[CVPR'25, oral] Thinking in Space How Multimodal Large Language Models See, Remember and Recall Spaces

1. Link: https://vision-x-nyu.github.io/thinking-in-space.github.io/
2. Arthurs and institution: Jihan Yang, Shusheng Yang, Anjali W. Gupta1, Rilyn Han, Li Fei-Fei, Saining Xie from NYU, Yale and Stanford.

TL;DR A benchmark with 5000 QA pairs testing the spatiotemporal understanding ability of MLLMs, finds out 1) COT is useless 2)MLLMs are good at think locally, but not globally.

Thoughts and critisims

1. Why do humans score so low on abs. dist, obj. size and room size? It doesn't quite make sense to me.
2. The authors set the confidence threshold of MRA in range(0.5, 0.95). When using MLLMs for robot planning/manipulation tasks, is a high confidence threshold necessary?

Related works

Problem formulation

Contributions

Key concepts

Taxonomy

1. the taxonomy is built on some cognitive science results

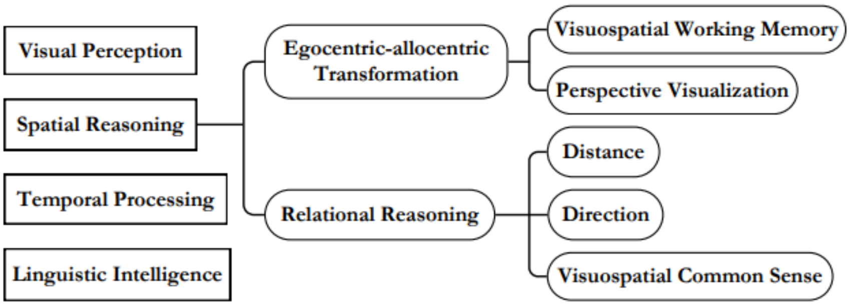
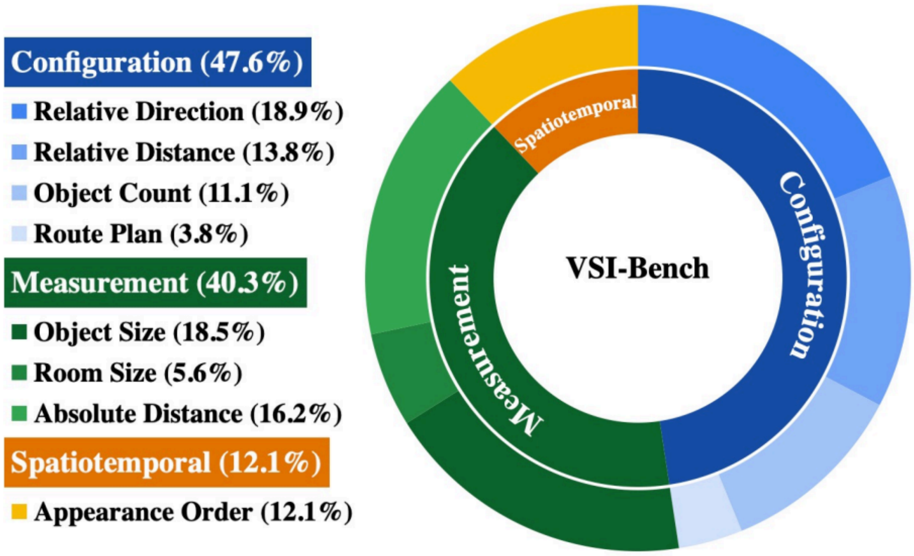
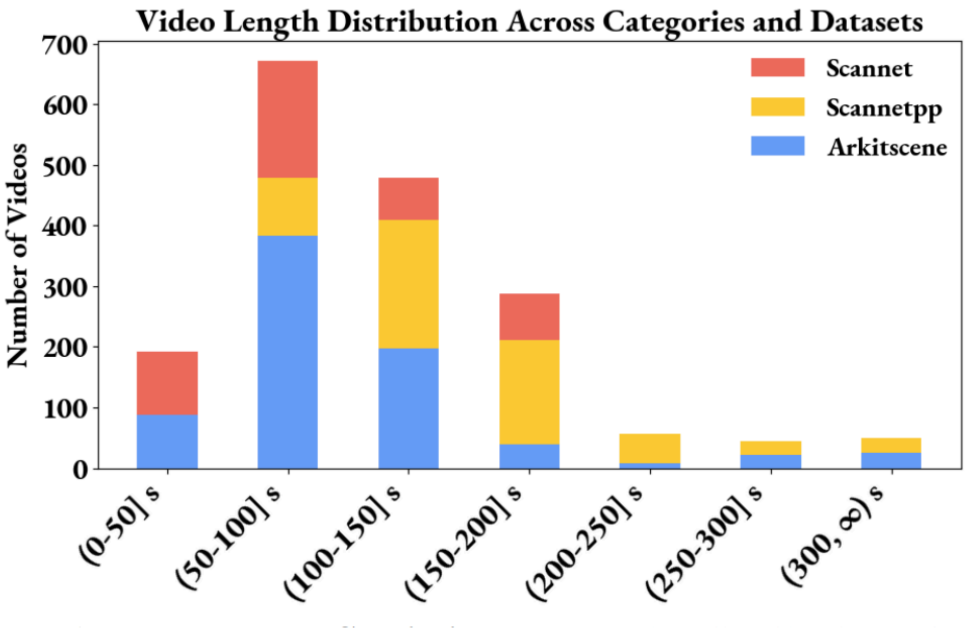


Figure 2. A taxonomy of **visual-spatial intelligence** capabilities

2. Definitions
    1. Relational reasoning: ability to identify, via distance and direction, relationships between objects
    2. Egocentric-allocentric transformation: shifting between a self-centered (egocentric) view and an environment-centered (allocentric) one.

Benchmark





Statics

1. 5,000 questionanswer pairs derived from 288 real videos
    1. ScanNet(++)/ARKitScenes
2. Tasks
    1. configurational: test a model’s understanding of the configuration of a space and are more intuitive for humans
        1. obj count
        2. relative distance
        3. relative direction
        4. route plan
    2. measurement estimation: of value to any embodied agent
        1. object size
        2. room size
        3. abs distance

        3. spatiotemporal: test a model's memory of a space as seen in video
            1. appearance order
3. Answer type
    1. MCA
    2. numerical

Construction

1. Data Collection and Unification
    1. route plan is human-annotated
    2. the others are generated based on previous labels.
    3. human verification is implemented at all key stages for filtering low-quality videos, annotations, and ambiguous QA pairs

Metric

1. MCA: Accuracy
2. Numerical: Mean Relative Accuracy

$$\mathcal{MRA} = \frac{1}{10} \sum_{\theta \in \mathcal{C}} \mathbb{1} \left(\frac{|\hat{y} - y|}{y} < 1 - \theta \right).$$

Results

Methods	Rank	Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
			Numerical Answer				Multiple-Choice Answer			
Baseline										
Chance Level (Random)	-	-	-	-	-	-	25.0	36.1	28.3	25.0
Chance Level (Frequency)	-	34.0	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
`VSI-Bench` *(tiny) Perf.*										
†Human Level	-	79.2	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0
†Gemini-1.5 Flash	-	45.7	50.8	33.6	56.5	45.2	48.0	39.8	32.7	59.2
†Gemini-1.5 Pro	-	48.8	49.6	28.8	58.6	49.4	46.0	48.1	42.0	68.0
†Gemini-2.0 Flash	-	45.4	52.4	30.6	66.7	31.8	56.0	46.3	24.5	55.1
Proprietary Models (API)										
GPT-4o	3	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash	2	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	1	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-source Models										
InternVL2-2B	11	27.4	21.8	24.9	22.0	35.0	33.8	44.2	30.5	7.1
InternVL2-8B	5	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B	3	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
LongVILA-8B	12	21.6	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5
VILA-1.5-8B	9	28.9	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8
VILA-1.5-40B	7	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9
LongVA-7B	8	29.2	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7
LLaVA-NeXT-Video-7B	4	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B	1	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
LLaVA-OneVision-0.5B	10	28.0	46.1	28.4	15.4	28.3	28.9	36.9	34.5	5.8
LLaVA-OneVision-7B	6	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B	2	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6

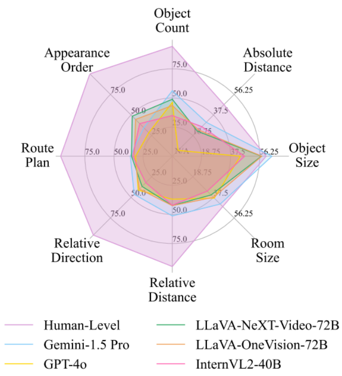


Table 1. **Evaluation on `VSI-Bench`. Left:** Dark gray indicates the best result among all models and light gray indicates the best result among open-source models. † indicates results on `VSI-Bench` (tiny) set. **Right:** Results including the top-3 open-source models.

Finds

How MLLMs Think in Space Linguistically

1. probing via self-explanation

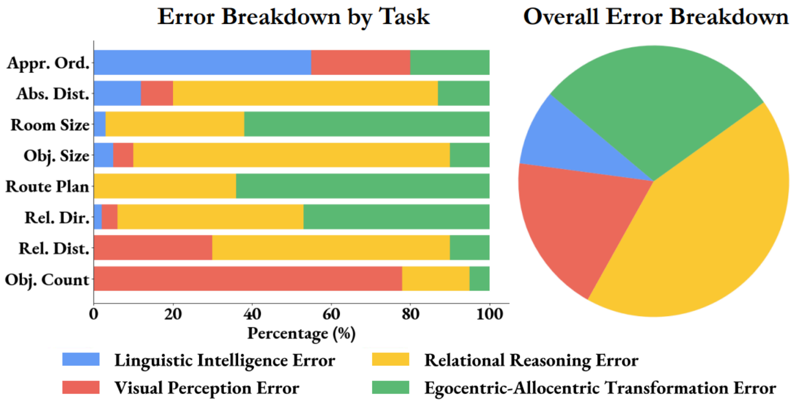
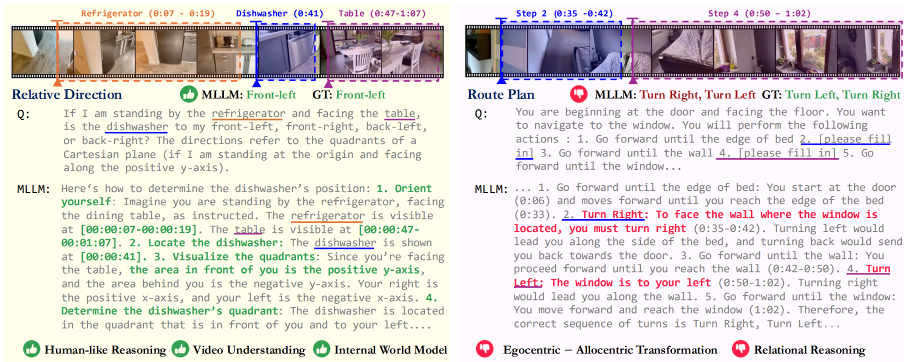
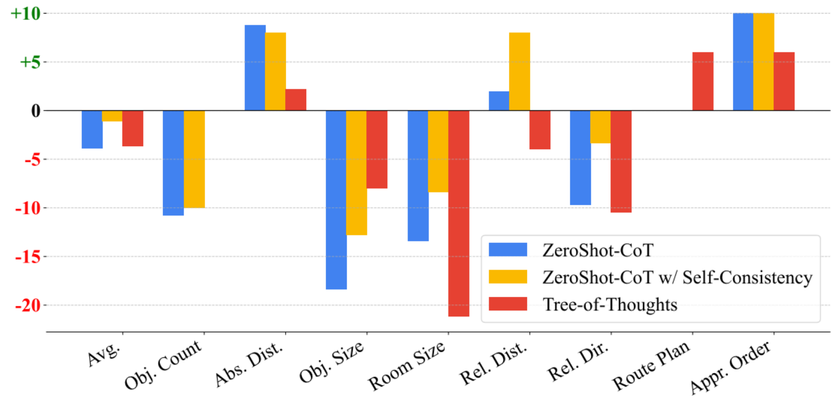


Figure 8. Human-conducted analysis of errors by type. Over 70% of errors stem from faulty spatial reasoning capabilities.

    1.
    2. good at human-like reasoning, video understanding and internal world model
    3. bad at ego-allocentric transformation and relational reasoning
    4. 

2. errors
    1. Visual perception error: stemming from unrecognized objects or misclassified object categories;
    2. Linguistic intelligence error: caused by logical, mathematical reasoning, or language understanding defects
    3. Relational reasoning error: errors in spatial relationship reasoning, i.e., distance, direction, and size
    4. Egocentric-allocentric transformation error: resulting from an incorrect allocentric spatial layout or improper perspective-taking
3. conclusion
    1. Spatial reasoning is the primary bottleneck for MLLM performance on VSI-Bench
    2. Linguistic prompting techniques, although effective in language reasoning and general visual tasks, are harmful for spatial

reasoning



How MLLMs Think in Space Visually

1. ask MLLMs the position of objects in a 10 x 10 grid

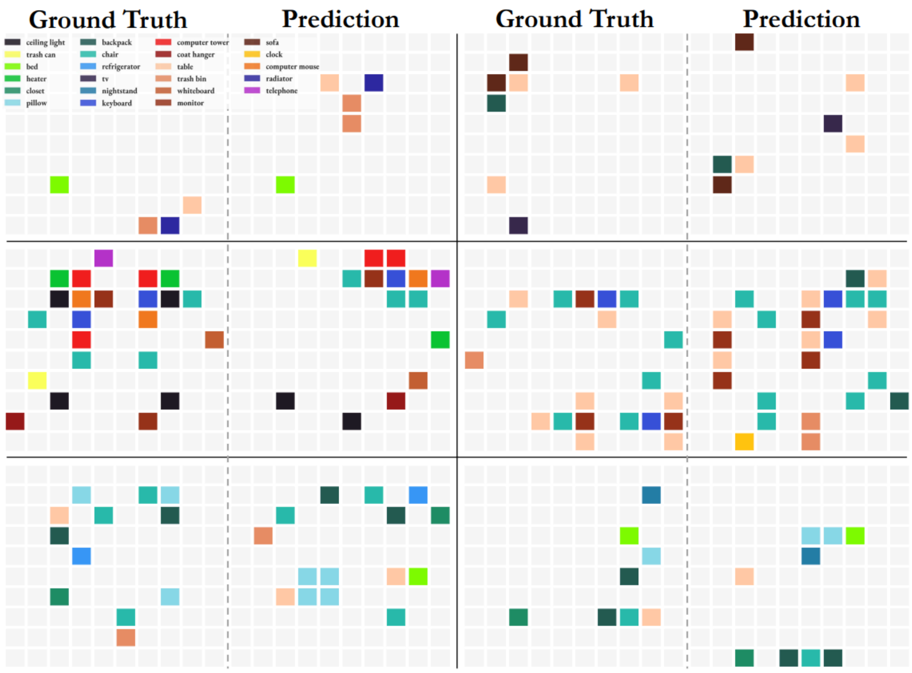


Figure 10. Visualization of cognitive maps from MLLM and GT.

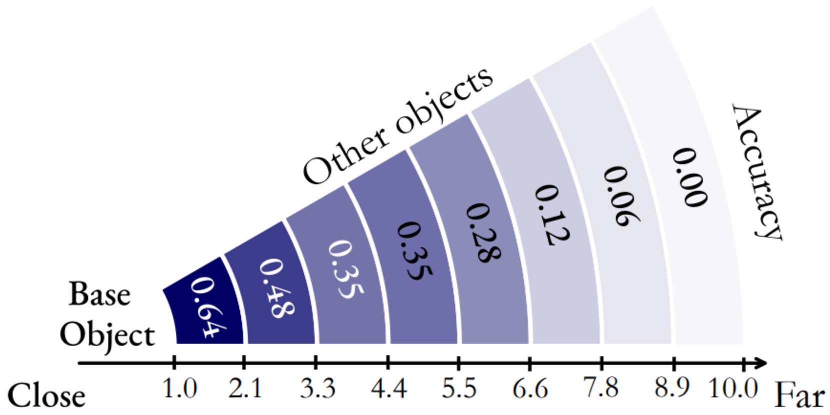


Figure 11. **Locality of the MLLM's predicted cognitive maps.** The MLLM's map-distance accuracy decreases dramatically with increasing object distance.

2. conclusion: When remembering spaces, a MLLM forms a series of local world models in its mind from a given video, rather than a unified global model.
3. MLLMs do better jobs if ask them to generate a congitive map first.