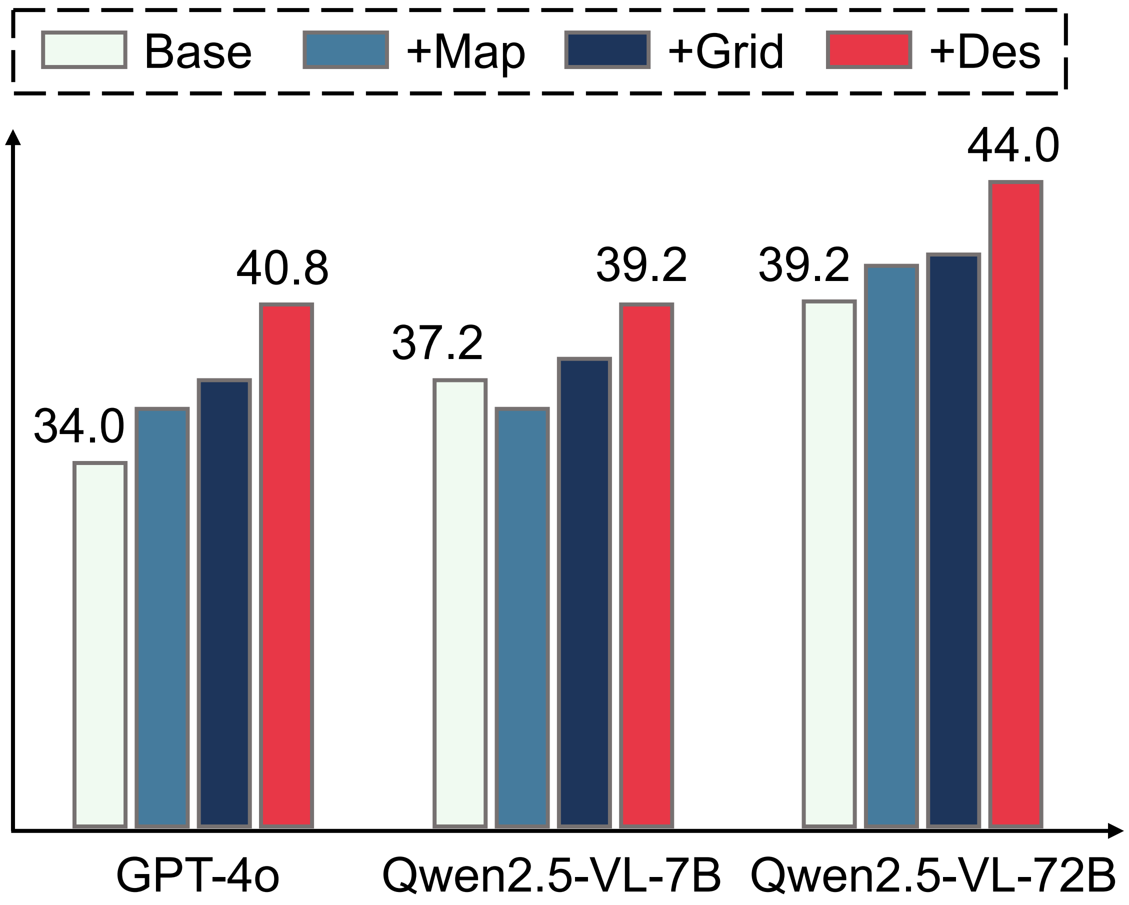
➢ SpatialMind Prompting Strategy



➢ SpatialMind Prompting Strategy



➢ SpatialMind Prompting Strategy



➢ Motivation

- **Spatial Uncertainty.** In the absence of explicit depth information, models must infer 3D structure from inherently limited 2D observations. This process is further complicated by occlusions, perspective distortions, and texture ambiguities, all of which introduce significant spatial uncertainty

- **Data Scarcity.** Existing datasets for this task are limited in both scale and diversity, restricting the ability of VLMs to acquire robust spatial knowledge and perceptual capabilities. Moreover, these datasets involve scans of real-world scenes, which leads to poor scalability.

➢ Motivation

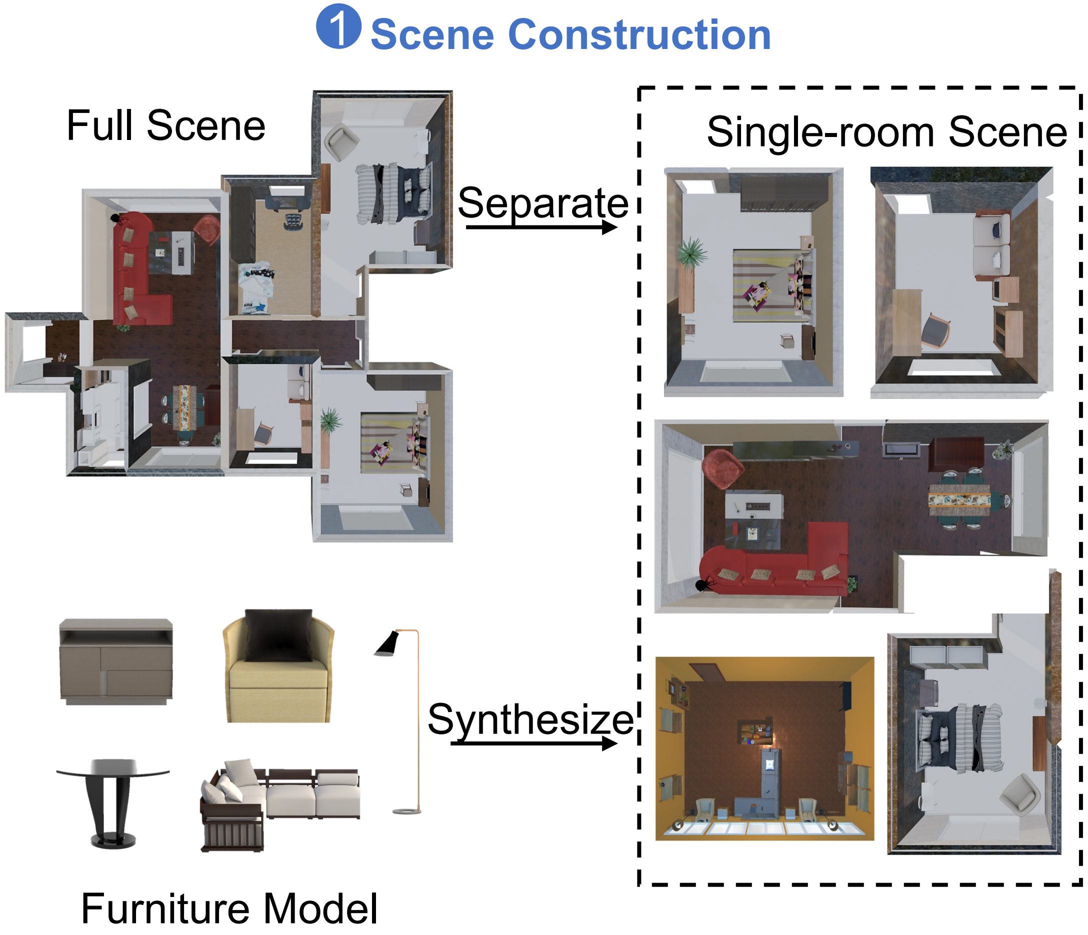
- **Spatial Uncertainty.** In the absence of explicit depth information, models must infer 3D structure from inherently limited 2D observations. This process is further complicated by occlusions, perspective distortions, and texture ambiguities, all of which introduce significant spatial uncertainty
- **Data Scarcity.** Existing datasets for this task are limited in both scale and diversity, restricting the ability of VLMs to acquire robust spatial knowledge and perceptual capabilities. Moreover, these datasets involve scans of real-world scenes, which leads to poor scalability.

⬇

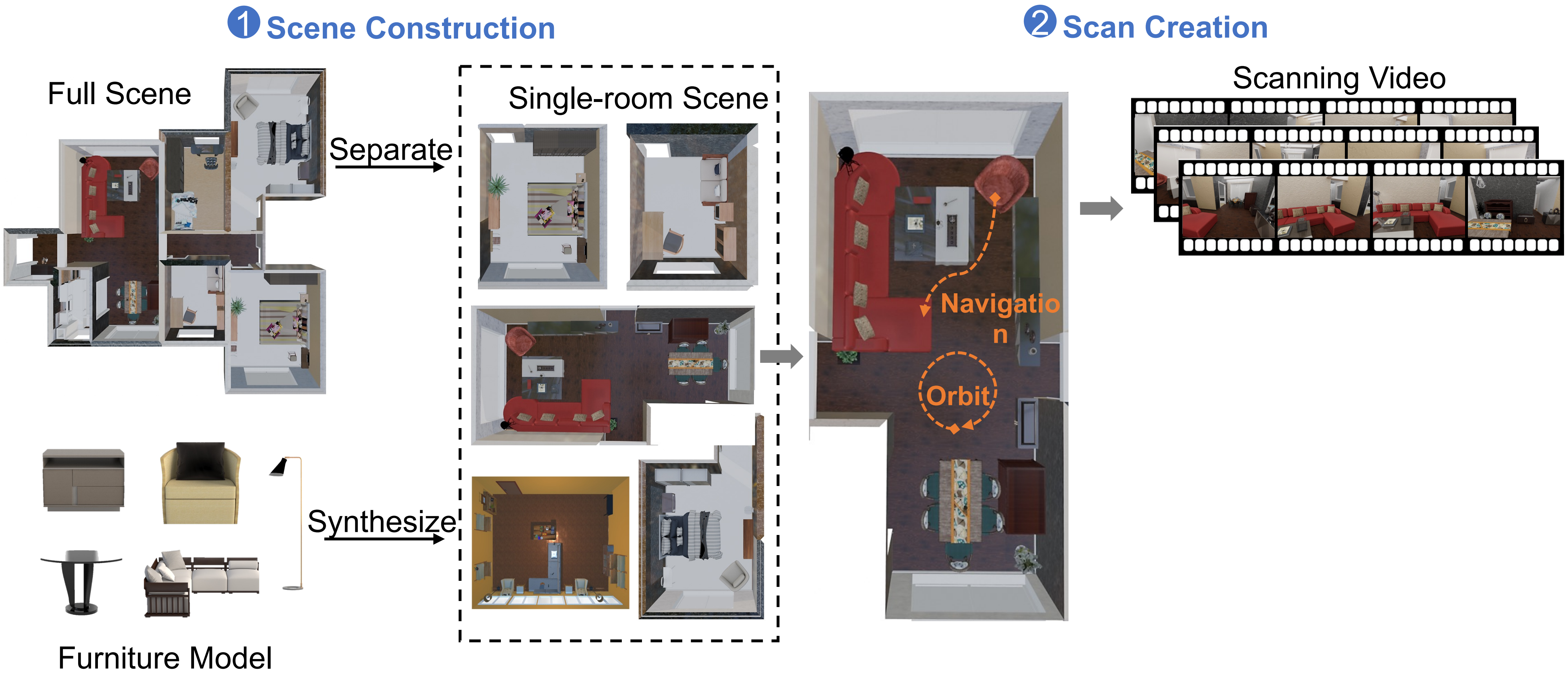
ScanForgeQA Dataset Construction

for scalable and extensible data sources

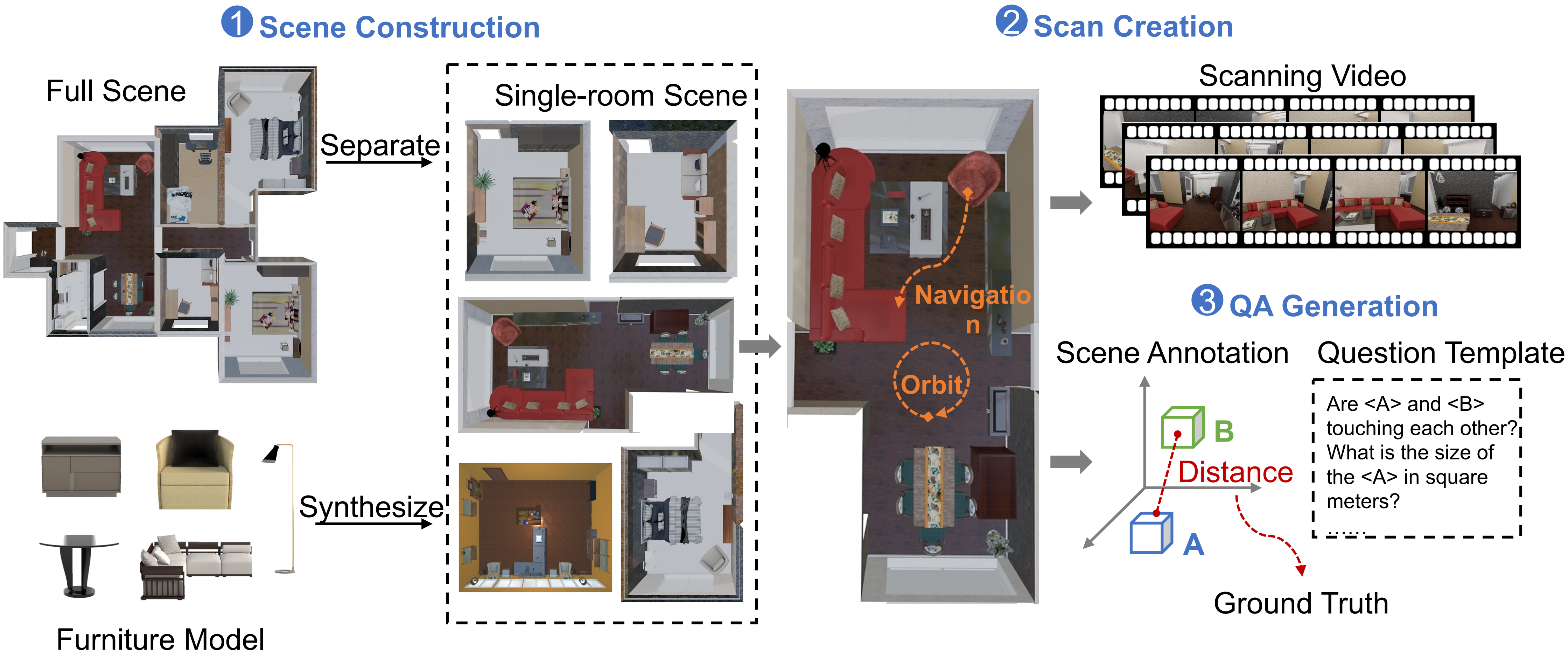
➢ ScanForgeQA Dataset Construction



➤ ScanForgeQA Dataset Construction



➢ ScanForgeQA Dataset Construction



➢ Performance Comparison

Results on VSI-Bench

Method	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	Avg	Δ
Close-source										
Human Level†	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0	79.2	-
Gemini-1.5 Pro†	49.6	28.8	58.6	49.4	46.0	48.1	42.0	68.0	48.8	-
Gemini-1.5 Pro	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6	45.4	-
+SpatialMind	63.9	51.8	70.2	47.3	56.3	45.9	42.6	44.3	52.8	↑ **7.4%**
GPT-4o	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5	34.0	-
+SpatialMind	40.0	27.1	62.7	40.9	41.0	39.6	37.1	38.5	40.8	↑ **6.8%**
Open-source										
InternVL2-8B	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6	34.6	-
+SpatialMind	35.8	28.9	49.7	44.4	37.2	34.8	35.1	45.5	38.9	↑ **4.3%**
+ScanForgeQA	45.3	33.4	54.8	45.0	41.1	36.1	33.4	43.0	41.5	↑ **6.9%**
+Both	47.0	32.8	53.2	46.6	39.8	36.8	37.9	47.5	42.7	↑ **8.1%**
InternVL2-40B	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6	36.0	-
+SpatialMind	36.4	30.0	49.1	41.8	43.8	36.1	35.6	50.0	40.4	↑ **4.4%**
+ScanForgeQA	51.0	29.2	52.7	38.1	47.2	36.4	35.9	47.6	42.3	↑ **6.3%**
+Both	52.2	30.5	54.4	41.0	50.5	37.0	40.2	50.3	44.5	↑ **8.5%**
Qwen2.5-VL-7B	40.3	22.2	50.1	38.9	38.0	40.7	31.4	35.9	37.2	-
+SpatialMind	45.1	25.2	52.1	41.4	38.7	41.6	34.7	34.5	39.2	↑ **2.0%**
+ScanForgeQA	53.2	30.5	56.8	44.9	42.3	44.0	37.3	37.7	43.3	↑ **6.1%**
+Both	55.0	29.5	57.3	44.0	43.5	44.3	38.3	39.2	43.9	↑ **6.7%**
Qwen2.5-VL-72B	37.9	28.6	57.4	49.8	45.5	38.4	20.6	35.4	39.2	-
+SpatialMind	42.3	32.0	61.7	53.8	48.2	43.9	30.4	39.3	44.0	↑ **4.8%**
+ScanForgeQA	45.2	32.7	63.3	52.4	50.1	41.7	32.8	40.2	44.8	↑ **5.6%**
+Both	48.6	34.4	68.9	54.7	53.4	43.9	30.1	42.7	47.1	↑ **7.9%**

➢ Performance Comparison

Results on OpenEQA, ScanQA, and SQA3D datasets

Method	OpenEQA Acc/Score	ScanQA BLEU-1	SQA3D EM-1
Qwen2.5-VL-7B	50.1/3.1	32.5	17.2
+SpatialMind	53.7/3.2	33.1	19.8
+ScanForgeQA	56.2/3.3	34.8	23.3
+Both	58.6/3.5	37.9	24.5
Qwen2.5-VL-72B	53.8/3.2	35.4	34.8
+SpatialMind	55.7/3.2	38.0	39.2
+ScanForgeQA	59.1/3.4	42.5	43.0
+Both	60.4/3.4	44.1	46.3

➢ Ablation Studies

On fine-tuning data and prompting strategy

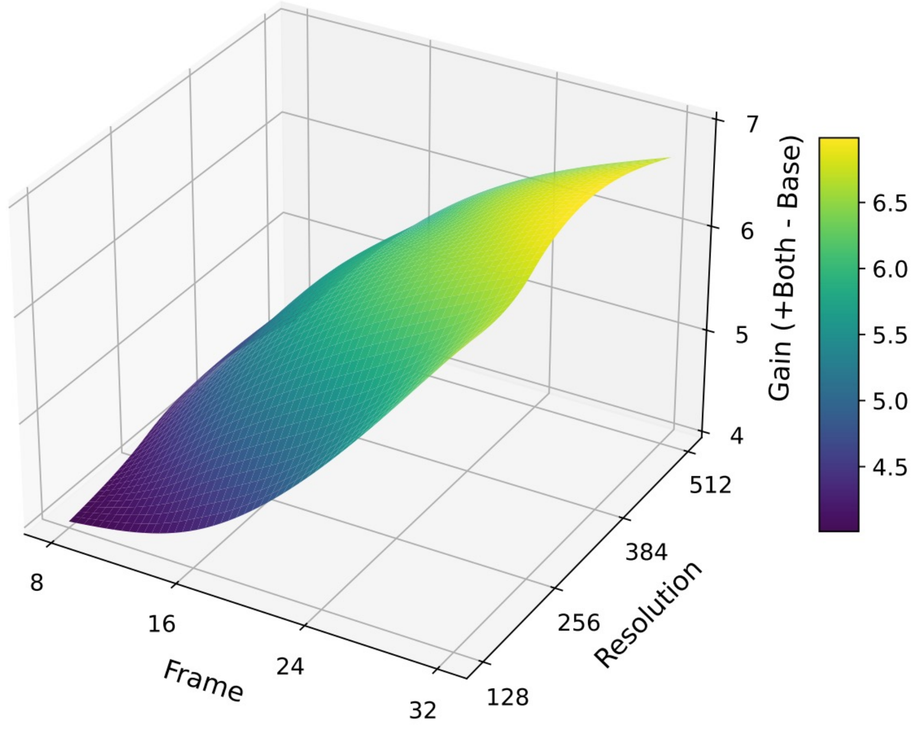
Method	Room Size	Avg
Qwen2.5-VL-7B	38.9	37.2
+SQA3D	38.8	38.9
+ScanQA	38.5	39.1
+ScanForgeQA	44.9	43.3
Qwen2.5-VL-72B	49.8	39.2
+CoT-Question	50.6	41.3
+CoT-Scene	52.1	42.7
+SpatialMind	53.8	44.0

➢ Ablation Studies

On fine-tuning data and prompting strategy

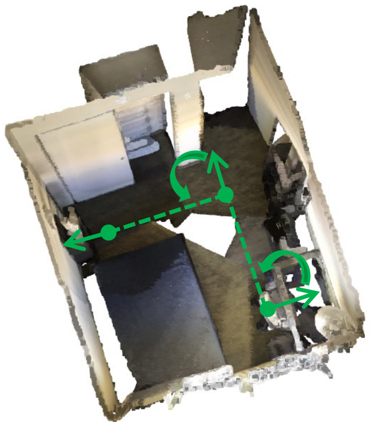
Method	Room Size	Avg
Qwen2.5-VL-7B	38.9	37.2
+SQA3D	38.8	38.9
+ScanQA	38.5	39.1
+ScanForgeQA	44.9	43.3
Qwen2.5-VL-72B	49.8	39.2
+CoT-Question	50.6	41.3
+CoT-Scene	52.1	42.7
+SpatialMind	53.8	44.0

On frames and resolution



➢ Case Studies

(a) Route Plan



You are a robot beginning at the chair and facing to lamp. You want to navigate to the lamp on the cabinet. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): **1. [please fill in] 2. Go forward until the wardrobe. 3. [please fill in]. 4. Go foward until the lamp**. You reached the final destination.

Qwen2.5-VL-7B: Turn Back, Turn Left

+Both (Ours) : Turn Left, Turn Left

(b) Appearance Order

What will be the first-time appearance order of the following categories in the video: **door, towel, refrigerator, microwave**?

Qwen2.5-VL-7B: door, towel, refrigerator, microwave

+Both (Ours) : towel, microwave, refrigerator, door

Thank you for watching!

Paper

Code

E-mail