SpatialTree: How Spatial Intelligence Branches Out in MLLMs

Yuxi Xiao[12*], Longfei Li[13*], Shen Yan[1], Xinhang Liu[4], Sida Peng[2], Yunchao Wei[3], Xiaowei Zhou[2], Bingyi Kang[1]

Bytedance Seed, Zhejiang University, Beijing Jiaotong University, HKUST

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

*What does **"spatial intelligence"** even mean for MLLMs?*

It spans low-level perception to embodied action — today, studied in scattered, task-specific benchmarks.

1 PRIOR WORK: ORGANIZED BY TASK

- **Low-level Perception** — *BLINK · 3DSR-Bench*
- **Spatial Reasoning** — *VSI-Bench · MMSI*
- **Embodied Agent** — *EmbodiedBench*

Benchmarks grow task-by-task — abilities tested in isolation, overlapping, with no shared structure.

2 THE REAL QUESTION: HOW DO ABILITIES RELATE?

- **DEPENDENCY** — *Independent, or compositional?*
- **TRANSFER** — *Do low-level skills lift higher ones?*
- **PREREQUISITE** — *Which abilities underpin others?*

Task-centric benchmarks score skills in isolation — blind to how abilities interrelate and transfer.

3 OUR SHIFT: A CAPABILITY-CENTRIC HIERARCHY

"Each stage builds on the stage before it." — Jean Piaget, cognitive development

Capability-first, we recast it as SpatialTree — a hierarchical taxonomy:

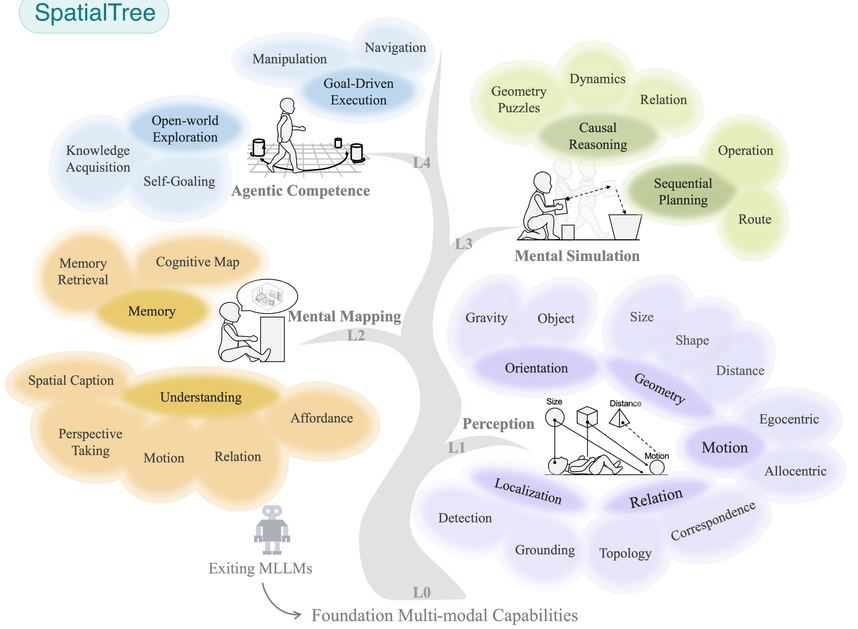
- **L1 Perception** — *Native perception of space.*
- **L2 Mental Mapping** — *Mapping spatial concepts to language.*
- **L3 Mental Simulation** — *Thinking about space in the mind.*
- **L4 Agentic Competence** — *Executing actions in space.*

27 sub-abilities across 4 levels · full framework in the next column ⟶

This work then studies all three: structure · transfer · training

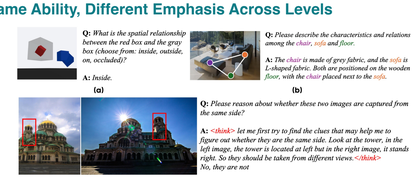
Scan the code, or visit **spatialtree.github.io** · *Paper · Code · Leaderboard*

SpatialTree



Four-layer hierarchy rooted in foundational multi-modal capabilities (L0) — branching from Basic Perception (L1) to Agentic Competence (L4).

Same Ability, Different Emphasis Across Levels



(a) *Q: What is the spatial relationship between the red box and the gray box (choose from: inside, outside, on, occluded)?*
A: Inside.

(b) *Q: Please describe the characteristics and relations among the chair, sofa and floor.*
A: The chair is made of grey fabric, and the sofa is L-shaped fabric. Both are positioned on the wooden floor, with the chair placed next to the sofa.

(c) *Q: Please reason about whether these two images are captured from the same side?*
A: <think> let me first try to find the clues that may help me to figure out whether they are the same side. Look at the tower, in the left image, the tower is located at left but in the right image, it stands right. So they should be taken from different views. </think> No, they are not

The same ability "Relation" shifts focus across levels:

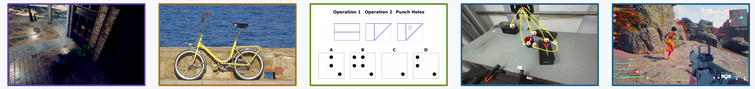
(a) L1 Perception
basic spatial relations (inside / outside)

(b) L2 Understanding
object attributes & mutual relationships

(c) L3 Causal Reasoning
visual cues + logical inference for complex tasks

Sample Cases Across the Hierarchy

Concrete cases our data engine builds — L1 perception to L4 agentic action



- **L1 · Orientation** — *Estimate camera roll & pitch vs. gravity*
- **L2 · Affordance** — *Where to hold, sit & pedal to ride?*
- **L3 · Geometry Puzzle** — *Two folds + a punch — unfold to? (A–D)*
- **L4 · Manipulation** — *Pick the next gripper action*
- **L4 · 3D-Game Nav** — *Plan the moves to reach the target view*

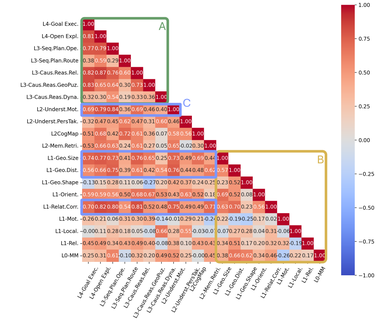
Benchmark

Methods	Rank	Avg.	L1 Perception Geom. [0.40]	Motion [0.15]	Rel. [0.15]	Local. [0.20]	Orient. [0.10]	L2 Mental Mapping Underst. [0.70]	Memory [0.30]	L3 Mental Simulation Caus. Reas. [0.65]	Seq. Plan. [0.35]	L4 Agentic Competence Goal Exec. [0.50]	Open Expl. [0.50]
Non-Thinking Models													
GPT-4o	13	31.9	23.9	38.6	29.8	24.2	36.2	31.2	43.6	29.3	40.5	25.8	39.2
Gemini2.5 Flash NT	10	35.8	31.6	29.3	30.8	35.2	45.4	36.4	53.7	28.4	36.9	27.6	45.7
Gemini2.5 Pro NT	5	41.4	36.2	30.0	33.2	47.0	48.5	43.3	55.2	39.6	47.5	29.2	46.0
Thinking Models													
Seed1.5vl	6	41.3	39.2	36.6	24.6	44.7	44.8	46.5	38.5	36.6	47.0	28.4	25.9
Seed1.8	3	50.3	42.5	42.7	49.9	54.5	52.4	56.3	51.0	49.7	53.6	26.0	70.6
GLM-4.5V	9	36.0	35.3	24.0	32.5	34.5	43.7	34.5	34.1	33.8	41.0	26.8	49.7
Gemini2.5-Pro	4	50.1	47.8	32.6	44.4	61.6	47.9	50.5	61.5	47.6	58.2	28.3	63.3
Gemini2.5-Flash	7	41.1	35.6	29.3	40.5	56.2	44.6	43.1	51.4	36.4	49.1	26.9	46.6
Gemini3-Pro	2	56.5	54.5	47.4	62.5	74.0	53.9	60.5	50.3	56.0	68.2	29.9	68.5
Gemini3-Flash	1	57.8	50.1	40.7	62.9	62.4	54.6	62.0	66.8	58.4	68.6	31.6	70.8
Open-source Models													
Qwen2.5VL-7B	17	27.5	17.8	22.6	23.9	20.6	31.6	34.7	15.2	28.4	39.8	24.5	31.1
Qwen2.5VL-32B	16	27.9	21.4	30.0	25.8	22.0	35.1	29.0	21.7	32.8	37.0	14.1	38.6
Qwen2.5VL-72B	12	33.0	24.4	22.0	28.6	37.5	37.3	38.9	35.1	32.6	38.4	23.9	38.6
Qwen3VL-30B	11	35.3	31.9	26.0	32.3	20.7	39.2	37.8	48.1	32.7	44.2	25.8	40.9
Qwen3VL-235B	8	40.0	33.9	27.4	35.1	35.4	38.9	43.7	53.4	37.3	44.7	28.8	48.9
Kimi-VL-A3B	19	24.4	13.8	23.3	31.6	27.0	21.4	24.9	28.3	26.9	27.8	15.7	32.6
Specialized SI Model													
SenseNova-base-Qwen3-VL-8B	14	30.2	25.5	28.6	29.9	21.5	35.6	35.3	45.2	22.9	35.3	25.0	32.6
VST-base-Qwen25-VL-7B	18	26.2	20.2	31.2	29.6	22.1	28.9	26.3	24.8	24.8	31.2	25.4	31.6
SpaceR-base-Qwen25-VL-7B	15	28.1	15.7	25.4	24.8	25.1	35.0	32.0	15.2	26.8	33.4	31.4	38.8

SpatialTree-Bench. Dark gray indicates the best result among all models and light gray indicates the best result among open-source models. NT denotes the non-thinking model. **Avg.** is computed using the weights in brackets [·].

4 Levels (L1–L4)
27 Sub-abilities
19 MLLMs evaluated
9 Metrics

Analysis



Correlation Findings :

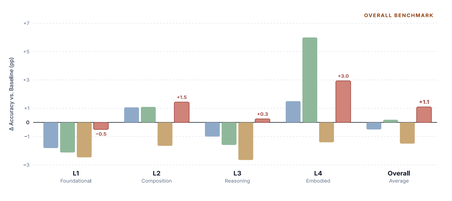
A. L3/L4 higher-level capabilities — tightly coupled

B. L1 foundational — largely independent.

C. Salient L1 abilities that drive higher-level tasks.

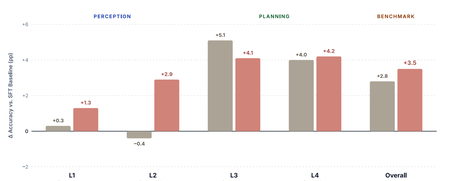
Correlation → causation. If low-level abilities underpin higher levels (B, C), training them should transfer *upward* — verified by SFT (Findings 1–2) and scaled by RL (Finding 3). ⟶

Ability Transfer Probe



Finding 1: Single-ability L1 SFT induces cross-level transfer, while yielding limited or slightly negative effects on same-level abilities.

Finding 2: The holistic integration across multiple fundamental abilities achieves synergistic gains far exceeding their individual effects.



Finding 3: Auto-think — suppress reasoning for perception, amplify it for planning — unlocks consistent gains across the hierarchy.