Scaling Spatial Intelligence with Multimodal Foundation Models

Zhongang Cai

SenseTime Research

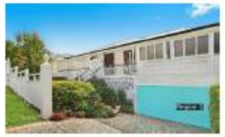
Spatial Intelligence

From Words to Worlds: Spatial Intelligence is AI’s Next Frontier

Huge Performance Gap

GPT-5-thinking

MM1

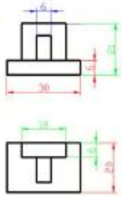


Question: What is the height of region 1 in meters?
GT: 2.7m.

Answer: ☑

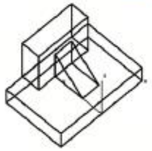
2m.

MR2

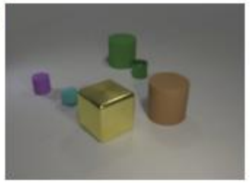


Question: Given the front, side and top-down view of a 3D object, analyze its structure and reconstruct it in 3D axis.

Answer: ☑

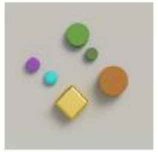


MR3

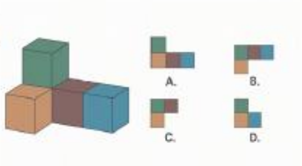


Question: Generate a 90 degrees top-down view of this scene.

Answer: ☑



MR4

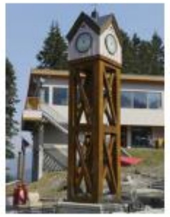


Question: Which option is the correct top-down view of the object?
GT: B.

Answer: ☒

A.

SR5



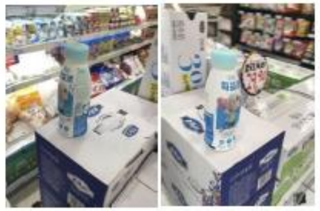
Question: Which object is higher in the 3D world space, the clock or the house in the back?
GT: The house in the back.

Answer: ☒

Clock.

GPT-5-thinking

PT6

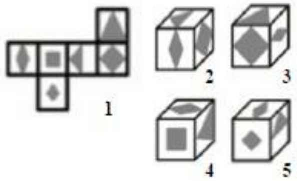


Question: The images are frames from a video. The first image is from the beginning of the video and the second image is from the end. Is the camera moving left or right when shooting the video?
GT: Left.

Answer: ☒

Right.

DA7

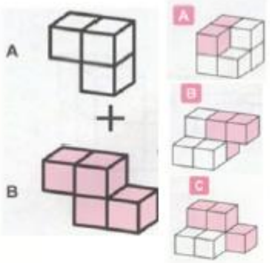


Question: Flip the shape in image 1 to form a 3D cube. Which of the image 2, 3, 4, 5 is a possible view of the formed cube?
GT: Image 4.

Answer: ☒

Image 2.

DA8

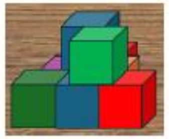


Question: Which of A, B, C is possible to be built when rotating and combining the two 3D structure in image 1?
GT: C.

Answer: ☒

A and B.

CR9

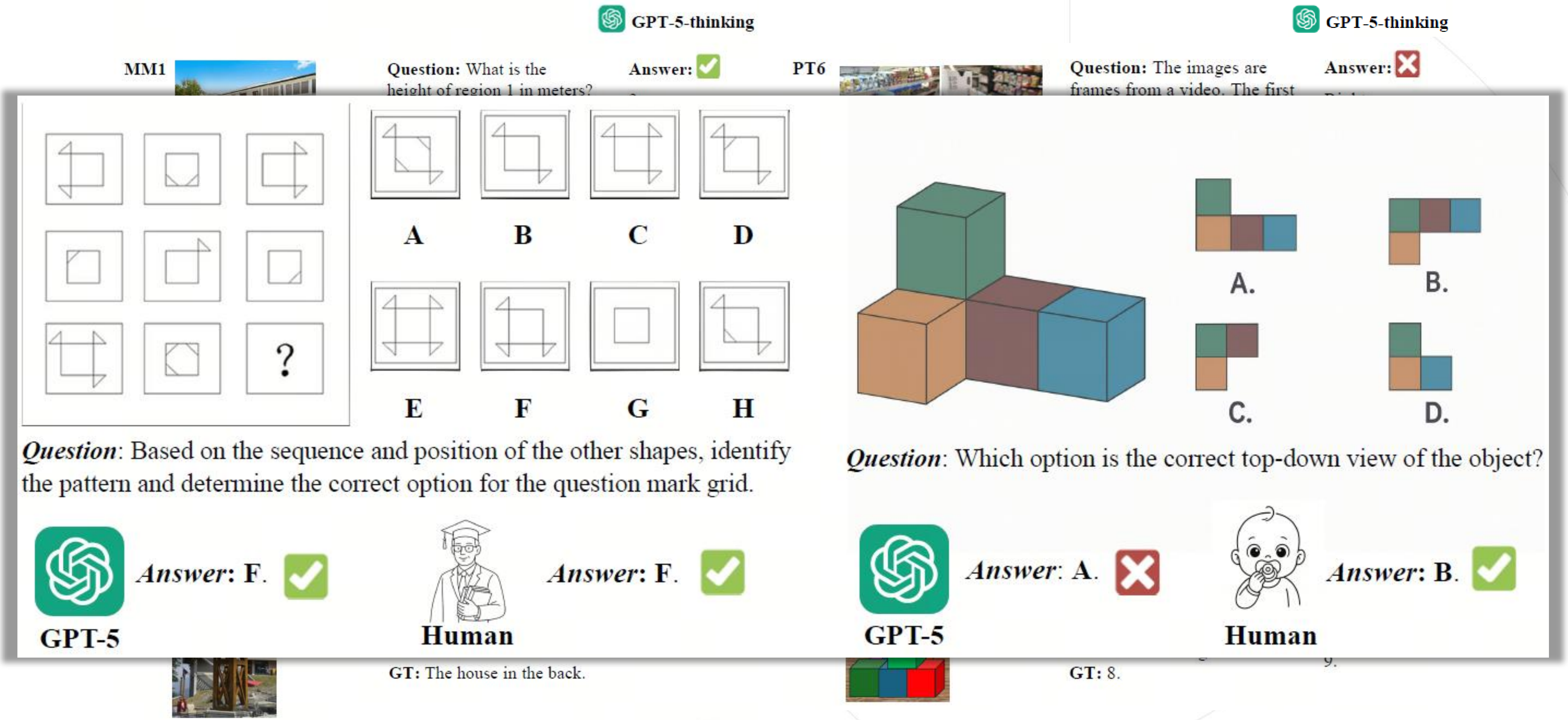


Question: How many 3D blocks in the image?
GT: 8.

Answer: ☒

9.

Huge Performance Gap


GPT-5-thinking
MM1
Question: What is the height of region 1 in meters?
Answer: ✓
PT6
Question: The images are frames from a video. The first
GPT-5-thinking
Answer: ☒
GT: The house in the back.
GT: 8.
9.
A
B
C
D
E
F
G
H
?
A.
B.
C.
D.

Question: Based on the sequence and position of the other shapes, identify the pattern and determine the correct option for the question mark grid.

GPT-5 — *Answer*: F. ✓

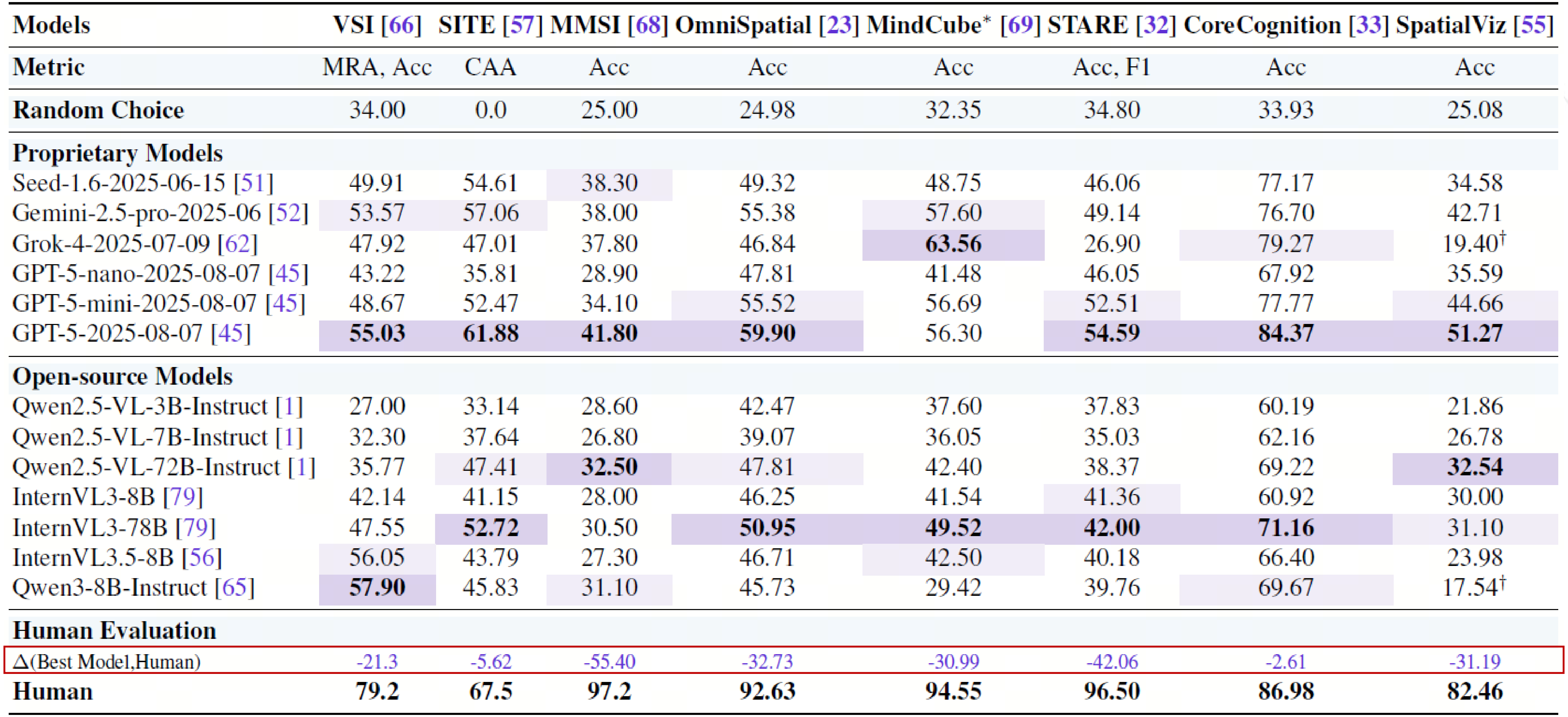
Human — *Answer*: F. ✓

Question: Which option is the correct top-down view of the object?

GPT-5 — *Answer*: A. ☒

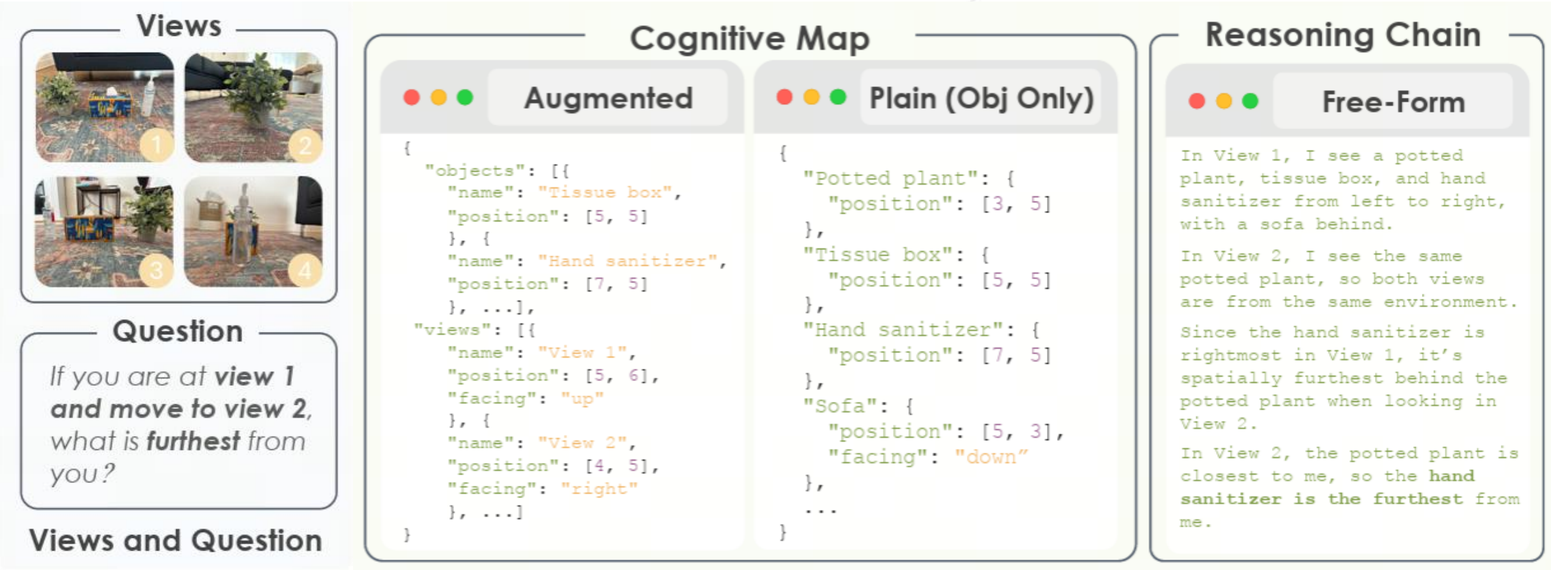
Human — *Answer*: B. ✓

Huge Performance Gap against Humans



Models	**VSI [66]**	**SITE [57]**	**MMSI [68]**	**OmniSpatial [23]**	**MindCube* [69]**	**STARE [32]**	**CoreCognition [33]**	**SpatialViz [55]**
Metric	MRA, Acc	CAA	Acc	Acc	Acc	Acc, F1	Acc	Acc
Random Choice	34.00	0.0	25.00	24.98	32.35	34.80	33.93	25.08
Proprietary Models								
Seed-1.6-2025-06-15 [51]	49.91	54.61	38.30	49.32	48.75	46.06	77.17	34.58
Gemini-2.5-pro-2025-06 [52]	53.57	57.06	38.00	55.38	57.60	49.14	76.70	42.71
Grok-4-2025-07-09 [62]	47.92	47.01	37.80	46.84	**63.56**	26.90	79.27	19.40†
GPT-5-nano-2025-08-07 [45]	43.22	35.81	28.90	47.81	41.48	46.05	67.92	35.59
GPT-5-mini-2025-08-07 [45]	48.67	52.47	34.10	55.52	56.69	52.51	77.77	44.66
GPT-5-2025-08-07 [45]	**55.03**	**61.88**	**41.80**	**59.90**	56.30	**54.59**	**84.37**	**51.27**
Open-source Models								
Qwen2.5-VL-3B-Instruct [1]	27.00	33.14	28.60	42.47	37.60	37.83	60.19	21.86
Qwen2.5-VL-7B-Instruct [1]	32.30	37.64	26.80	39.07	36.05	35.03	62.16	26.78
Qwen2.5-VL-72B-Instruct [1]	35.77	47.41	**32.50**	47.81	42.40	38.37	69.22	**32.54**
InternVL3-8B [79]	42.14	41.15	28.00	46.25	41.54	41.36	60.92	30.00
InternVL3-78B [79]	47.55	**52.72**	30.50	**50.95**	**49.52**	**42.00**	**71.16**	31.10
InternVL3.5-8B [56]	56.05	43.79	27.30	46.71	42.50	40.18	66.40	23.98
Qwen3-8B-Instruct [65]	**57.90**	45.83	31.10	45.73	29.42	39.76	69.67	17.54†
Human Evaluation								
Δ(Best Model,Human)	-21.3	-5.62	-55.40	-32.73	-30.99	-42.06	-2.61	-31.19
Human	**79.2**	**67.5**	**97.2**	**92.63**	**94.55**	**96.50**	**86.98**	**82.46**

Chain-of-Thought?



*If you are at **view 1** and move to view 2, what is **furthest** from you?*

```
{
  "objects": [{
    "name": "Tissue box",
    "position": [5, 5]
    }, {
    "name": "Hand sanitizer",
    "position": [7, 5]
    }, ...],
 "views": [{
    "name": "View 1",
    "position": [5, 6],
    "facing": "up"
    }, {
    "name": "View 2",
    "position": [4, 5],
    "facing": "right"
    }, ...]
}
```

```
{
  "Potted plant": {
    "position": [3, 5]
  },
  "Tissue box": {
    "position": [5, 5]
  },
  "Hand sanitizer": {
    "position": [7, 5]
  },
  "Sofa": {
    "position": [5, 3],
    "facing": "down"
  },
  ...
}
```

In View 1, I see a potted plant, tissue box, and hand sanitizer from left to right, with a sofa behind.

In View 2, I see the same potted plant, so both views are from the same environment.

Since the hand sanitizer is rightmost in View 1, it's spatially furthest behind the potted plant when looking in View 2.

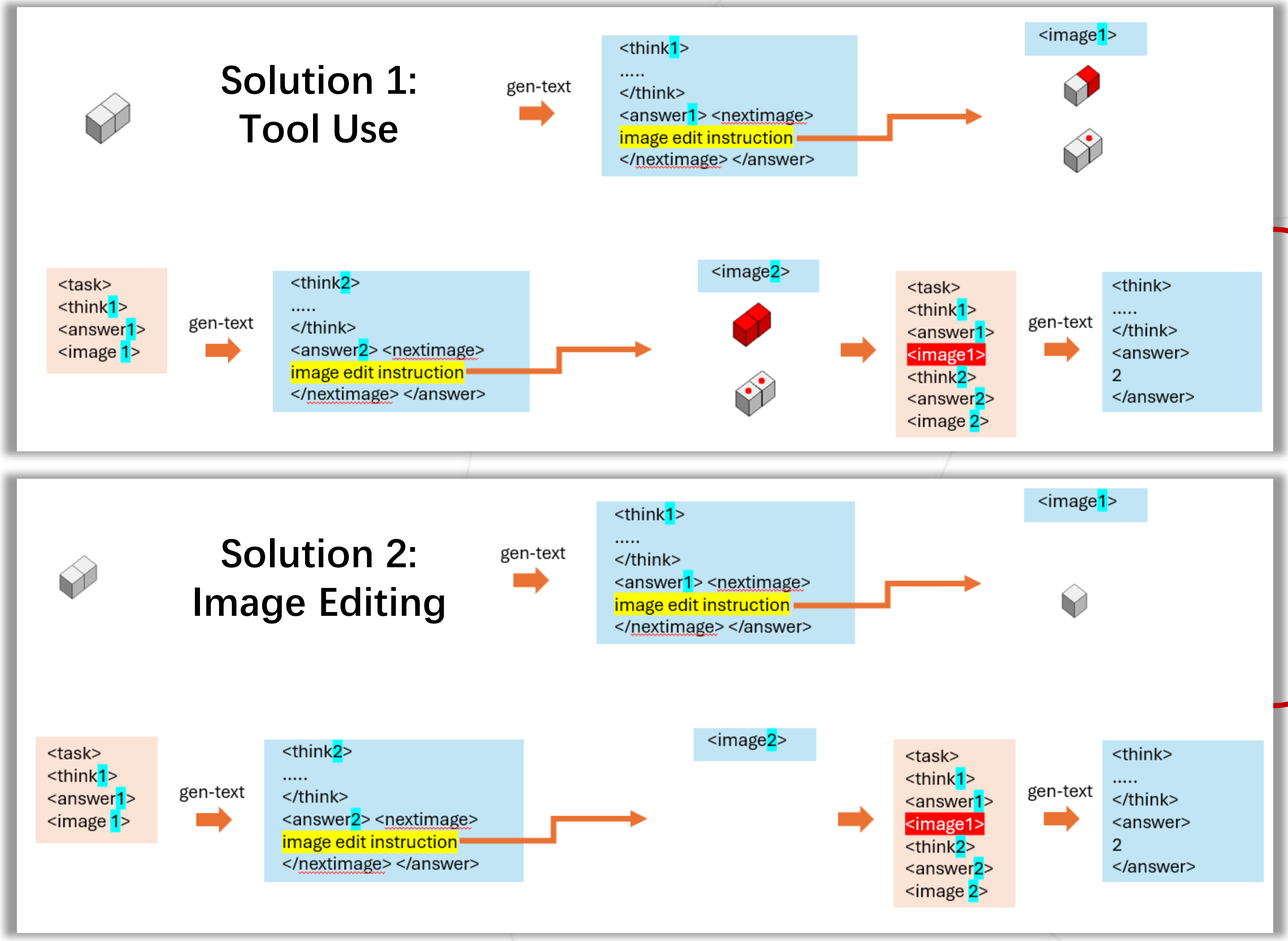
In View 2, the potted plant is closest to me, so the **hand sanitizer is the furthest** from me.

Saining Xie, Manling Li, Jiajun Wu, Li Fei-Fei, et al, *MindCube*, 2025

CoT Style	Average # Output Token	VSI-Bench: Obj. Rel. Direction			
		Overall	Easy	Medium	Hard
InternVL3-8B	1	39.3	48.8	47.0	21.9
No CoT	3.4	54.9	62.2	55.8	46.6
CoT-GPT-5	1070.7	40.0	41.4	43.1	36.1
CoT-MindCube-Aug-CGMap	1490.6	39.9	45.9	42.7	33.7
CoT-SenseNova-SI-CGMap	2262.8	47.9	53.9	51.3	41.0

Text-based CoT is not efficient or generalizable!

Visual Chain-of-Thought?



Solution 1:
Tool Use

gen-text

<think1>
.....
</think>
<answer1> <nextimage>
image edit instruction
</nextimage> </answer>

<image1>

<task>
<think1>
<answer1>
<image 1>

gen-text

<think2>
.....
</think>
<answer2> <nextimage>
image edit instruction
</nextimage> </answer>

<image2>

<task>
<think1>
<answer1>
<image1>
<think2>
<answer2>
<image 2>

gen-text

<think>
.....
</think>
<answer>
2
</answer>

Solution 2:
Image Editing

gen-text

<think1>
.....
</think>
<answer1> <nextimage>
image edit instruction
</nextimage> </answer>

<image1>

<task>
<think1>
<answer1>
<image 1>

gen-text

<think2>
.....
</think>
<answer2> <nextimage>
image edit instruction
</nextimage> </answer>

<image2>

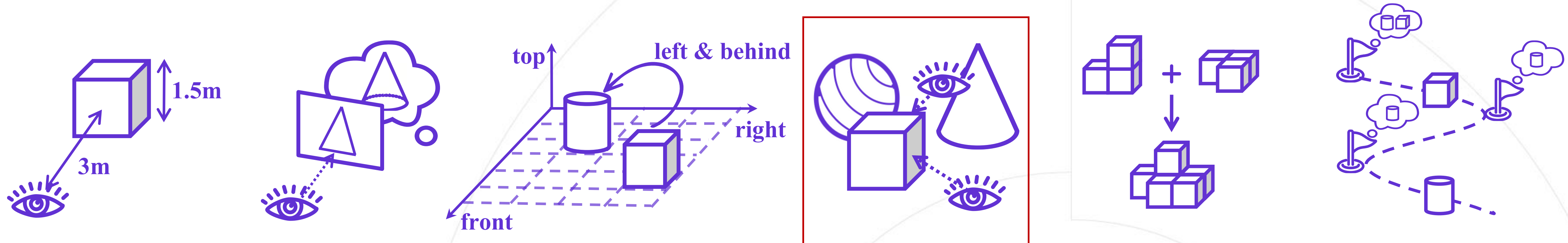
<task>
<think1>
<answer1>
<image1>
<think2>
<answer2>
<image 2>

gen-text

<think>
.....
</think>
<answer>
2
</answer>

Visual CoT did not achieve stability or generalizability either

Fundamental Capabilities



Metric Measurement

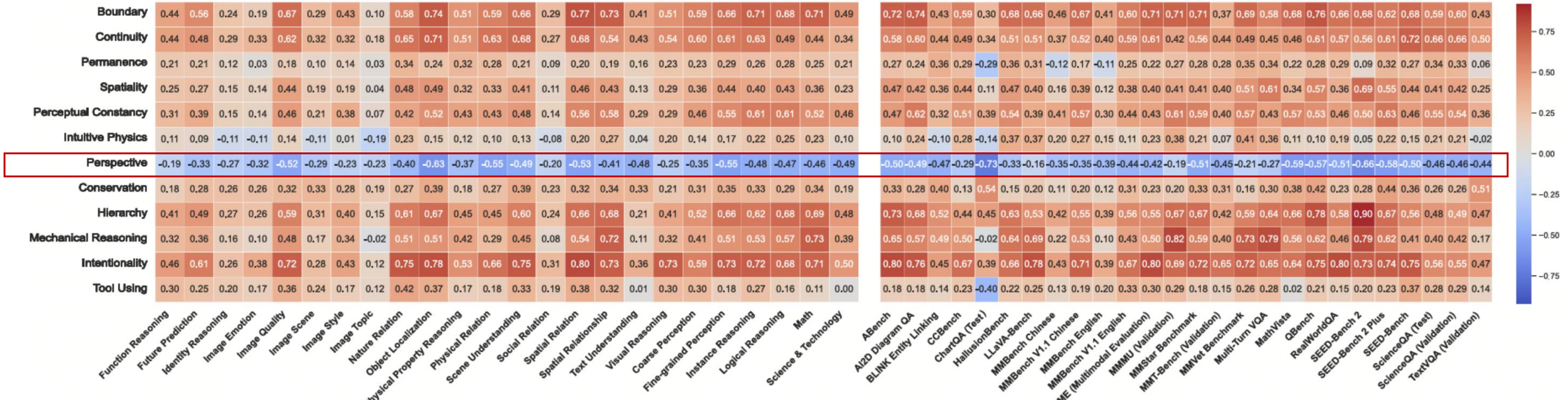
Mental Reconstruction

Spatial Relations

Perspective-taking

Deformation & Assembly

Comprehensive Reasoning

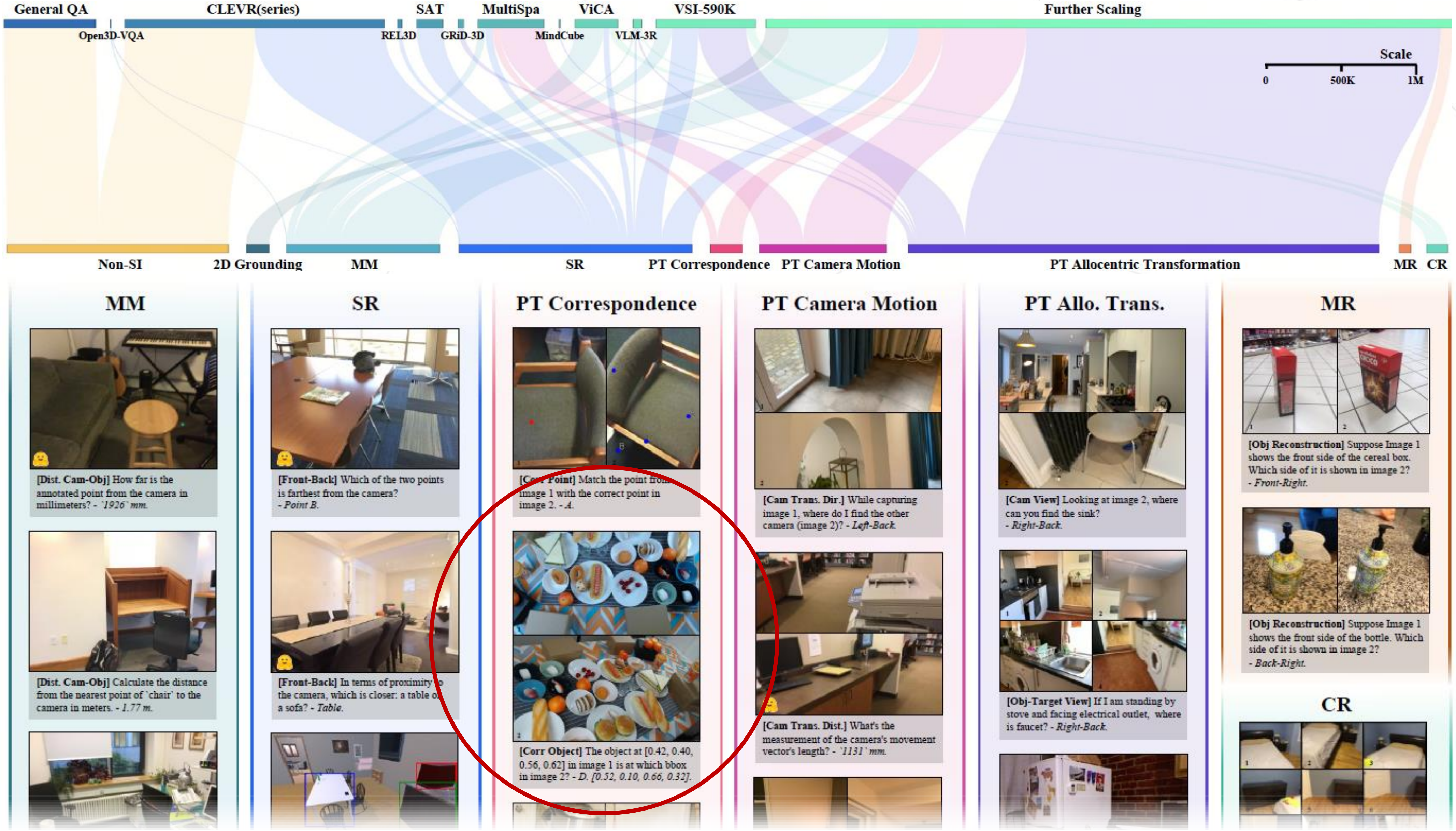


	Function Reasoning	Future Prediction	Identity Reasoning	Image Emotion	Image Quality	Image Scene	Image Style	Image Topic	Nature Relation	Object Localization	Physical Property Reasoning	Physical Relation	Scene Understanding	Social Relation	Spatial Relation	Spatial Relationship	Text Understanding	Visual Reasoning	Coarse Perception	Fine-grained Perception	Instance Reasoning	Logical Reasoning	Math	Science & Technology
Boundary	0.44	0.56	0.24	0.19	0.67	0.29	0.43	0.10	0.58	0.74	0.51	0.59	0.66	0.29	0.77	0.73	0.41	0.51	0.59	0.66	0.71	0.68	0.71	0.49
Continuity	0.44	0.48	0.29	0.33	0.62	0.32	0.32	0.18	0.65	0.71	0.51	0.63	0.68	0.27	0.68	0.54	0.43	0.54	0.60	0.61	0.63	0.49	0.44	0.34
Permanence	0.21	0.21	0.12	0.03	0.18	0.10	0.14	0.03	0.34	0.24	0.32	0.28	0.21	0.09	0.20	0.19	0.16	0.23	0.23	0.29	0.26	0.28	0.25	0.21
Spatiality	0.25	0.27	0.15	0.14	0.44	0.19	0.19	0.04	0.48	0.49	0.32	0.33	0.41	0.11	0.46	0.43	0.13	0.29	0.36	0.44	0.40	0.43	0.36	0.23
Perceptual Constancy	0.31	0.39	0.15	0.14	0.46	0.21	0.38	0.07	0.42	0.52	0.43	0.43	0.48	0.14	0.56	0.58	0.29	0.29	0.46	0.55	0.61	0.61	0.52	0.46
Intuitive Physics	0.11	0.09	-0.11	-0.11	0.14	-0.11	0.01	-0.19	0.23	0.15	0.12	0.10	0.13	-0.08	0.20	0.27	0.04	0.20	0.14	0.17	0.22	0.25	0.23	0.10
Perspective	-0.19	-0.33	-0.27	-0.32	-0.52	-0.29	-0.23	-0.23	-0.40	-0.63	-0.37	-0.55	-0.49	-0.20	-0.53	-0.41	-0.48	-0.25	-0.35	-0.55	-0.48	-0.47	-0.46	-0.49
Conservation	0.18	0.28	0.26	0.26	0.32	0.33	0.28	0.19	0.27	0.39	0.18	0.27	0.39	0.23	0.32	0.34	0.33	0.21	0.31	0.35	0.33	0.29	0.34	0.19
Hierarchy	0.41	0.49	0.27	0.26	0.59	0.31	0.40	0.15	0.61	0.67	0.45	0.45	0.60	0.24	0.66	0.68	0.21	0.41	0.52	0.66	0.62	0.68	0.69	0.48
Mechanical Reasoning	0.32	0.36	0.16	0.10	0.48	0.17	0.34	-0.02	0.51	0.51	0.42	0.29	0.45	0.08	0.54	0.72	0.11	0.32	0.41	0.51	0.53	0.57	0.73	0.39
Intentionality	0.46	0.61	0.26	0.38	0.72	0.28	0.43	0.12	0.75	0.78	0.53	0.66	0.75	0.31	0.80	0.73	0.36	0.73	0.59	0.73	0.72	0.68	0.71	0.50
Tool Using	0.30	0.25	0.20	0.17	0.36	0.24	0.17	0.12	0.42	0.37	0.17	0.18	0.33	0.19	0.38	0.32	0.01	0.30	0.30	0.18	0.27	0.16	0.11	0.00

	ABench	AI2D Diagram QA	BLINK Entity Linking	CCBench	ChartQA (Test)	HallusionBench	LLaVA-Bench	MMBench Chinese	MMBench V1.1 Chinese	MMBench English	MMBench V1.1 English	MME (Multimodal Evaluation)	MMMU (Validation)	MMStar Benchmark	MMT-Bench (Validation)	MMVet Benchmark	Multi-Turn VQA	MathVista	QBench	RealWorldQA	SEED-Bench 2	SEED-Bench 2 Plus	SEED-Bench	ScienceQA (Test)	ScienceQA (Validation)	TextVQA (Validation)
Boundary	0.72	0.74	0.43	0.59	0.30	0.68	0.66	0.46	0.67	0.41	0.60	0.71	0.71	0.71	0.37	0.69	0.58	0.68	0.76	0.66	0.68	0.62	0.68	0.59	0.60	0.43
Continuity	0.58	0.60	0.44	0.49	0.34	0.51	0.51	0.37	0.52	0.40	0.59	0.61	0.42	0.56	0.44	0.49	0.45	0.46	0.61	0.57	0.56	0.61	0.72	0.66	0.66	0.50
Permanence	0.27	0.24	0.36	0.29	-0.29	0.36	0.31	-0.12	0.17	-0.11	0.25	0.22	0.27	0.28	0.28	0.35	0.34	0.22	0.28	0.29	0.09	0.32	0.27	0.34	0.33	0.06
Spatiality	0.47	0.42	0.36	0.44	0.11	0.47	0.40	0.16	0.39	0.12	0.38	0.40	0.41	0.41	0.40	0.51	0.61	0.34	0.57	0.36	0.69	0.55	0.44	0.41	0.42	0.25
Perceptual Constancy	0.47	0.62	0.32	0.51	0.39	0.54	0.39	0.41	0.57	0.30	0.44	0.43	0.61	0.59	0.40	0.57	0.43	0.57	0.53	0.46	0.50	0.63	0.46	0.55	0.54	0.36
Intuitive Physics	0.10	0.24	-0.10	0.28	-0.14	0.37	0.37	0.20	0.27	0.15	0.11	0.23	0.38	0.21	0.07	0.41	0.36	0.11	0.10	0.19	0.05	0.22	0.15	0.21	0.21	-0.02
Perspective	-0.50	-0.49	-0.47	-0.29	-0.73	-0.33	-0.16	-0.35	-0.35	-0.39	-0.44	-0.42	-0.19	-0.51	-0.45	-0.21	-0.27	-0.59	-0.57	-0.51	-0.66	-0.58	-0.50	-0.46	-0.46	-0.44
Conservation	0.33	0.28	0.40	0.13	0.54	0.15	0.20	0.11	0.20	0.12	0.31	0.23	0.20	0.33	0.31	0.16	0.30	0.38	0.42	0.23	0.28	0.44	0.36	0.26	0.26	0.51
Hierarchy	0.73	0.68	0.52	0.44	0.45	0.63	0.53	0.42	0.55	0.39	0.56	0.55	0.67	0.67	0.42	0.59	0.64	0.66	0.78	0.58	0.90	0.67	0.56	0.48	0.49	0.47
Mechanical Reasoning	0.65	0.57	0.49	0.50	-0.02	0.64	0.69	0.22	0.53	0.10	0.43	0.50	0.82	0.59	0.40	0.73	0.79	0.56	0.62	0.46	0.79	0.62	0.41	0.40	0.42	0.17
Intentionality	0.80	0.76	0.45	0.67	0.39	0.66	0.78	0.43	0.71	0.39	0.67	0.80	0.69	0.72	0.65	0.72	0.65	0.64	0.75	0.80	0.73	0.74	0.75	0.56	0.55	0.47
Tool Using	0.18	0.18	0.14	0.23	-0.40	0.22	0.25	0.13	0.19	0.20	0.33	0.30	0.29	0.18	0.15	0.26	0.28	0.02	0.21	0.15	0.20	0.23	0.37	0.28	0.29	0.14

Hokin Deng *et al.*, *CoreCognition*, 2025

Scaling Spatial Intelligence



[Dist. Cam-Obj] How far is the annotated point from the camera in millimeters? - *`1926` mm.*

[Dist. Cam-Obj] Calculate the distance from the nearest point of `chair` to the camera in meters. - *1.77 m.*

[Front-Back] Which of the two points is farthest from the camera? - *Point B.*

[Front-Back] In terms of proximity to the camera, which is closer: a table or a sofa? - *Table.*

[Corr Point] Match the point from image 1 with the correct point in image 2. - *A.*

[Corr Object] The object at [0.42, 0.40, 0.56, 0.62] in image 1 is at which bbox in image 2? - *D. [0.52, 0.10, 0.66, 0.32].*

[Cam Trans. Dir.] While capturing image 1, where do I find the other camera (image 2)? - *Left-Back.*

[Cam Trans. Dist.] What's the measurement of the camera's movement vector's length? - *`1131` mm.*

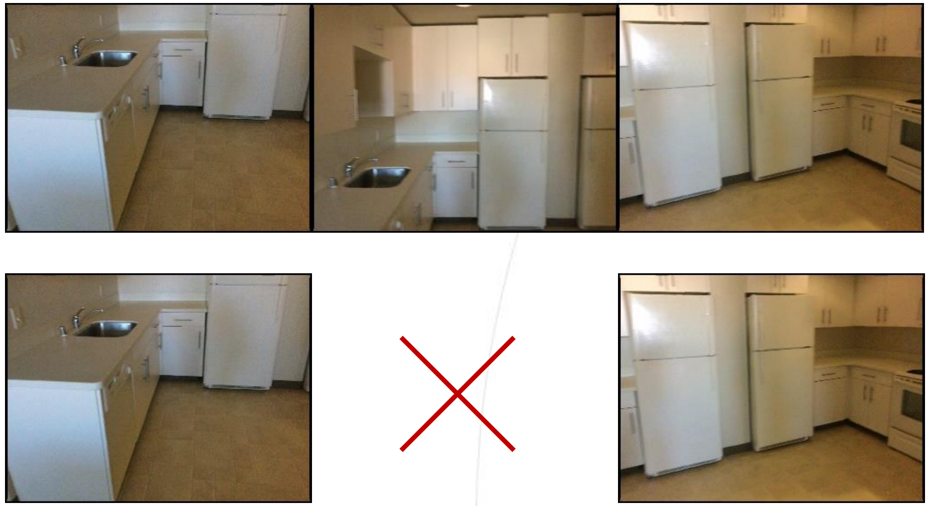
[Cam View] Looking at image 2, where can you find the sink? - *Right-Back.*

[Obj-Target View] If I am standing by stove and facing electrical outlet, where is faucet? - *Right-Back.*

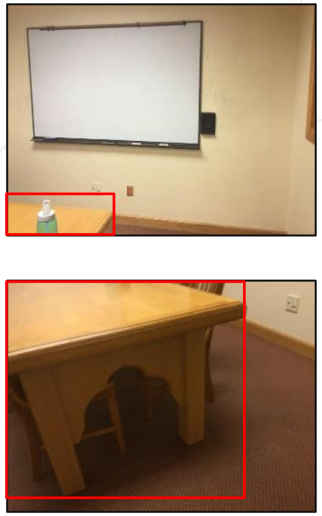
[Obj Reconstruction] Suppose Image 1 shows the front side of the cereal box. Which side of it is shown in image 2? - *Front-Right.*

[Obj Reconstruction] Suppose Image 1 shows the front side of the bottle. Which side of it is shown in image 2? - *Back-Right.*

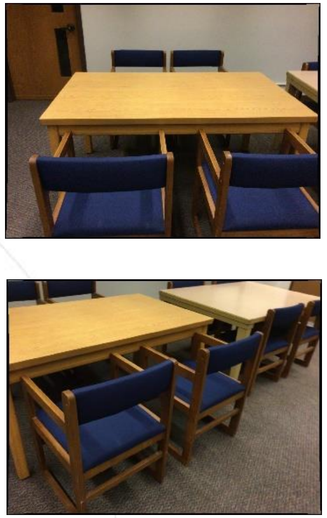
A Few Tips for Good Spatial Data



Minimize Redundant Information

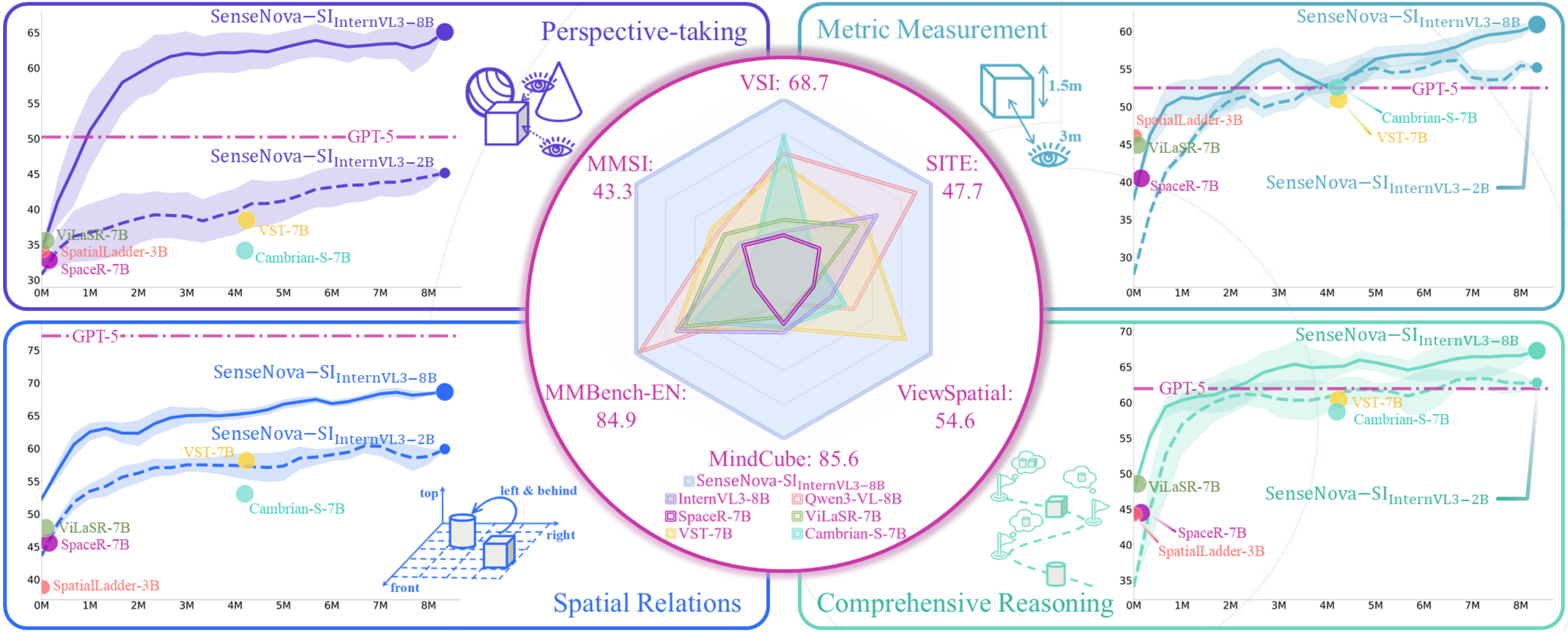


Avoid Insufficient Information

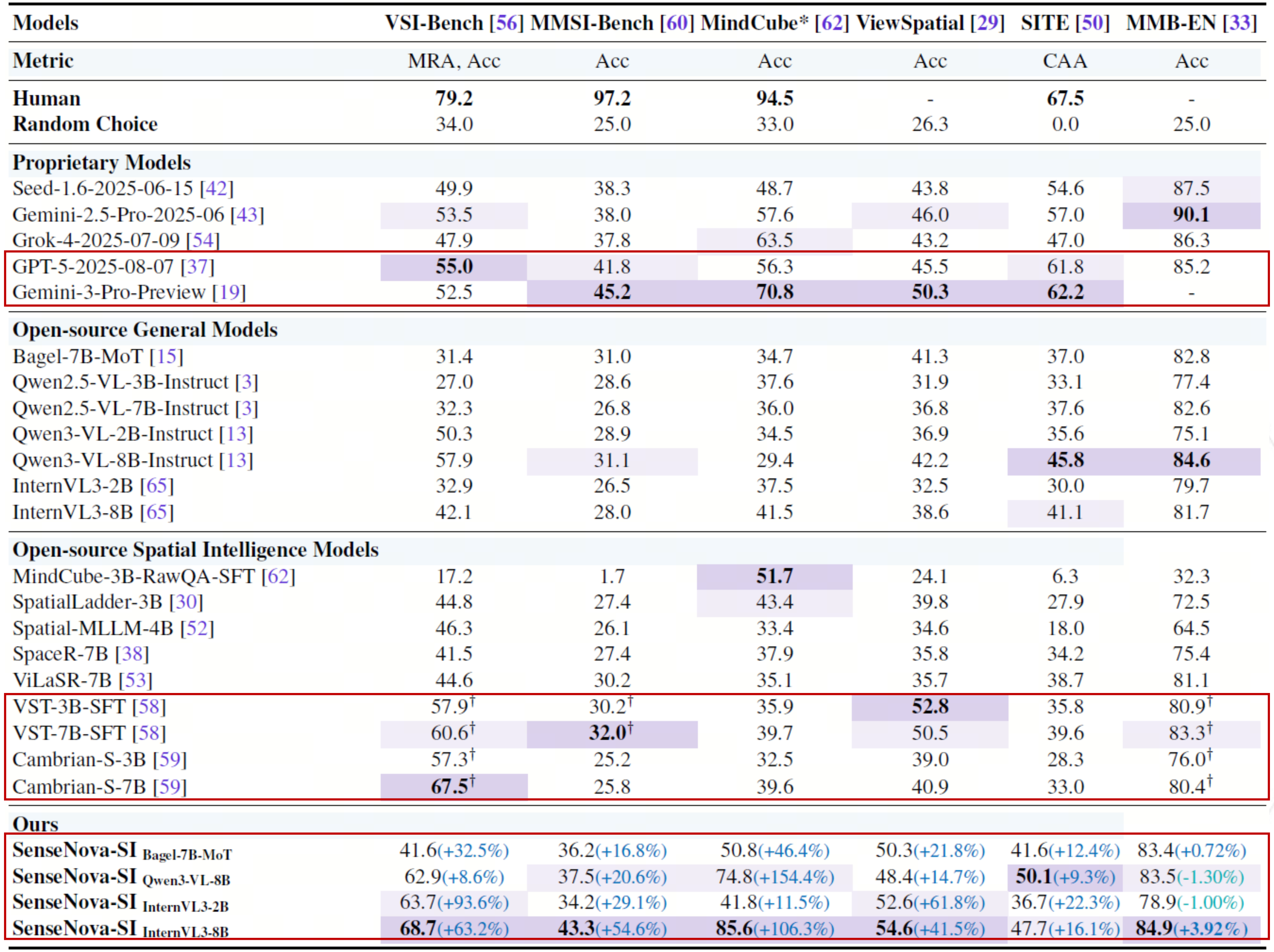


Avoid Ambiguity

Scaling Spatial Intelligence



Scaling Spatial Intelligence



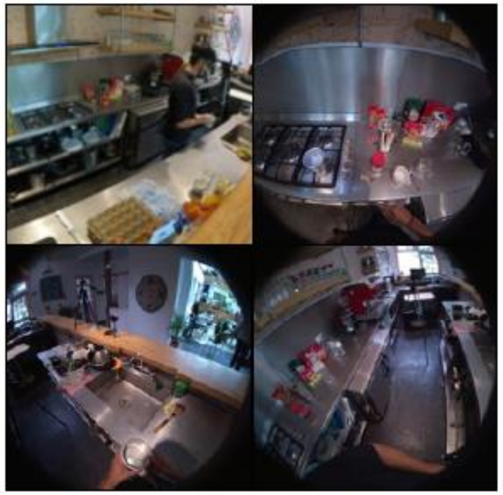
Models	VSI-Bench [56]	MMSI-Bench [60]	MindCube* [62]	ViewSpatial [29]	SITE [50]	MMB-EN [33]
Metric	MRA, Acc	Acc	Acc	Acc	CAA	Acc
Human	**79.2**	**97.2**	**94.5**	-	**67.5**	-
Random Choice	34.0	25.0	33.0	26.3	0.0	25.0
Proprietary Models						
Seed-1.6-2025-06-15 [42]	49.9	38.3	48.7	43.8	54.6	87.5
Gemini-2.5-Pro-2025-06 [43]	53.5	38.0	57.6	46.0	57.0	**90.1**
Grok-4-2025-07-09 [54]	47.9	37.8	63.5	43.2	47.0	86.3
GPT-5-2025-08-07 [37]	**55.0**	41.8	56.3	45.5	61.8	85.2
Gemini-3-Pro-Preview [19]	52.5	**45.2**	**70.8**	**50.3**	**62.2**	-
Open-source General Models						
Bagel-7B-MoT [15]	31.4	31.0	34.7	41.3	37.0	82.8
Qwen2.5-VL-3B-Instruct [3]	27.0	28.6	37.6	31.9	33.1	77.4
Qwen2.5-VL-7B-Instruct [3]	32.3	26.8	36.0	36.8	37.6	82.6
Qwen3-VL-2B-Instruct [13]	50.3	28.9	34.5	36.9	35.6	75.1
Qwen3-VL-8B-Instruct [13]	57.9	31.1	29.4	42.2	**45.8**	**84.6**
InternVL3-2B [65]	32.9	26.5	37.5	32.5	30.0	79.7
InternVL3-8B [65]	42.1	28.0	41.5	38.6	41.1	81.7
Open-source Spatial Intelligence Models						
MindCube-3B-RawQA-SFT [62]	17.2	1.7	**51.7**	24.1	6.3	32.3
SpatialLadder-3B [30]	44.8	27.4	43.4	39.8	27.9	72.5
Spatial-MLLM-4B [52]	46.3	26.1	33.4	34.6	18.0	64.5
SpaceR-7B [38]	41.5	27.4	37.9	35.8	34.2	75.4
ViLaSR-7B [53]	44.6	30.2	35.1	35.7	38.7	81.1
VST-3B-SFT [58]	57.9†	30.2†	35.9	**52.8**	35.8	80.9†
VST-7B-SFT [58]	60.6†	**32.0**†	39.7	50.5	39.6	83.3†
Cambrian-S-3B [59]	57.3†	25.2	32.5	39.0	28.3	76.0†
Cambrian-S-7B [59]	**67.5**†	25.8	39.6	40.9	33.0	80.4†
Ours						
SenseNova-SI $_{\text{Bagel-7B-MoT}}$	41.6(+32.5%)	36.2(+16.8%)	50.8(+46.4%)	50.3(+21.8%)	41.6(+12.4%)	83.4(+0.72%)
SenseNova-SI $_{\text{Qwen3-VL-8B}}$	62.9(+8.6%)	37.5(+20.6%)	74.8(+154.4%)	48.4(+14.7%)	**50.1**(+9.3%)	83.5(-1.30%)
SenseNova-SI $_{\text{InternVL3-2B}}$	63.7(+93.6%)	34.2(+29.1%)	41.8(+11.5%)	52.6(+61.8%)	36.7(+22.3%)	78.9(-1.00%)
SenseNova-SI $_{\text{InternVL3-8B}}$	**68.7**(+63.2%)	**43.3**(+54.6%)	**85.6**(+106.3%)	**54.6**(+41.5%)	47.7(+16.1%)	**84.9**(+3.92%)

Best Proprietary

Best Expert

SenseNova-SI

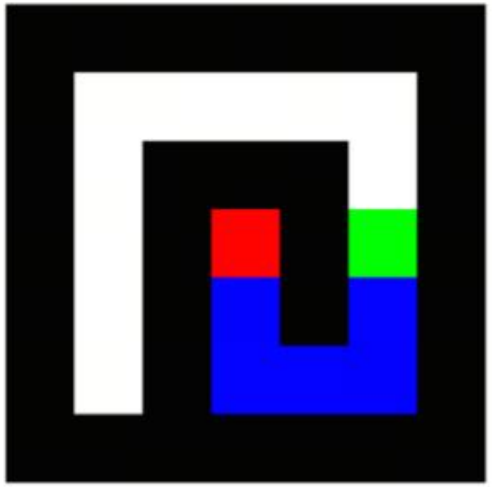
Early Signs of Emergent Intelligence



[Obj-Orient View (Ego-Exo4D)]

Which egocentric view image correctly matches the exocentric view?

- *B.*

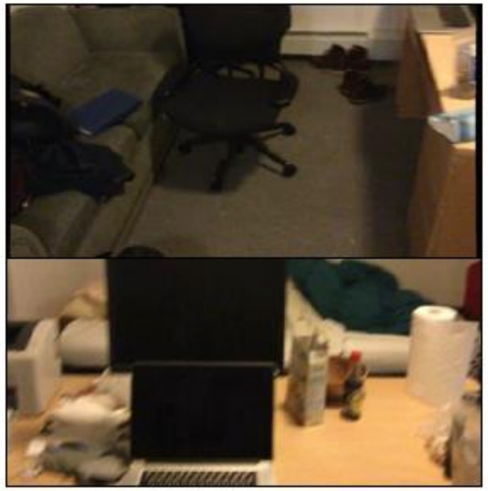


[SITE Mov&Nav-Maze]

How many right turns are there in the provided path (marked by Blue) from S (green block) to E (red block)?

– *C. 2.*

90.4%

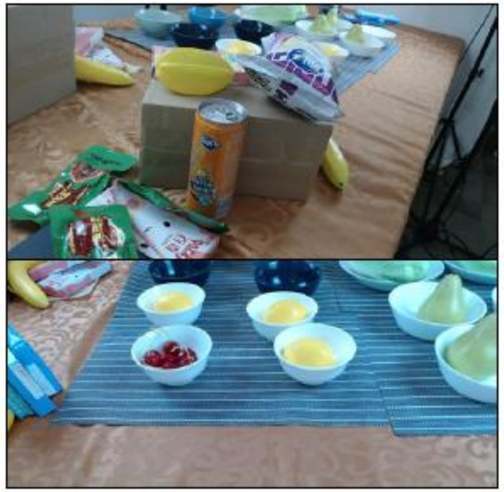


[MMSI Pos-Cam-Cam]

Camera was facing the west side of the room when the first picture was taken, which direction is the camera facing in the room when the second picture is taken?

– *C. North.*

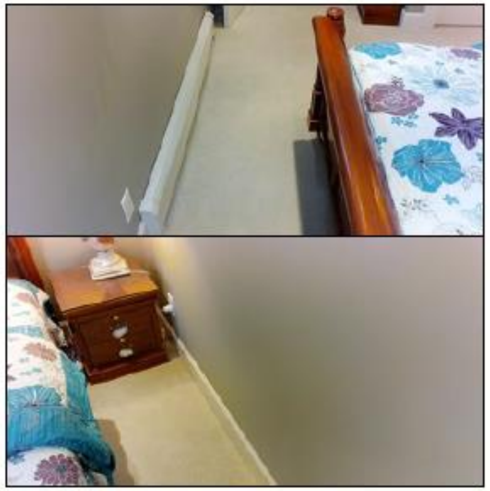
23.8%



[Cam Rotation (MessyTable)]

From image 1 to 2. What is the rotation direction?

- *Rotated to the Left.*

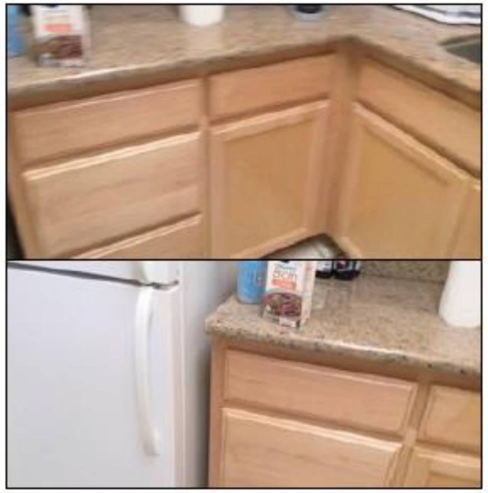


[MMSI Pos-Cam-Cam]

Camera coordinate system +Y up, -Z forward, right-handed. How can the first image be obtained from the second image?

– *A. Rotate a positive angle around the Y axis.*

66.7%



[MMSI Attr-Appr]

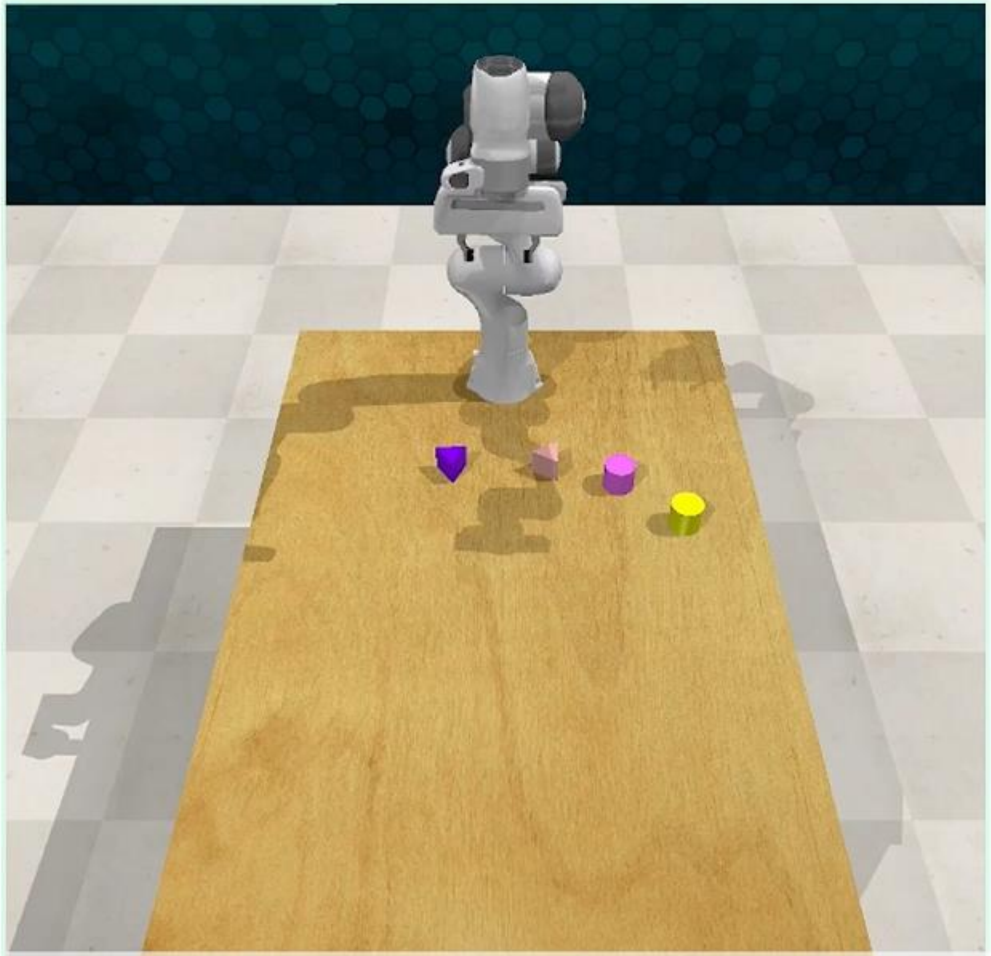
How many different drawers with a width greater than their height appear in total?

– *D. 5.*

25.6%

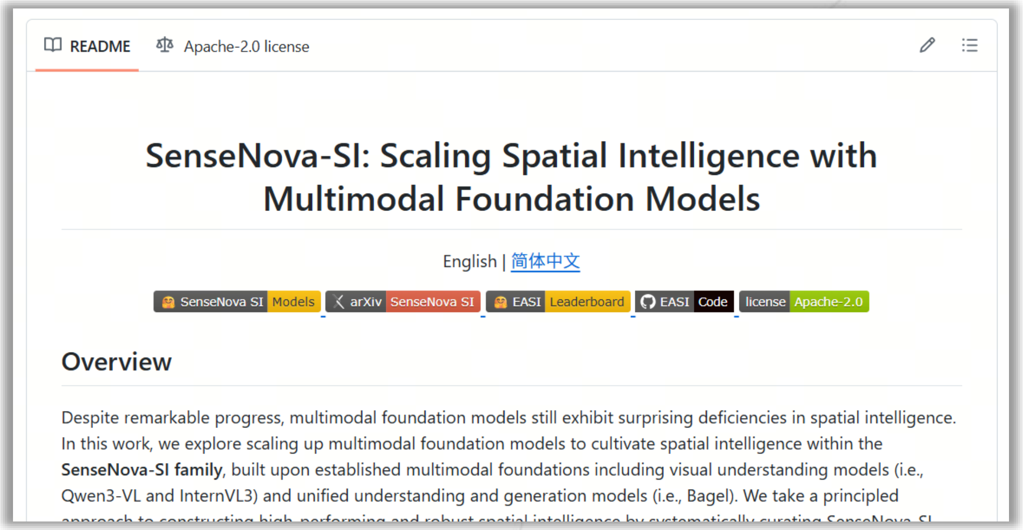
Downstream Task

SSI-InternVL3-8B



Stack the left triangular prism on top of the right cylinder.

Open-source



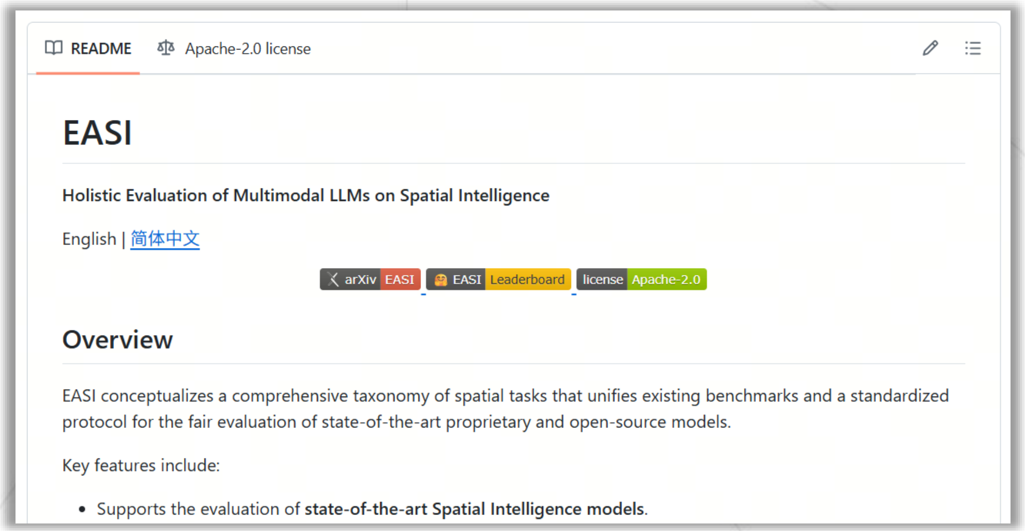
SenseNova-SI: Scaling Spatial Intelligence with Multimodal Foundation Models

English | 简体中文

Overview

Despite remarkable progress, multimodal foundation models still exhibit surprising deficiencies in spatial intelligence. In this work, we explore scaling up multimodal foundation models to cultivate spatial intelligence within the **SenseNova-SI family**, built upon established multimodal foundations including visual understanding models (i.e., Qwen3-VL and InternVL3) and unified understanding and generation models (i.e., Bagel). We take a principled approach to constructing high-performing and robust spatial intelligence by systematically curating SenseNova-SI

Code for SenseNova-SI



EASI

Holistic Evaluation of Multimodal LLMs on Spatial Intelligence

English | 简体中文

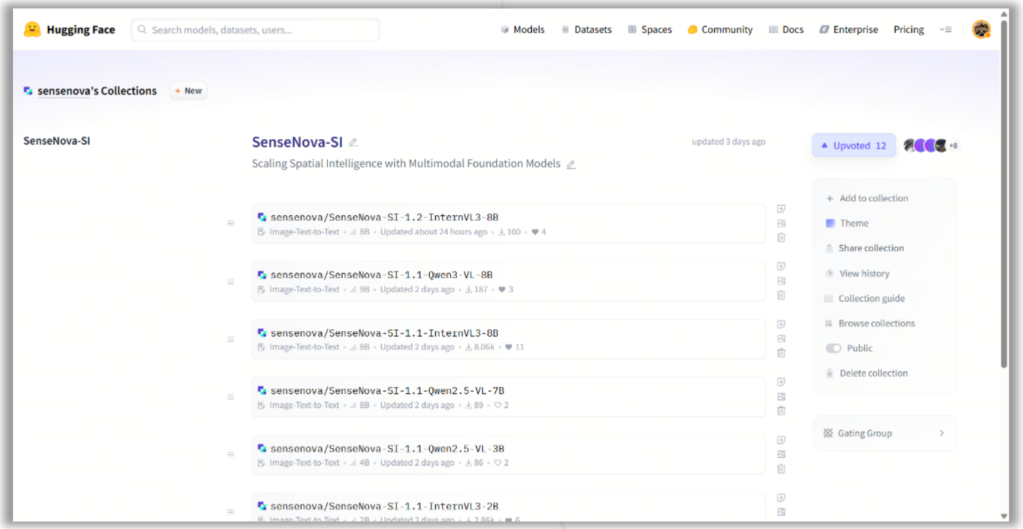
Overview

EASI conceptualizes a comprehensive taxonomy of spatial tasks that unifies existing benchmarks and a standardized protocol for the fair evaluation of state-of-the-art proprietary and open-source models.

Key features include:

- Supports the evaluation of **state-of-the-art Spatial Intelligence models**.

EASI: Evaluation Toolbox



SenseNova-SI

Scaling Spatial Intelligence with Multimodal Foundation Models

updated 3 days ago

sensenova/SenseNova-SI-1.2-InternVL3-8B
Image-Text-to-Text · 8B · Updated about 24 hours ago · 100 · 4

sensenova/SenseNova-SI-1.1-Qwen3-VL-8B
Image-Text-to-Text · 9B · Updated 2 days ago · 187 · 3

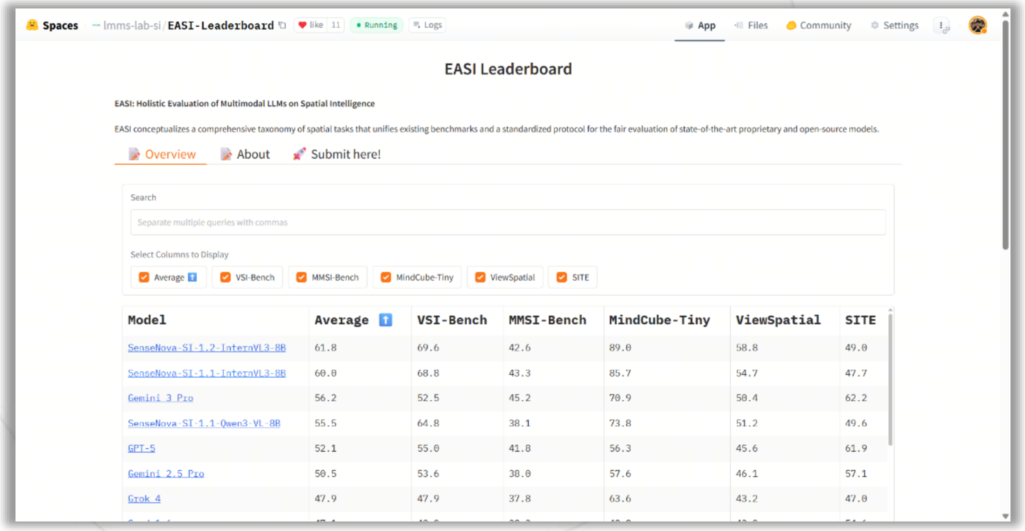
sensenova/SenseNova-SI-1.1-InternVL3-8B
Image-Text-to-Text · 8B · Updated 2 days ago · 8.06k · 11

sensenova/SenseNova-SI-1.1-Qwen2.5-VL-7B
Image-Text-to-Text · 8B · Updated 2 days ago · 89 · 2

sensenova/SenseNova-SI-1.1-Qwen2.5-VL-3B
Image-Text-to-Text · 4B · Updated 2 days ago · 86 · 2

sensenova/SenseNova-SI-1.1-InternVL3-2B
Image-Text-to-Text · 2B · Updated 2 days ago · 2.86k · 6

Weights for SenseNova-SI



EASI Leaderboard

EASI: Holistic Evaluation of Multimodal LLMs on Spatial Intelligence

EASI conceptualizes a comprehensive taxonomy of spatial tasks that unifies existing benchmarks and a standardized protocol for the fair evaluation of state-of-the-art proprietary and open-source models.

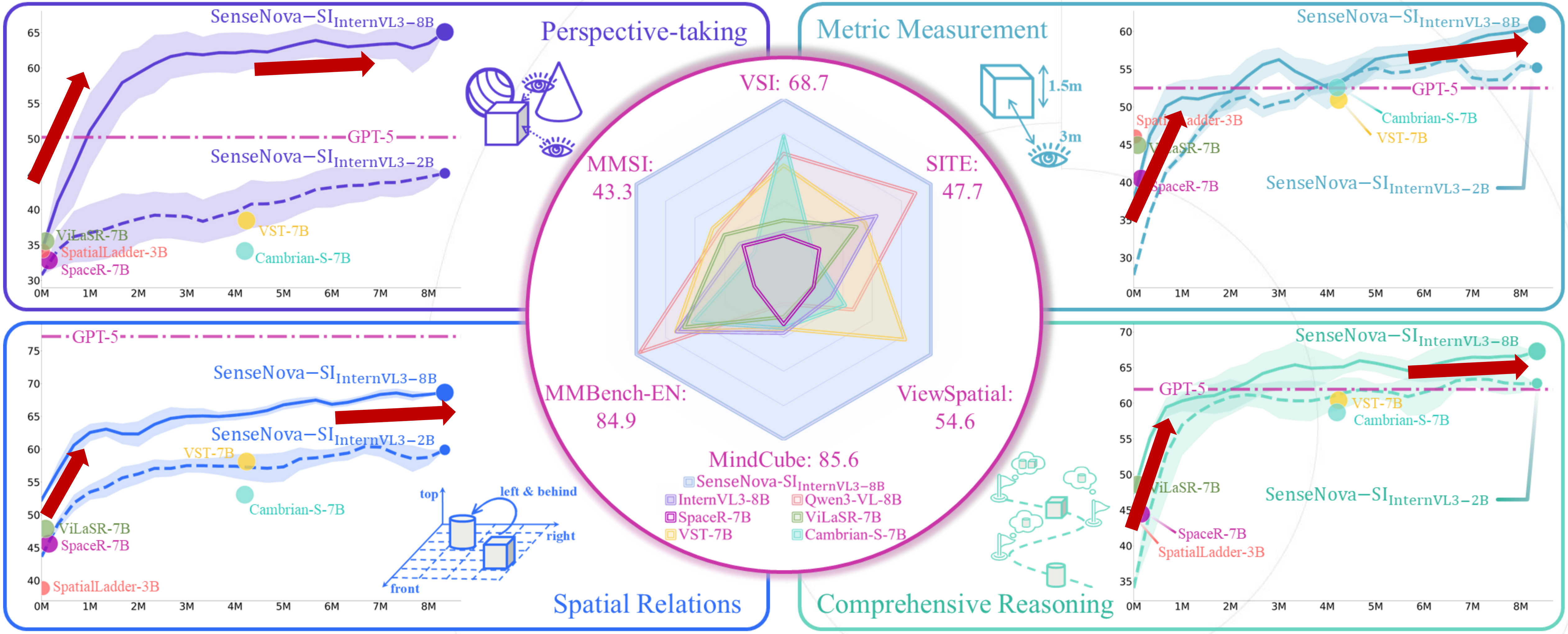
Select Columns to Display

☑ Average ☑ VSI-Bench ☑ MMSI-Bench ☑ MindCube-Tiny ☑ ViewSpatial ☑ SITE

Model	Average	VSI-Bench	MMSI-Bench	MindCube-Tiny	ViewSpatial	SITE
SenseNova-SI-1.2-InternVL3-8B	61.8	69.6	42.6	89.0	58.8	49.0
SenseNova-SI-1.1-InternVL3-8B	60.0	68.8	43.3	85.7	54.7	47.7
Gemini 3 Pro	56.2	52.5	45.2	70.9	50.4	62.2
SenseNova-SI-1.1-Qwen3-VL-8B	55.5	64.8	38.1	73.8	51.2	49.6
GPT-5	52.1	55.0	41.8	56.3	45.6	61.9
Gemini 2.5 Pro	50.5	53.6	38.0	57.6	46.1	57.1
Grok 4	47.9	47.9	37.8	63.6	43.2	47.0

EASI Leaderboard

What is Next?


Perspective-taking
$\text{SenseNova-SI}_{\text{InternVL3-8B}}$
$\text{SenseNova-SI}_{\text{InternVL3-2B}}$
GPT-5
ViLaSR-7B
SpatialLadder-3B
SpaceR-7B
VST-7B
Cambrian-S-7B
65
60
55
50
40
35
30
0M
1M
2M
3M
4M
5M
6M
7M
8M
Metric Measurement
1.5m
3m
$\text{SenseNova-SI}_{\text{InternVL3-8B}}$
$\text{SenseNova-SI}_{\text{InternVL3-2B}}$
GPT-5
SpatialLadder-3B
ViLaSR-7B
SpaceR-7B
Cambrian-S-7B
VST-7B
60
55
50
45
40
35
30
0M
1M
2M
3M
4M
5M
6M
7M
8M
VSI: 68.7
MMSI: 43.3
SITE: 47.7
MMBench-EN: 84.9
ViewSpatial: 54.6
MindCube: 85.6
□ $\text{SenseNova-SI}_{\text{InternVL3-8B}}$
□ InternVL3-8B
□ Qwen3-VL-8B
□ SpaceR-7B
□ ViLaSR-7B
□ VST-7B
□ Cambrian-S-7B
GPT-5
$\text{SenseNova-SI}_{\text{InternVL3-8B}}$
$\text{SenseNova-SI}_{\text{InternVL3-2B}}$
VST-7B
Cambrian-S-7B
ViLaSR-7B
SpaceR-7B
SpatialLadder-3B
75
70
65
60
55
50
45
40
0M
1M
2M
3M
4M
5M
6M
7M
8M
top
left & behind
right
front
Spatial Relations
$\text{SenseNova-SI}_{\text{InternVL3-8B}}$
$\text{SenseNova-SI}_{\text{InternVL3-2B}}$
GPT-5
VST-7B
Cambrian-S-7B
ViLaSR-7B
SpaceR-7B
SpatialLadder-3B
70
65
60
55
50
45
40
35
0M
1M
2M
3M
4M
5M
6M
7M
8M
Comprehensive Reasoning

Thank you!