ESI-Bench: Towards Embodied Spatial Intelligence that Closes the Perception-Action Loop

Yining Hong[1] **Jiageng Liu**[2] **Han Yin**[1] **Manling Li**[3] **Leonidas Guibas**[1] **Fei-Fei Li**[1] **Jiajun Wu**[1] **Yejin Choi**[1]
[1]Stanford University [2]UCLA [3]Northwestern University

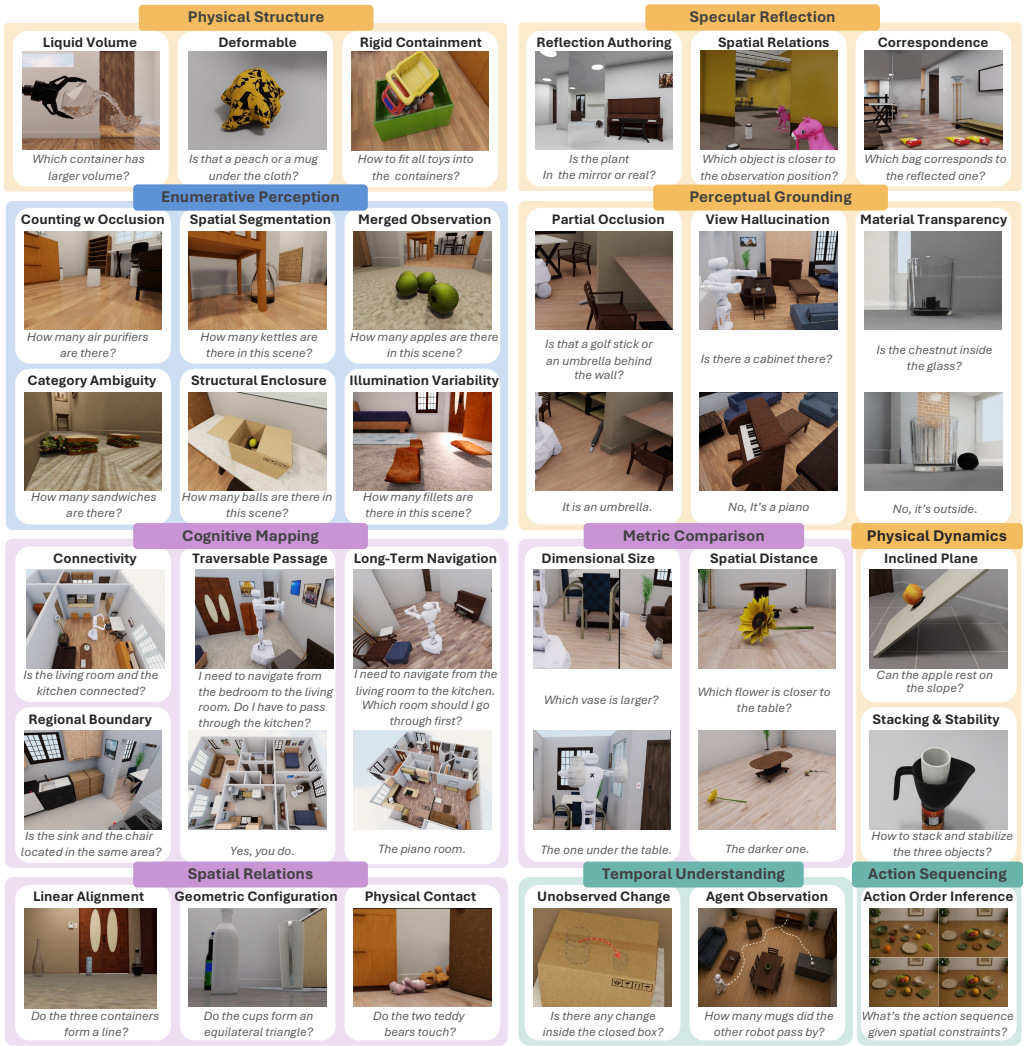


Figure 1: **ESI-Bench** is a comprehensive benchmark for embodied spatial intelligence, spanning 10 task categories and 30 subcategories organized around Spelke's four core knowledge systems [Spelke and Kinzler, 2007]: object persistance, layout and geometry, number representation, and agents and goal-directed actions.

Abstract

Spatial intelligence unfolds through a perception–action loop: agents act to acquire observations, and reason about how observations vary as a function of action. Rather than passively processing what is seen, they actively uncover what is unseen—occlusion, dynamics, containment, and functionality—beyond the reach

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

of passive sensing. We take a step beyond prior formulations of spatial intelligence, which often emphasize passive perception or assume access to oracle observations, by recasting the observer as an actor. We introduce ESI-BENCH, a comprehensive benchmark for embodied spatial intelligence spanning 10 task categories and 30 subcategories built on OmniGibson, grounded in Spelke's core knowledge systems. Agents must decide what abilities to deploy — perception, locomotion, and manipulation — and how to act to answer questions that cannot be resolved from passive observation alone. We conduct extensive experiments on state-of-the-art MLLMs and find that active exploration substantially outperforms passive counterparts, with agents spontaneously discovering emergent spatial strategies without explicit instruction, while passive multi-view adds noise rather than signal despite consuming far more images. Most failures stem not from weak perception but from action blindness, and their coupling drives cascading failures where bad actions produce bad views which produce worse actions. While explicit 3D grounding stabilizes reasoning on depth-sensitive tasks, imperfect reconstruction proves more harmful than 2D baselines by actively distorting spatial relations. Human studies further reveal that unlike humans who seek falsifying viewpoints and revise beliefs under contradiction, models commit prematurely with high confidence regardless of evidence quality, exposing a metacognitive gap that neither better perception nor more embodied interaction alone can close. All data, codes are publicly available at our project page `https://esi-bench.github.io/`.

1 Introduction

Perception is often characterized in cognitive science as *perceptually guided action* [Varela et al., 1991, Gibson, 1979]: a perception–action loop where knowing how observations change as a function of action [O'Regan and Noë, 2001], and which actions elicit effective and informative sensing, are often more challenging than sensing itself. This is especially critical in spatial intelligence, which concerns not only what is *seen*, but also what is *unseen*: latent physical properties such as occlusion, dynamics, containment, and functionality that are inaccessible to passive sensing and must be actively unveiled through interaction.

selective or prioritized sensing? sentences not parsable: perception-action loop where; also which predominantly treat make (1) and (2) more different from passive sensing to proactive sensing mirage -> hallucination. maybe not incomplete. ambiguity teaser refer to figure epistemic calibration -> spell it out. explain more We take a step beyond prior formulations of spatial intelligence Liu et al. [2023], Yang et al. [2024, 2025d], which predominantly treat perception as passive or assume access to oracle observations, by recasting the observer as an actor. Our work contrasts with prior works in three key ways: (1) *from spatial sensing to spatial competence*, where agents are evaluated not only on what they can perceive, but on whether they know how to act to perceive it; (2) *selective sensing*, where agents must determine which observations are worth acquiring, prioritizing task-relevant information over redundant or uninformative inputs; and (3) *resolving perceptual mirages*, where agents must reason through incomplete or misleading observations to infer the hidden spatial structures and underlying physical constraints beyond what is directly observed.

We introduce `ESI-BENCH`, the first comprehensive benchmark for embodied spatial intelligence spanning 10 task categories and 30 subcategories, to address the critical perception-action gap in existing benchmarks by focusing on questions unanswerable from passive observation. Our category design follows Spelke's core knowledge systems [Spelke and Kinzler, 2007], which identify four faculties of spatial intelligence: object persistence, layout and geometry, number representation, and agents and intentional action. Building on these theoretical foundations, we conduct human surveys to identify the most challenging spatial tasks that require embodied interaction and manipulation within each faculty,narrowing these into a structured taxonomy spanning diverse forms of spatial reasoning, as illustrated in the teaser across all 10 categories and 30 subcategories. These tasks only become meaningful when an agent has a body, a belief state, and physical stakes in the outcome: the agent must determine what abilities to deploy (perception, locomotion, manipulation), which actions to take (where to move, what to probe, how to manipulate), and how to execute them in the right order.

We conduct extensive experiments on state-of-the-art MLLMs across three paradigms: passive single-view, passive multi-view, and active exploration, alongside a ground-truth oracle that separates perception errors from action errors. Our experiments reveal multiple key insights: (1) *active exploration unlocks emergent spatial strategies*: without explicit instruction, active agents spontaneously discover diverse action compositions, driving substantial gains over passive counterparts while passive multi-view, despite consuming far more images, adds noise rather than signal; (2) *action blindness dominates perceptual blindness*: for most tasks the bottleneck is not perception but action selection, as models given oracle viewpoints succeed dramatically, yet certain tasks expose a hard perceptual ceiling where even perfect viewpoints cannot overcome the fundamental limits of 2D recognition; and (3) *failures cascade and compound*: suboptimal actions produce uninformative views, which trigger worse subsequent actions, creating a compounding chain of errors that cannot be recovered within the step budget.

We further investigate whether explicit 3D representations can help. We find that while explicit 3D grounding stabilizes reasoning on depth-sensitive tasks by recovering information that 2D projections fundamentally lose, imperfect reconstructions prove more harmful than 2D baselines, as geometric artifacts actively distort fine-grained spatial relations and mislead downstream reasoning. Finally, human studies expose a critical gap in epistemic calibration. We observe that unlike humans who actively seek falsifying viewpoints, explore orthogonal angles, and revise their beliefs when contradicted, models commit prematurely with uniformly high confidence regardless of evidence quality, anchoring to first impressions and ignoring contradictory observations, a metacognitive failure that neither better perception nor more embodied interaction alone can close.

2 Related Work

Spatial Reasoning Methods in MLLMs. Recent advances in multimodal large language models have improved spatial understanding, yet most methods still focus on interpreting a fixed observation set rather than choosing which observations to acquire. One line of work injects geometric priors into otherwise 2D backbones: SpatialVLM Chen et al. [2024] constructs training data with synthesized 3D spatial annotations and metric-depth supervision; SpatialBot Cai et al. [2024] leverages RGB-D inputs and depth-oriented question answering; SpatialRGPT Cheng et al. [2024] learns region-level representations from 3D scene graphs with a depth plugin; and newer geometry-grounded models such as Spatial-MLLM Wu et al. [2025a] and VLM-3R Fan et al. [2026] further strengthen 3D reasoning by combining visual geometry priors with monocular video or reconstructive instruction tuning. A second line of work treats spatial inference as an explicit reasoning process: SpatialCoT Liu et al. [2025] aligns vision–language inputs with spatial coordinates for chain-of-thought spatial grounding, while VILASR Wu et al. [2025b] interleaves textual reasoning with visual drawing operations. Together, these methods improve how models reason over observed content, but the observation set remains fixed as input. ESI-Bench instead asks whether a spatial reasoner can actively choose the observations its perception modules need.

Benchmarks for Spatial Reasoning. Evaluation of spatial intelligence has scaled with model capability, yet most benchmarks still operate under fixed observation settings. Early work such as VSR Liu et al. [2023] formulates spatial reasoning as classification over a predefined set of spatial relations from single images, while BLINK Fu et al. [2024] and 3DSRBench Ma et al. [2025] extend evaluation to core visual perception and fine-grained 3D spatial reasoning. SpatialScore Wu et al. [2026a] further consolidates VGBench and multiple datasets into a unified benchmark with a tool-augmented evaluation agent. More recent benchmarks expand the observation space to multi-view and temporal regimes, including egocentric video understanding (e.g., VSI-Bench Yang et al. [2025a]), multi-image spatial consistency (e.g., MMSI-Bench Yang et al. [2025d]), and spatial mental modeling from partial observations (e.g., MindCube Wang et al. [2026]). Cambrian-S Yang et al. [2025c] pushes toward long-horizon settings with VSI-SUPER, targeting visual spatial recall and continual spatial counting over extended video. In parallel, benchmarks such as PhysBench Chow et al. [2025] and CausalSpatial Ma et al. [2026] probe latent physical structure and object-centric dynamics from passive multimodal inputs. Across this progression, inputs become increasingly rich—multiple images, videos, partial coverage, and temporal dynamics—yet the observation process remains largely fixed, with limited support for agent-driven view selection. As a result, these benchmarks provide limited signals for disentangling perceptual limitations from failures in active information

acquisition. ESI-Bench retains the diagnostic spirit of this line of work but treats the observer as an active agent, making the utility of each observation contingent on the model's decisions.

Embodied Evaluation and Active Perception. A separate line of work places models inside simulators and evaluates them as embodied agents. EmbodiedBench Yang et al. [2025b] and EmbodiedEval Cheng et al. [2025] measure end-to-end task success across diverse navigation, interaction, and QA scenarios, while OpenEQA Majumdar et al. [2024]—spanning both episodic-memory (EM-EQA) and active (A-EQA) settings—and the more recent EXPRESS-Bench Jiang et al. [2025] formalize embodied question answering with explicit attention to exploration quality, the latter introducing an Exploration–Answer Consistency (EAC) metric to discourage disembodied reasoning. Spatial cognition under embodiment has been examined more directly by EmbSpatial-Bench Du et al. [2024] and ESPIRE Zhao et al. [2026], which provide diagnostic assessments of egocentric spatial reasoning, with ESPIRE further decomposing each task into localization and execution to enable fine-grained error analysis. On the side of active perception itself, several recent efforts develop methods that adaptively select what to observe: Vision in Action Xiong et al. [2025] learns active observation strategies from human demonstrations; Thinking in 360° Yu et al. [2025] studies humanoid head-rotation visual search over panoramic environments; and SaPaVe Liu et al. [2026a] and ActiveVLA Liu et al. [2026b] integrate viewpoint selection into vision–language–action policies for manipulation. The closest concurrent benchmark is CHAIN Wu et al. [2026b], which evaluates closed-loop physical reasoning over interlocking mechanical puzzles and 3D stacking and packing in a physics engine; CHAIN and ESI-Bench are complementary in scope—CHAIN focuses on tabletop physical reasoning, while ESI-Bench targets the broader range of spatial faculties an embodied agent must exercise across object, geometry, number, and agent reasoning, including latent properties such as containment, occlusion, transparency, reflection, and unobserved scene change that remain only partially covered by current embodied evaluation. Table 1 situates ESI-Bench within this landscape by combining active embodiment with hidden-state probing for spatial reasoning.

Table 1: Comparison of spatial reasoning and embodied benchmarks. **Action:** L = locomotion, P = active perception, M = manipulation. **Core Knowledge**: Obj = object, Geom = geometry, Num = number, Phys = physics, Agt = agent. Hidden-State Probing denotes whether the benchmark targets latent properties (e.g., occlusion, containment, dynamics) inaccessible to passive observation.

Benchmark	Active Obs.	Action	Hidden-State	Core Knowledge
VSR Liu et al. [2023]	✗	–	✗	Geom
BLINK Fu et al. [2024]	✗	–	✗	Geom
3DSRBench Ma et al. [2025]	✗	–	✗	Obj, Geom
VSI-Bench Yang et al. [2025a]	✗	–	✗	Obj, Geom, Num
MMSI-Bench Yang et al. [2025d]	✗	–	✗	Obj, Geom
MindCube Wang et al. [2026]	✗	–	✓	Geom, Phys
PhysBench Chow et al. [2025]	✗	–	✓	Obj, Geom, Phys
CausalSpatial Ma et al. [2026]	✗	–	✓	Obj, Phys
EmbSpatial-Bench Du et al. [2024]	✗	–	✗	Geom
OpenEQA Majumdar et al. [2024]	✓	L	✗	Obj, Geom
EmbodiedBench Yang et al. [2025b]	✓	L, M	✗	Obj, Geom, Agt
EmbodiedEval Cheng et al. [2025]	✓	L, M	✗	Obj, Geom, Agt
EXPRESS-Bench Jiang et al. [2025]	✓	L	✗	Geom, Agt
ESPIRE Zhao et al. [2026]	✓	L, M	✗	Obj, Geom, Agt
CHAIN Wu et al. [2026b]	✓	M	✓	Obj, Geom, Phys
ESI-Bench (ours)	✓	**L, P, M**	✓	**Obj, Geom, Num, Phys, Agt**

3 ESI-BENCH

3.1 Benchmark Setup

Task Definition Each task in ESI-Bench is defined by a tuple $(\mathcal{S}, p_0, q, y^*)$, where $\mathcal{S}$ is a 3D scene instantiated from the BEHAVIOR-1K scene pool with a fixed set of pre-loaded objects, p_0 is the

initial pose of the agent within the scene, q is a natural language question about a spatial property of the scene, and y^* is the ground-truth answer. The environment is formalized as $\mathcal{E} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T \rangle$, where $\mathcal{A}$ is the action space, $\mathcal{O}$ is the egocentric visual observation space, and $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition function governing scene evolution. Given $(\mathcal{S}, p_0, q)$, the agent receives observation $o_t \in \mathcal{O}$ at each timestep t, issues action $a_t \in \mathcal{A}$, and transitions to a new scene state via T, producing a trajectory $\tau = (o_0, a_0, o_1, a_1, \ldots)$ until it commits to a final answer $\hat{y}$. The agent produces $\hat{y}$ as free-form text; while the answer space is open, the phrasing of q implicitly suggests the expected format: binary yes/no for relational tasks, a category name for comparative tasks, an integer count for enumerative tasks, an ordering for procedural tasks, and so on. A response is correct if $\hat{y} = y^*$.

Simulation Environment. We build ESI-BENCH on BEHAVIOR-1K within the OmniGibson simulator. BEHAVIOR-1K provides 51 fully interactive 3D scenes spanning residential, commercial, and institutional environments (houses, gardens, offices, restaurants, schools, and stores), totaling over 300 rooms across all scenes, with over 9,000 object instances across 1,829 categories, each annotated with rich physical properties including friction, mass, and articulation. OmniGibson is built on top of NVIDIA's Isaac Sim and PhysX 5, providing the broad simulation capabilities essential for embodied spatial evaluation: rigid-body contact physics, particle-based fluid simulation, transparency rendering, realistic lighting and reflections, and extended object states including fill levels, and toggled states. For each task instance (described in the next paragraph), we randomly sample a scene from the BEHAVIOR-1K scene pool and select rooms based on room type and task category requirements, determined from a combined room-object list. We load the selected room into OmniGibson, allow the physics simulation to settle, and query the simulator state to extract a structured scene graph containing per-object bounding boxes, categories, spatial relationships, room assignments, and object states such as fillable capacity, toggled state,and contact flags. This scene graph serves as the foundation for scenario construction: it provides the object inventory from which task-relevant objects are selected, the spatial layout from which camera positions are computed, and the ground-truth geometric state and object states from which labels are automatically derived.

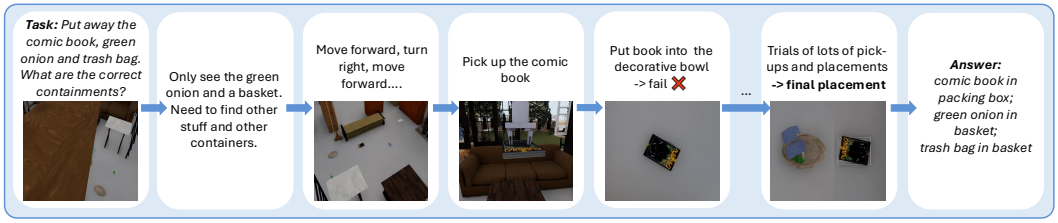
Action Space. Active agents operate over a unified action vocabulary spanning perception, locomotion and manipulation, summarized in Table **??**. We impose a maximum step budget of $T_{\max} = 30$.

3.2 Task Construction

Task Proposal. GPT-4o is prompted with the scene graph alongside task category requirements to select task-relevant objects from a random sample of 200 candidate categories drawn from the full BEHAVIOR-1K inventory, applying task-specific physical criteria to choose the most appropriate categories and resolve a specific model instance per category. Beyond object selection, GPT-4o also determines the initial positions of both the objects and the agent within the scene, and generates a ground-truth action trajectory providing the optimal sequence of actions needed to resolve the task. The selected objects and their spatial configuration implicitly define the task, with the ground-truth answer y^* derived directly from the resulting scene state.

Scene Instantiation. Taking the GPT-4o-proposed object selections and initial positions as input, selected objects are loaded into the scene at data generation time. Before placement, a conflict check is performed against existing scene objects via bounding box intersection tests to ensure no overlap with pre-existing scene content. Objects are then placed on supporting surfaces via physics-based kinematic sampling and allowed to settle under simulation for a fixed number of steps. After settling, the configuration is checked through a battery of tests: placement stability via bounding box re-querying, per-view object existence verification via segmentation masks, and contact flag validation where applicable. Configurations that fail any check are rejected.

Agent Trajectory Collection. The agent is initialized at the GPT-4o-proposed pose, determined by a combination of randomization and pre-defined placement principles designed to intentionally withhold the full scene configuratiosn and properties from the agent. At initialization, the same battery of checks applied during scene instantiation is performed: per-view object existence verification via segmentation masks, bounding box re-querying, and contact flag validation where applicable. The proposed actions are then executed step by step in the environment, with each step rendered to produce egocentric visual observations and verified through the same battery of checks. Trajectories where any step fails verification are discarded.



Action	Description
Locomotion	
`move_forward` / `move_backward`	Translate agent along viewing axis
`move_left` / `move_right`	Translate agent laterally
`move_up` / `move_down`	Translate agent vertically
Perception	
`turn_left` / `turn_right`	Rotate agent horizontally
`turn_up` / `turn_down`	Rotate agent vertically
Manipulation	
`pick_up obj`	Pick up object
`put obj inside obj`	Put object inside object
`put obj on obj`	Put object on object
`fill obj with water`	Fill object with water
`pour obj from obj to obj`	Pour object from object to object
Terminal	
`answer(`$\hat{y}$`,` c`)`	Commit to final answer with confidence c

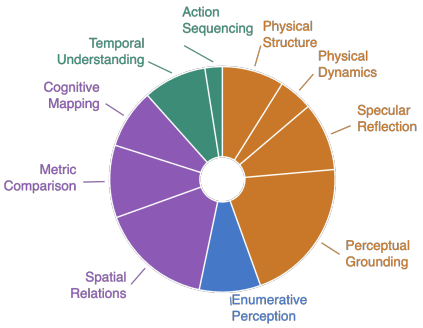


Figure 2: **Top:** Dataset example. **Bottom-left:** ESI-BENCH agent action space. **Bottom-right:** Distribution of task categories.

Metadata Saving. Upon completion of agent trajectory collection, all task instance information is saved to a structured JSON file. This includes scene name, room name, floor name, per-object category, model instance, verified initial position and quaternion, agent initial pose and quat p_0, per-view object existence flags derived from segmentation masks, question, and ground-truth answer and action trajectory. This metadata file forms a self-contained and fully reproducible record of the task instance.

Human Verification. All generated task instances are examined by human annotators using the rendered per-step observations alongside the metadata. Annotators verify along three axes: correctness, ensuring the initial position and whole trajectory is physically valid; answerability, ensuring the task is solvable through interaction and the spatial configuration is unambiguous; and non-triviality, ensuring the task cannot be resolved through visual bias or prior knowledge without interaction and presents genuine spatial uncertainty. Task instances that fail any criterion are discarded.

3.3 Task Categories and Statistics

ESI-Bench comprises N total scenarios across 10 categories and 30 subcategories, summarized in Table 2. Figure **??** reports the task contribution. Each category targets a distinct spatial faculty structurally inaccessible to passive sensing. Physical Capacity and Physical Dynamics tasks require manipulation to reveal hidden physical properties such as containment capacity and gravitational stability. Reflective Reasoning and Perceptual Grounding tasks require active repositioning to disambiguate viewpoint-dependent phenomena where the correct answer changes critically with the observer's position. Metric Comparison and Enumerative Perception tasks require locomotion and manipulation to overcome forced-perspective distortions where objects appear merged or equal in size from any fixed view. Spatial Relations tasks require navigation to vantage points that break projective symmetry, since configurations such as collinearity and contact are indistinguishable from most angles. Cognitive Mapping tasks require multi-step locomotion to construct topological representations of space that no single observation can provide. Temporal Scene Understanding tasks require manipulation and interaction to trigger or observe state changes, while Procedural Sequencing tasks require executing and reasoning over ordered action sequences to determine causal dependencies. Across all categories, the correct answer emerges not from any single image but from the agent's capacity to act selectively and reason over the result.

potentially appendix if boring.... or should be interesting. select more interesting ones....

Table 2: **ESI-Bench task taxonomy.** Ten categories spanning 30 subcategories, each targeting a distinct spatial faculty requiring embodied action to resolve.

Category	Subcategory	Description	Example
Physics Capacity	Rigid Containment	Plan the placement of multiple objects across multiple containers	How can all the toys be fit into the boxes?
	Liquid Volume	Compare liquid-holding capacity of containers	Which container has larger volume?
	Deformable Fitting	Whether a deformable container can conform to an object	How to fold the sweater to make it fit?
Physics Dynamics	Inclined Plane	Predict object motion and stability on slopes	Can the apple sit stable on the slope?
	Stacking Stability	Whether objects stack or balance given shape, mass and geometry	How to stack and stabilize these three objects?
Specular Reflection	Reflection Authorization	Distinguish real objects from mirror reflections	Is the object real or reflected?
	Spatial Relationship	Infer relations among objects across mirror and real-world views	What relations does the mirror reveal?
	Scene Correspondence	Identify which objects appear in mirror given the real-world scene	Which of the 3 snacks appears in the mirror?
Perceptual Grounding	Partial Occlusion	Reason about objects hidden behind other scene elements	Golf Stick or Umbrella behind the Wall?
	Viewpoint Hallucination	Detect objects whose visibility changes critically with viewing angle	Cabinet or piano from this view?
	Material Transparency	Reason about objects seen through transparent surfaces	Is the object inside the glass or not?
Metric Comparison	Dimensional Size	Compare relative sizes of objects	Which vase is larger?
	Spatial Distance	Compare relative distances with respect to a reference object	Which flower is closer to the table?
Enumerative Perception	Occluded Counting	Count objects partially obscured by other scene elements	How many balls are under the blanket?
	Spatial Segmentation	Count objects separated across distinct spatial regions	One or two cylinders separated by the post?
	Category Ambiguity	Count visually similar objects requiring fine-grained distinction	How many apples among the balls?
	Merged Observation	Count groups that appear visually merged from a single viewpoint	How many separate stacks of books?
	Illumination Variability	Count objects under challenging or non-uniform lighting	Objects in the dim scene?
	Structural Enclosure	Count objects hidden within enclosed or covered spaces	Objects in the microwave?
Spatial Relations	Linear Alignment	Whether objects are arranged along a common axis	Do the glasses form a line?
	Geometric Configuration	Identify the shape formed by a set of objects	Do the cups form an equilateral triangle?
	Physical Contact	Detect whether two or more objects are in direct contact	Are the teddy bears touching each other?
Cognitive Mapping	Topology & Connectivity	Whether two locations or regions are mutually reachable	Is region A connected to C?
	Traversable Passage	Identify navigable corridors or passageways between regions	Passage between the rooms?
	Regional Boundary	Identify and delineate distinct functional spatial regions	Boundaries of this region?
	Long-Horizon Navigation	Plan multi-step navigation trajectories toward a distant goal	What room is behind the wall?
Temporal Scene	Unobserved State Change	Infer scene changes that occurred during an unobserved interval	What changed when I looked away?
	Multi-Agent Interaction	Reason about scene dynamics from other agents	Other robot's world model?
Procedural Sequencing	Action Ordering	Determine the correct procedural ordering of a sequence of actions	What is the action sequence for assembly?

4 Experiments

4.1 Models and Evaluation Setup

We evaluate models across four paradigms, organized by the degree of action and perceptual access granted to the agent: (1) **Passive Single-View** provides a single fixed observation from the initial pose, establishing a baseline under conditions identical to existing spatial benchmarks; (2) **Passive Multi-View** provides a set of pre-defined randomly sampled views designed to broadly cover the full environment, simulating exhaustive passive scene coverage without any agent action or viewpoint selection; (3) **Active Exploration** places the agent the initial pose with full access to the action space, requiring it to gather evidence through deliberate movement and interaction before committing to a final answer; Ground-Truth Passive provides the sequence of views rendered along the ground-truth action trajectory, serving as an oracle ablation that *separates perception errors from action errors*. Comparing (1) vs. (2) reveals whether exhaustive passive coverage helps; comparing (2) vs. (4) isolates the benefit of action-guided over passive coverage; comparing (4) vs. (3) isolates whether failures stem from the agent's inability to select informative actions or from perceptual limitations given the perfect views themselves.

We evaluate two families of models. 2D vision-language models include GPT-5 and Gemini 3.1, each taking egocentric visual observations as input, evaluated across all four paradigms. 3D-augmented models include VGGT+Gemini, where explicit 3D scene representations are reconstructed from multi-view observations via VGGT, from which scene graphs are constructed and provided to the language model; and Ground-Truth 3D+Gemini, where perfect point clouds derived directly from simulator state are used to construct scene graphs provided to the language model, serving as an oracle ablation for 3D grounding. Human performance is additionally collected on a subset of tasks to establish a human upper bound. All models are evaluated zero-shot with no task-specific fine-tuning, and results are reported as accuracy on matched subsets to ensure fair cross-paradigm comparison.

4.2 What Really Holds Spatial Intelligence Back: Seeing or Acting?

Table 3 reports accuracy across all models, paradigms, and task categories. Without any explicit instruction, active agents *spontaneously discover emergent spatial strategies*: to determine whether

Table 3: **ESI-Bench results (accuracy %).** *2D + VLM*: GPT-5, Gemini 3.1. *3D + LLM*: VGGT + Gemini, Ground-truth 3D+ Gemini.

	2D + VLM								3D + LLM								Human Performance			
	GPT-5				Gemini 3.1				VGGT + Gemini				GT 3D + Gemini							
Category / Subcategory	*Passive Single*	*Passive Multiple*	*Active Multiple*	*GT Passive*	*Passive Single*	*Passive Multiple*	*Active Multiple*	*GT Passive*	*Passive Single*	*Passive Multiple*	*Active Multiple*	*GT Passive*	*Passive Single*	*Passive Multiple*	*Active Multiple*	*GT Passive*	*Passive Single*	*Passive Multiple*	*Active Multiple*	*GT Passive*
Perceptual Grounding																				
Partial Occlusion	30.5	32.9	62.4	91.5	14.6	20.7	78.5	95.1	45.4	40.2	65.9	80.4	52.6	53.6	79.4	93.8	38.1	37.1	79.4	86.6
View Hallucination	11.7	20.2	60.1	87.8	39.9	32.9	68.1	91.1	56.3	59.0	76.1	87.6	61.6	60.6	74.1	89.0	52.2	45.8	80.8	83.8
Material Transparency	30.3	36.7	66.1	96.3	44.0	45.0	52.3	88.0	37.4	29.4	31.8	90.9	27.8	31.2	60.4	100	41.3	44.0	93.6	97.2
Physical Structure																				
Rigid Containment	45.0	42.5	42.5	95.0	47.5	40.0	67.5	97.5	27.5	37.5	57.5	72.5	45.0	42.5	65.0	95.0	47.5	42.5	92.5	100
Liquid Volume	66.2	66.2	81.6	86.0	69.9	67.6	80.9	86.8	65.4	56.7	71.3	77.2	74.5	74.5	83.1	86.8	79.4	66.9	89.7	86.8
Deformable	42.9	41.8	55.1	75.5	34.7	41.8	43.9	78.6	43.9	42.9	49.0	81.6	98.0	98.0	99.0	99.0	57.1	60.2	76.5	82.7
Physical Dynamics																				
Inclined Plane	57.4	60.7	77.0	86.9	65.6	62.3	83.6	88.5	67.2	62.3	80.3	86.9	63.9	63.9	83.6	91.8	60.7	62.3	83.6	83.6
Stacking & Stability	34.8	37.1	62.9	86.5	38.2	36.0	78.7	84.3	34.8	39.3	55.1	62.9	27.2	33.7	68.5	86.5	36.0	39.3	84.3	86.5
Specular Reflection																				
Reflection Authoring	68.7	70.7	70.3	73.7	60.6	60.6	64.9	67.0	52.5	54.6	53.5	55.6	55.6	58.6	55.6	60.6	94.9	88.9	96.0	96.0
Spatial Relations	50.4	38.9	54.8	58.4	43.4	42.5	44.2	46.9	41.6	39.8	41.6	43.4	38.9	35.4	54.2	41.6	78.8	78.8	84.1	87.6
Correspondence	39.8	51.1	52.3	56.8	37.5	40.9	42.0	42.0	37.5	43.2	48.9	48.9	31.8	39.8	42.3	48.9	85.2	80.7	89.8	92.0
Enumerative Perception																				
Counting w Occlusion	3.3	3.3	13.3	56.7	3.3	3.3	10.0	63.3	0.0	0.0	13.3	53.3	33.3	80.0	83.3	100.0	6.7	10.0	53.3	76.7
Spatial Segmentation	3.3	6.7	26.7	63.3	3.3	3.3	16.7	70.0	3.3	3.3	23.3	56.7	43.3	80.0	66.7	86.7	13.3	23.3	63.3	70.0
Merged Observation	38.3	35.0	51.7	50.0	43.3	51.7	61.7	75.0	63.3	51.7	55.0	65.0	100.0	100.0	93.3	100.0	60.0	60.0	65.0	73.3
Category Ambiguity	8.3	10.0	8.3	41.7	13.3	13.3	15.0	48.3	21.7	26.7	26.7	46.7	18.3	76.7	40.0	93.3	23.3	25.0	51.7	61.7
Structural Enclosure	5.0	10.0	22.5	67.5	2.5	0.0	10.0	52.5	0.0	0.0	12.5	52.5	40.0	60.0	50.0	100.0	10.0	22.5	42.5	77.5
Illumination Variability	6.0	22.0	22.0	46.0	12.0	16.0	22.0	58.0	10.0	28.0	30.0	40.0	20.0	84.0	42.0	96.0	20.0	34.0	62.0	66.0
Spatial Relations																				
Linear Alignment	27.7	31.9	42.6	60.6	47.9	44.6	67.0	77.7	45.7	38.3	53.2	59.6	73.4	73.4	84.0	89.4	51.1	39.4	73.4	79.8
Geometric Configuration	25.3	20.4	26.0	26.0	27.5	22.3	27.5	44.6	9.9	11.3	18.6	32.6	70.8	73.6	88.0	100	32.4	33.5	74.3	86.3
Physical Contact	40.0	41.7	64.2	90.0	60.8	55.8	70.0	74.2	59.2	41.7	59.2	78.2	65.8	67.5	70.8	72.5	35.8	40.8	88.3	90.8
Metric Comparison																				
Dimensional Size	42.5	44.9	67.7	80.3	44.3	41.3	68.3	82.6	50.4	40.1	59.3	80.8	58.6	60.5	69.9	85.9	48.5	56.9	82.6	91.9
Spatial Distance	53.9	49.1	58.6	73.7	52.6	50.5	59.9	80.3	50.7	47.3	55.0	71.4	59.6	61.2	64.5	96.7	57.9	59.9	69.1	78.9
Cognitive Mapping																				
Connectivity	68.3	70.0	68.3	78.3	51.7	48.3	60.0	85.0	66.7	66.7	73.3	83.3	65.0	65.0	71.7	86.7	68.3	68.3	81.7	91.7
Traversable Passage	68.3	66.7	71.7	73.3	66.7	61.7	73.3	78.3	68.3	63.3	66.7	78.3	71.7	73.3	71.7	80.0	70.0	71.7	78.3	85.0
Regional Boundary	65.0	63.8	65.0	67.5	63.8	62.5	65.0	65.0	65.0	65.0	62.5	67.5	60.0	68.8	62.5	68.8	61.3	62.5	67.5	70.0
Long-Term Navigation	40.0	38.3	41.7	50.0	33.3	35.0	33.3	41.7	36.7	40.0	43.3	51.6	36.7	41.7	43.3	40.0	35.0	36.7	51.7	61.7
Temporal Understanding																				
Unobserved Change	40.5	41.2	51.4	77.0	37.2	37.8	47.3	74.5	37.8	41.9	45.8	76.6	39.9	42.6	46.5	74.5	39.4	39.4	70.8	81.0
Agent Observation	40.6	30.1	51.0	72.7	37.6	36.1	58.6	65.4	32.7	27.9	34.2	56.5	36.3	33.3	76.7	87.8	38.3	42.1	83.5	90.2
Action Sequencing																				
Action Order Inference	36.4	37.7	41.6	67.5	44.2	46.8	54.5	72.7	35.1	31.1	50.6	58.4	44.2	46.8	51.9	75.3	40.3	41.6	74.0	81.8

a chestnut is inside a glass, agents independently develop four distinct approaches, moving behind the object, repositioning top-down, picking it up, and pouring it out (Figure 3a), none of which were prescribed. These emergent abilities drive consistent and substantial gains across tasks requiring deliberate repositioning: on Dimensional Size GPT-5 improves from 42.5% to 67.7%, on Physical Contact from 40.0% to 64.2%, and on Viewpoint Hallucination Gemini 3.1 jumps from 39.9% to 68.1%. In contrast, passive multi-view provides negligible or negative gains despite consuming far more images (*e.g.*, PT-5 drops from 53.9% to 49.1% on Spatial Distance compared with single-image passive baseline), confirming that observation quantity without selective action adds noise rather than signal.

Ground-truth passive trajectories reveal that the bottleneck is almost entirely in action selection, not perception: GPT-5 reaches 90.0% on Physical Contact and 95.0% on Rigid Containment under oracle views, and Gemini 3.1 jumps from 14.6% to 95.1% on Partial Occlusion. Geometric Configuration and Specular Reflection stand as exceptions where oracle views provide minimal benefit: GPT-5 reaches only 26.0% on Geometric Configuration under ground-truth passive, and models consistently fail to determine whether three objects form an equilateral triangle even from the perfect viewpoint; on Specular Reflection, models hallucinate objects in mirrors that do not exist, or fail to identify the

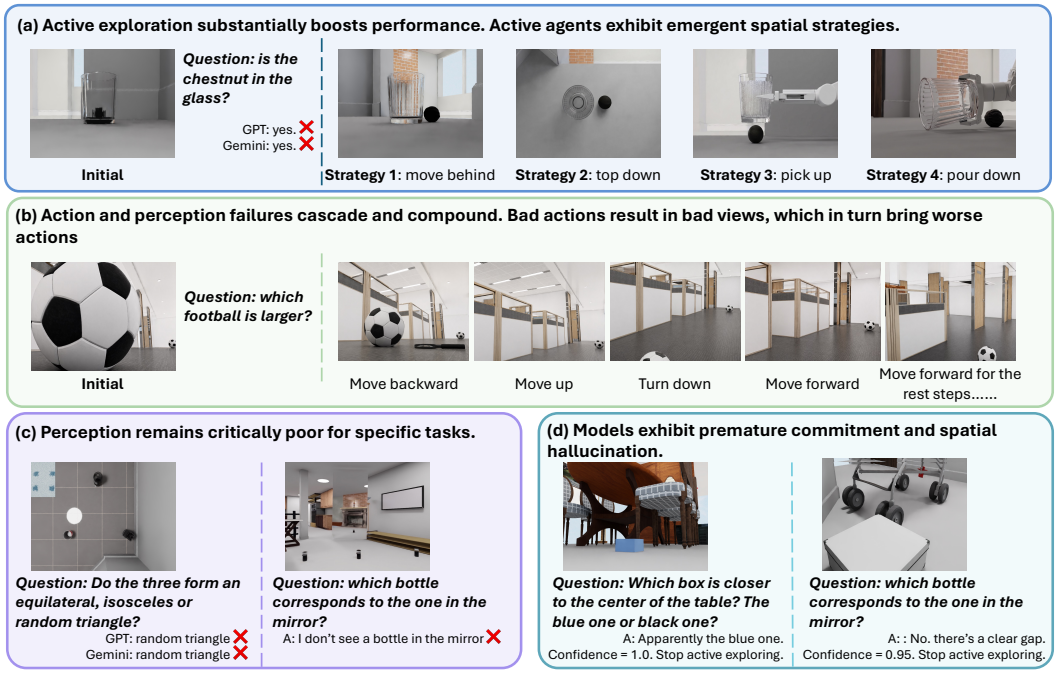


Figure 3: Qualitative results of ESI-Bench.

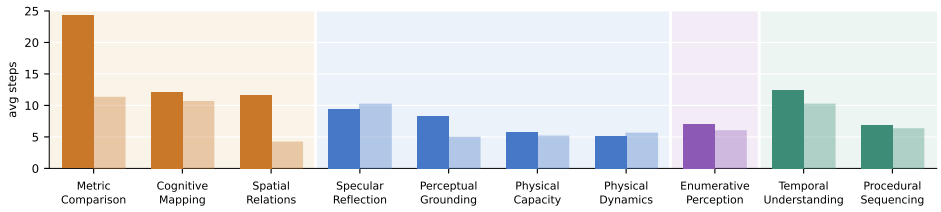


Figure 4: Average number of active exploration steps to reach a correct answer for GPT-5 (solid) and Gemini 3.1 (outlined), grouped by Spelke's four core knowledge systems.

correct real-world correspondence altogether (Figure 3c). These indicate hard perceptual limits that no action strategy can overcome.

The active-to-oracle gap confirms that actions and perception failures cascade and compound: on Occluded Counting the gap reaches 33.6%, and on Structural Enclosure 49.7%, confirming that bad actions result in bad views, which in turn bring worse actions that cannot be recovered within the step budget (Figure 3b).

Action Blindness Dominates Perceptual Blindness, while Their Coupling Drives Failure Cascades

- Without explicit instruction, agents spontaneously acquire spatial strategies, yielding substantial gains over passive multi-view, which often adds noise rather than signal.
- For most tasks, perception is not the bottleneck: with the right viewpoint, the agent succeeds; yet some tasks hit a hard perceptual ceiling that no action can overcome.
- Suboptimal actions produce bad views, which in turn produce worse actions, cascading into reasoning failures unrecoverable within the step budget.

4.3 When Does 3D Help, and When Does It Hurt?

Ground-Truth 3D+Gemini reaches 60.4% on Material Transparency versus 44.0% for Gemini 3.1, a 16.4 point improvement from 3D grounding alone, most pronounced on tasks where 2D projections

fundamentally lose depth information. However, VGGT+Gemini drops to 9.9% on Geometric Configuration versus Gemini 3.1's 27.5%, a 17.6 point degradation from reconstruction artifacts. Imperfect geometry actively misleads rather than merely failing to help: VGGT reconstructions distort fine-grained spatial relations, causing the model to reason over a corrupted scene graph rather than the true geometry, making 3D augmentation a strategy that amplifies failures.

3D Helps When Geometry Is Perfect, But Imperfect Reconstruction Actively Misleads

- Explicit 3D grounding improves performance on tasks where depth and occlusion render 2D projections fundamentally ambiguous.
- Imperfect 3D reconstruction actively degrades performance, proving more harmful than 2D baselines.

4.4 How Far Are Models From Human-Level Spatial Reasoning?

Human annotators consistently gather more observations before committing, actively seek viewpoints that falsify their current hypothesis, and reduce confidence under ambiguity. Models commit after fewer steps with uniformly high confidence regardless of evidence quality, manifesting as spatial hallucination: asserting object properties that directly contradict the scene state (Figure 3d). The gap is not perceptual but metacognitive: models lack the awareness to recognize when current observations are insufficient, a failure that neither better perception nor more embodied interaction alone can fully close.

Models Can See But Do Not Know When They Have Seen Enough

- Models commit prematurely with high confidence; humans treat uncertainty as a signal to keep looking, not a reason to answer.
- Humans seek viewpoints that falsify their hypothesis; models seek confirmation and move in the same direction repeatedly.
- Humans revise beliefs when contradicted; models anchor to their first impression, revealing a fundamental absence of belief updating.

5 Conclusion

We introduced ESI-Bench, a benchmark spanning 10 task categories and 30 subcategories that requires agents to close the perception-action loop rather than reason from pre-given observations. Our experiments show that active exploration substantially outperforms passive counterparts, while passive multi-view coverage adds noise rather than signal. The primary bottleneck across nearly all tasks is not perception but action selection: given oracle viewpoints, models succeed; given the freedom to act, they struggle to acquire them. Beyond accuracy, we identify a persistent gap in epistemic calibration: unlike humans, models commit prematurely with high confidence regardless of evidence quality, a failure that neither better perception nor more exploration alone can resolve. ESI-Bench establishes a clear frontier for embodied spatial intelligence, where closing the perception-action loop remains an open and fundamental challenge.

References

References

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 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 (CVPR)*, pages 14455–14465, June 2024.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language models, 2024. URL https://arxiv.org/abs/2406.01584.

Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, Lei Shi, and Maosong Sun. Embodiedeval: Evaluate multimodal llms as embodied agents, 2025. URL https://arxiv.org/abs/2501.11858.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding, 2025. URL https://arxiv.org/abs/2501.16411.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models, 2024. URL https://arxiv.org/abs/2406.05756.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction, 2026. URL https://arxiv.org/abs/2505.20279.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

Kaixuan Jiang, Yang Liu, Weixing Chen, Jingzhou Luo, Ziliang Chen, Ling Pan, Guanbin Li, and Liang Lin. Beyond the destination: A novel benchmark for exploration-aware embodied question answering, 2025. URL https://arxiv.org/abs/2503.11117.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023. URL https://arxiv.org/abs/2205.00363.

Mengzhen Liu, Enshen Zhou, Cheng Chi, Yi Han, Shanyu Rong, Liming Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Sapave: Towards active perception and manipulation in vision-language-action models for robotics, 2026a. URL https://arxiv.org/abs/2603.12193.

Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, Helong Huang, Guangjian Tian, Weichao Qiu, Xingyue Quan, Jianye Hao, and Yuzheng Zhuang. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning, 2025. URL https://arxiv.org/abs/2501.10074.

Zhenyang Liu, Yongchong Gu, Yikai Wang, Xiangyang Xue, and Yanwei Fu. Activevla: Injecting active perception into vision-language-action models for precise 3d robotic manipulation, 2026b. URL https://arxiv.org/abs/2601.08325.

Wenxin Ma, Chenlong Wang, Ruisheng Yuan, Hao Chen, Nanru Dai, S. Kevin Zhou, Yijun Yang, Alan Yuille, and Jieneng Chen. Causalspatial: A benchmark for object-centric causal spatial reasoning, 2026. URL https://arxiv.org/abs/2601.13304.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Celso M de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark, 2025. URL https://arxiv.org/abs/2412.07825.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

J. Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–973, 2001.

Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

Francisco J. Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1991.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Mindcube: Spatial mental modeling from limited views, 2026. URL https://arxiv.org/abs/2506.21458.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence, 2025a. URL https://arxiv.org/abs/2505.23747.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence, 2026a. URL https://arxiv.org/abs/2505.17012.

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.

Yuhao Wu, Maojia Song, Yihuai Lan, Lei Wang, Zhiqiang Hu, Yao Xiao, Heng Zhou, Weihua Zheng, Dylan Raharja, Soujanya Poria, and Roy Ka-Wei Lee. From perception to action: An interactive benchmark for vision reasoning, 2026b. URL https://arxiv.org/abs/2602.21015.

Haoyu Xiong, Xiaomeng Xu, Jimmy Wu, Yifan Hou, Jeannette Bohg, and Shuran Song. Vision in action: Learning active perception from human demonstrations, 2025. URL https://arxiv.org/abs/2506.15666.

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, 2024. URL https://arxiv.org/abs/2412.14171.

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, 2025a. URL https://arxiv.org/abs/2412.14171.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents, 2025b. URL https://arxiv.org/abs/2502.09560.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video, 2025c. URL https://arxiv.org/abs/2511.04670.

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, 2025d. URL https://arxiv.org/abs/2505.23764.

Heyang Yu, Yinan Han, Xiangyu Zhang, Baiqiao Yin, Bowen Chang, Xiangyu Han, Xinhao Liu, Jing Zhang, Marco Pavone, Chen Feng, Saining Xie, and Yiming Li. Thinking in 360°: Humanoid visual search in the wild, 2025. URL https://arxiv.org/abs/2511.20351.

Yanpeng Zhao, Wentao Ding, Hongtao Li, Baoxiong Jia, and Zilong Zheng. Espire: A diagnostic benchmark for embodied spatial reasoning of vision-language models, 2026. URL https://arxiv.org/abs/2603.13033.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **[TODO]**

   Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: **[TODO]**

    Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.