# ESI-BENCH

## Towards Embodied Spatial Intelligence that Closes the Perception-Action Loop (ArXiv'26)

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

[1]Stanford University [2]UCLA [3]Northwestern University

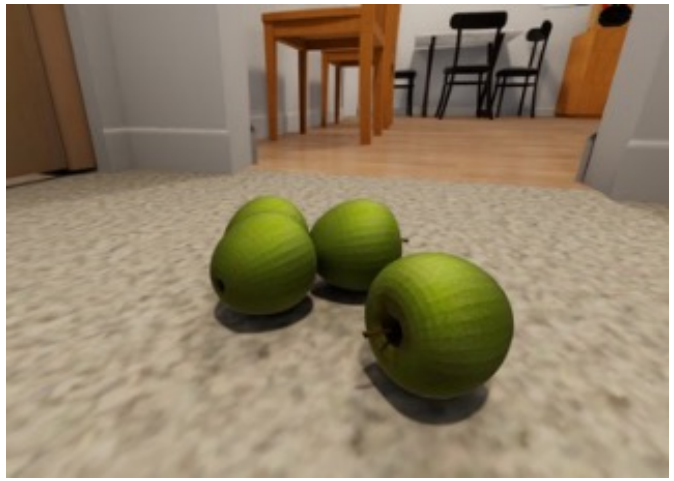
https://esi-bench.github.io/

**Sangoh Lee**

June 17, 2026

# 01
# Motivation

# Can you answer this — from a single photo?



**How many apples are in this scene?**



**Is anything hidden inside the container?**

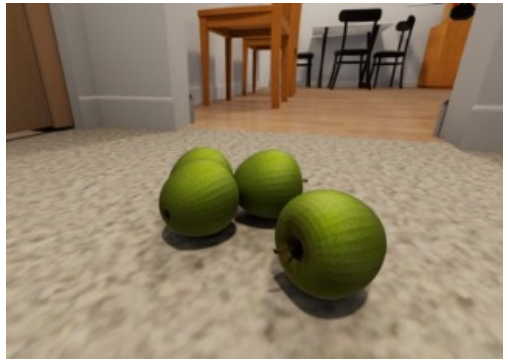


**Which flower is closer to the table?**

A fixed viewpoint underdetermines the answer. The information is insufficient in the pixels.

# Each question needs an action, not just a look

- **Count under occlusion:** move around the pile to separate merged / hidden instances
- **Hidden containment:** open the box and look inside, the state is not visible
- **Metric Comparison:** shift viewpoint sideways to break forced-perspective distortion







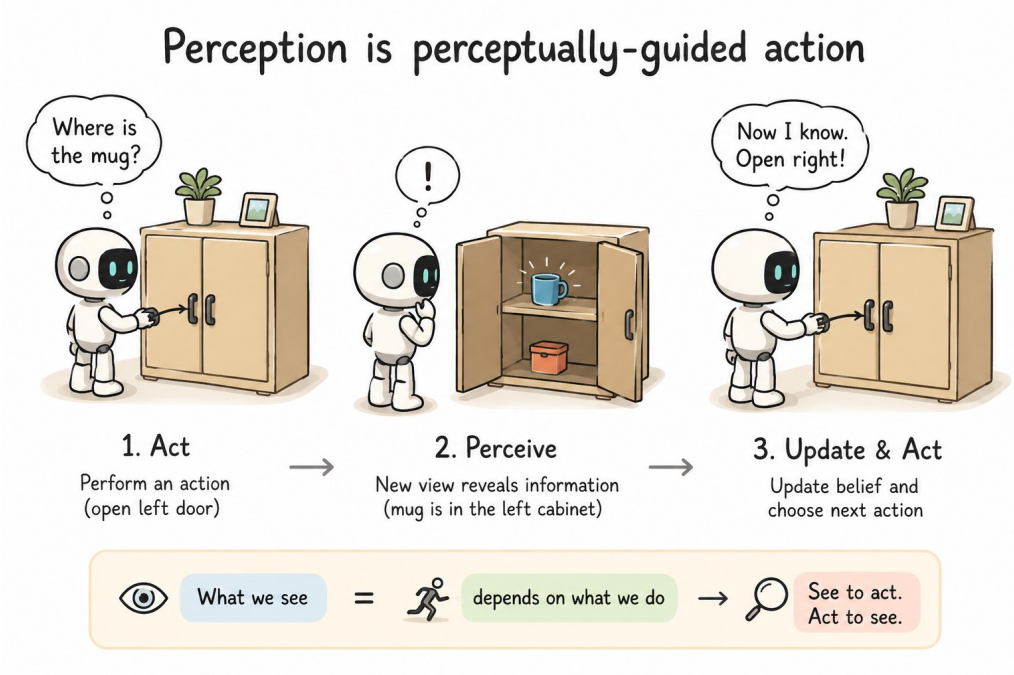
**Core idea**

The hard part is not perception — it is knowing **which action** yields the missing information.

# Perception is perceptually-guided action

- Cognitive science (Gibson; O'Regan & Noë): perceiving is an **activity**, not passive reception.
- What we see is governed by how observations change as a function of our actions.
- Spatial intelligence concerns not only **what is seen, but what is unseen**.
- Latent structure — occlusion, dynamics, containment, function — must be revealed by acting.

> "Knowing which actions elicit informative sensing is often harder than sensing itself."



# From observer to actor

**Passive sensing**

oracle observations are given; the agent only reads them



**Active acquisition**

the agent decides which observations are worth acquiring

## Three shifts this reframing forces:

### 01

### Sensing → competence

Not just what can be perceived, but which embodied abilities to deploy to solve the task.

### 02

### Selective sensing

Prioritize task-relevant information over redundant or uninformative inputs.

### 03

### Resolving ambiguity

Reason through misleading views to infer hidden structure and physical constraints.

# 02

# The Benchmark

# What exactly ESI-Bench has targeted?

Table 1: Comparison with spatial reasoning and embodied benchmarks. Here, *embodied* means that an agent is situated in an embodied environment and acts through an explicit action space in the environment. **Active?**: whether the benchmark supports active perception. **Action Space**: L = locomotion, P = perception, M = manipulation. **Core Knowledge**: Obj = object, Geom = geometry, Num = number, Phys = physics, Agt = agent. **Hidden**: whether the benchmark targets latent properties such as occlusion, containment, dynamics that are inaccessible from passive observation.

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

# Benchmark setups

**Each task in ESI-Bench:** $(\boldsymbol{S}, \boldsymbol{p_0}, \boldsymbol{q}, \boldsymbol{y^*})$

- $S$ = 3D scene instantiated from **BEHAVIOR-1k**
- $p_0$ = agent's initial pose
- $q$ = natural-language question
- $y^*$ = ground-truth answer

**Environment:** $\boldsymbol{E} = \langle \boldsymbol{S}, \boldsymbol{A}, \boldsymbol{O}, \boldsymbol{T} \rangle$

- $S$ = 3D scene instantiated from BEHAVIOR-1k
- $\mathrm{A}$ = action space
- $\mathrm{O}$ = egocentric observation space
- $\mathrm{T}$ = transition function from $S \times A$ to $S$

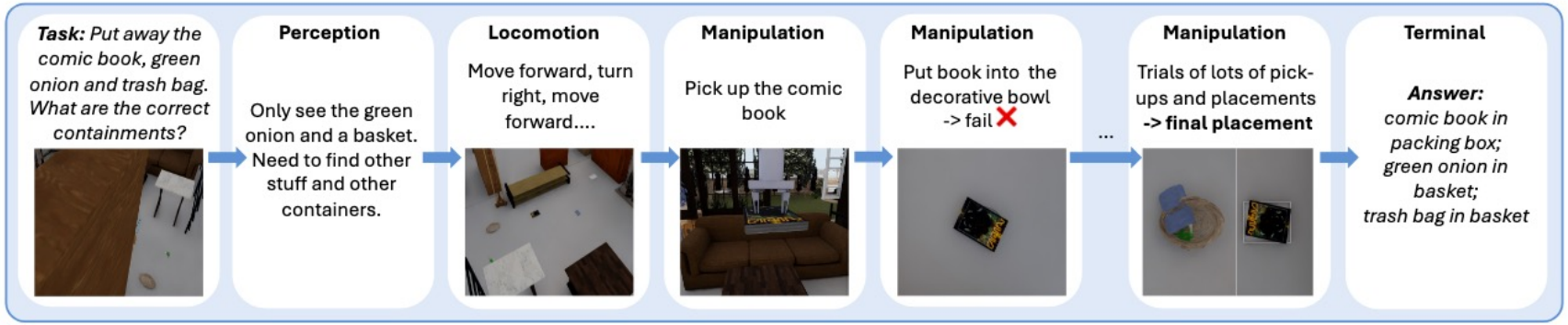
**Given $(\boldsymbol{S}, \boldsymbol{p_0}, \boldsymbol{q})$, the agent**

1. receives observation $\mathrm{o_t}$ at each timestep
2. issues action $\mathrm{a_t}$ (e.g., move, rotate, interact)
3. induces a trajectory $\tau = (o_0, a_0, o_1, a_1, \dots)$
4. repeats it until committing to a final answer $\hat{y}$ with confidence $\mathrm{c}$ (within budget 30)

# Benchmark setups: action space and trajectory example

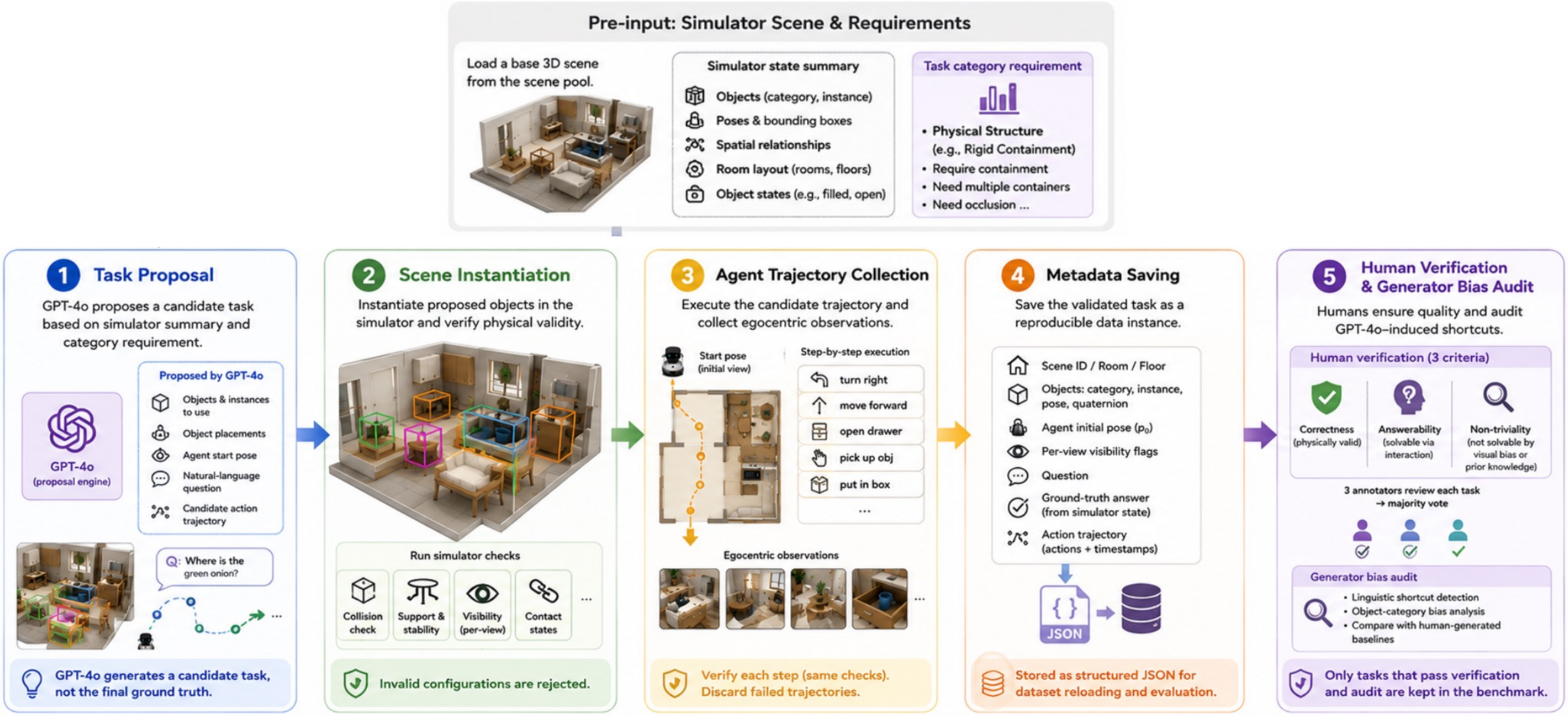
| Action | Description |
|---|---|
| *Locomotion* | |
| `move_forward` / `move_backward` | Translate agent along viewing axis |
| `move_left` / `move_right` | Translate agent laterally |
| `move_up` / `move_down` | Translate agent vertically |
| *Perception* | |
| `turn_left` / `turn_right` | Rotate 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 $c$ |

(b) ESI-BENCH agent action space.




(a) A dataset example including question, answer and trajectory, from the Rigid Containment sub-category.

# Task construction details



# Task Categories and Statistics

## 3081 tasks across 10 categories and 29 subcategories

Table 3: **ESI-BENCH task taxonomy and definitions** of 10 categories spanning 29 subcategories. Colors correspond to Spelke's four core knowledge systems: object representation, layout and geometry, number representation, and agents and goal-directed actions.

| Category | Subcategory | Description | Example |
|---|---|---|---|
| Physical Structure | 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 | Infer structure from deformable object | What's under the sweater, apple or mug? |
| Physical 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 Authoring | Distinguish real objects from mirror reflections | Is the object real or reflected? |
| | Spatial Relations | Infer relations among objects across mirror and real-world views | What relations does the mirror reveal? |
| | 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? |
| | View 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? |
| 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 | 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-Term Navigation | Plan multi-step navigation trajectories toward a distant goal | What room is behind the wall? |
| Enumerative Perception | Counting w Occlusion | 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? |
| Temporal Understanding | Unobserved Change | Infer scene changes that occurred during an unobserved interval | What changed when I looked away? |
| | Agent Observation | Reason about scene dynamics from other agents | Other robot's world model? |
| Action Sequencing | Action Order Inference | Determine the correct procedural ordering of a sequence of actions | What is the action sequence for assembly? |

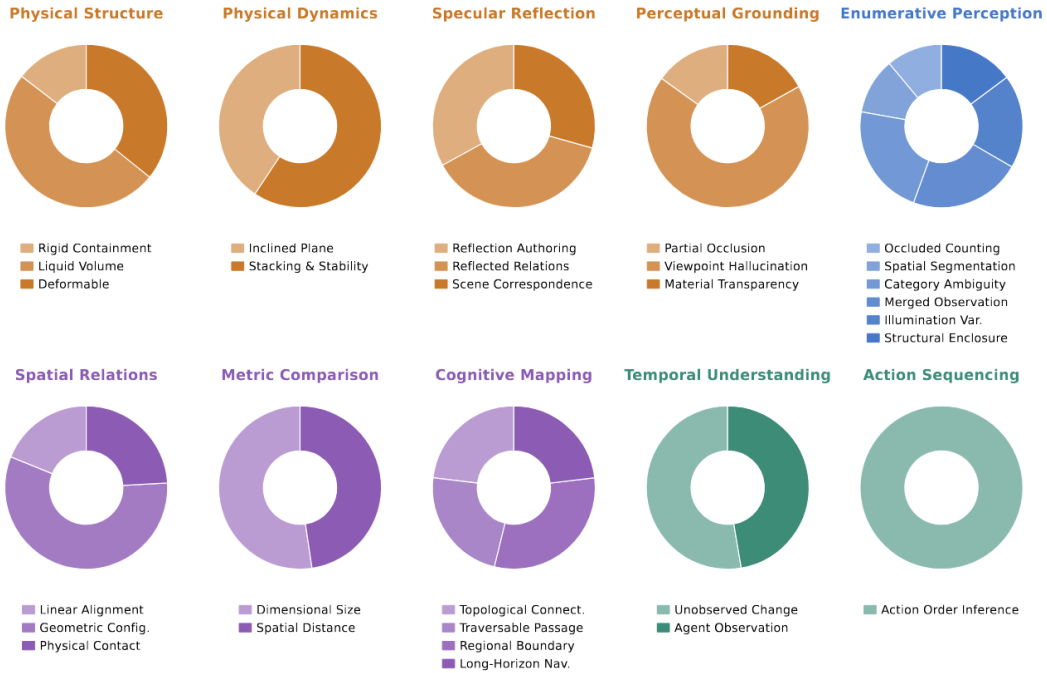
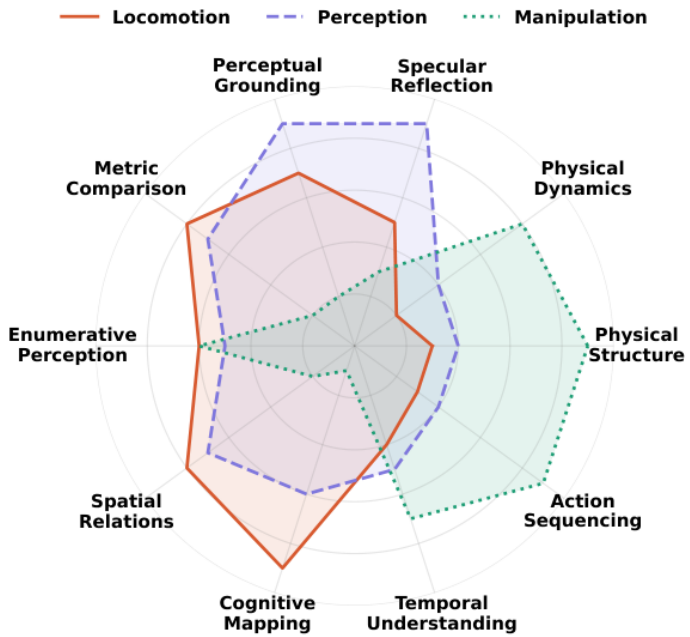



Figure 6: Subcategory distribution within each of the 10 ESI-BENCH task categories.

# What each task demands

| Category | Definition |
|---|---|
| **Physical Structure** | Reveal hidden containment capacity through manipulation |
| **Physical Dynamics** | Predict stability and motion under physical forces |
| **Specular Reflection** | Infer object relations from mirror reflections |
| **Perceptual Grounding** | Resolve viewpoint-dependent perceptual ambiguities |
| **Metric Comparison** | Overcome forced-perspective distortions in size and distan |
| **Spatial Relations** | Navigate to vantage points that break projective symmetry |
| **Cognitive Mapping** | Construct topological representations through locomotion |
| **Enumerative Perception** | Count objects despite occlusion, merging, or enclosure |
| **Temporal Understanding** | Infer state changes through interaction and observation |
| **Action Sequencing** | Reason over action sequences and causal dependencies |

(a) ESI-Bench task categories and definitions.




(b) Required action types per category.

Each category requires a distinct **combination of locomotion, perception, and manipulation** — not perception alone.

*Category definitions (left) and the embodied action types each requires (right) (Fig. 3).*

# 03

# Experimental Results

# Evaluation Setup

Evaluate models across 4 paradigms, organized by degree of action and perceptual access granted to the agent

1. **Passive Single-View**
   - Provide a single fixed observation from the initial pose
   - Identical setting to existing spatial benchmarks
2. **Passive Multi-View**
   - Set of 30 views along a **randomly-sampled trajectory** starting at the initial pose in the scene
   - Simulate exhaustive passive scene coverage without any agent active action or viewpoint
3. **Active Exploration**
   - Place the agent the initial pose with full access to the action space
   - Require it to gather evidence through deliberate movement and interaction
4. **Ground-Truth Passive**
   - Provide the sequence of views redarned from **GT action trajectory**
   - Serve as an oracle ablation separating perception errors from action errors

# Used Models

2 families of models: VLM only & 3D scene representations + VLM

1. **2D Vision-Language Models (VLMs):** give egocentric visual observations as input
   - GPT-5
   - Gemini 3.1
2. **3D-augmented VLMs:** 3D scene representations are reconstructed from multi-view observations via VGGT, and scene graphs from reconstructed result are fed to VLM
   - VGGT + Gemini 3.1
   - GT-3D + Gemini 3.1
3. **Cf) Human performance**



Figure 1. **VGGT** is a large feed-forward transformer with minimal 3D-inductive biases trained on a trove of 3D-annotated data. It accepts up to hundreds of images and predicts cameras, point maps, depth maps, and point tracks for all images at once in less than a second, which often outperforms optimization-based alternatives without further processing.

# Experimental Results

| Category / Subcategory | 2D + VLM | | | | | | | | 3D + LLM | | | | | | | | Human Performance | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GPT-5 | | | | Gemini 3.1 | | | | VGGT + Gemini | | | | GT 3D + Gemini | | | | | | | |
| | *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** | | | | | | | | | | | | | | | | | | | | |
| *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 |
| *Partial Occlusion* | 32.6 | 26.3 | 47.4 | 86.3 | 30.5 | 32.6 | 70.5 | 88.4 | 51.6 | 48.4 | 53.7 | 86.3 | 56.8 | 55.8 | 72.6 | 91.6 | 37.9 | 34.7 | 80.0 | 87.4 |
| *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 | 46.9 | 54.1 | 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 |
| **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 |
| **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 | 51.7 | 63.3 | 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 |
| **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 |

## Q1: What Really Holds Spatial Intelligence Back: Seeing or Acting?

1. Active exploration beats passive observation by a large margin, and agents discover spatial strategies with no instruction.

2. More passive images ≠ more signal: 30 random views barely help, sometimes hurt.

3. Most failures are "action blindness," not weak perception: bad actions yield bad views, which cascade into worse actions.

4. But for some tasks, perception itself is the hard limit that no action can fix.

# Experimental Results (cont'd)

| | 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** | | | | | | | | | | | | | | | | | | | | |
| *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 |
| *Partial Occlusion* | 32.6 | 26.3 | 47.4 | 86.3 | 30.5 | 32.6 | 70.5 | 88.4 | 51.6 | 48.4 | 53.7 | 86.3 | 56.8 | 55.8 | 72.6 | 91.6 | 37.9 | 34.7 | 80.0 | 87.4 |
| *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 | 46.9 | 54.1 | 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 |
| **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 |
| **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 | 51.7 | 63.3 | 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 |
| **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 |

## Q2: When does 3D help, and when does it hurt?

1. Accurate (ground-truth) 3D helps a lot, resolving ambiguities hard to infer from 2D.

2. But imperfect, reconstructed 3D is actively harmful — worse than 2D.
   - Why: noisy reconstruction distorts fine-grained spatial relations, so the LLM reasons over a corrupted scene graph.

3. 3D augmentation is a high-variance strategy that amplifies both successes and failures.

# Experimental Results (cont'd)

| | 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** | | | | | | | | | | | | | | | | | | | | |
| *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 |
| *Partial Occlusion* | 32.6 | 26.3 | 47.4 | 86.3 | 30.5 | 32.6 | 70.5 | 88.4 | 51.6 | 48.4 | 53.7 | 86.3 | 56.8 | 55.8 | 72.6 | 91.6 | 37.9 | 34.7 | 80.0 | 87.4 |
| *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 | 46.9 | 54.1 | 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 |
| **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 |
| **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 | 51.7 | 63.3 | 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 |
| **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 |

## Q3: How Far are Models from Human-level Reasoning?

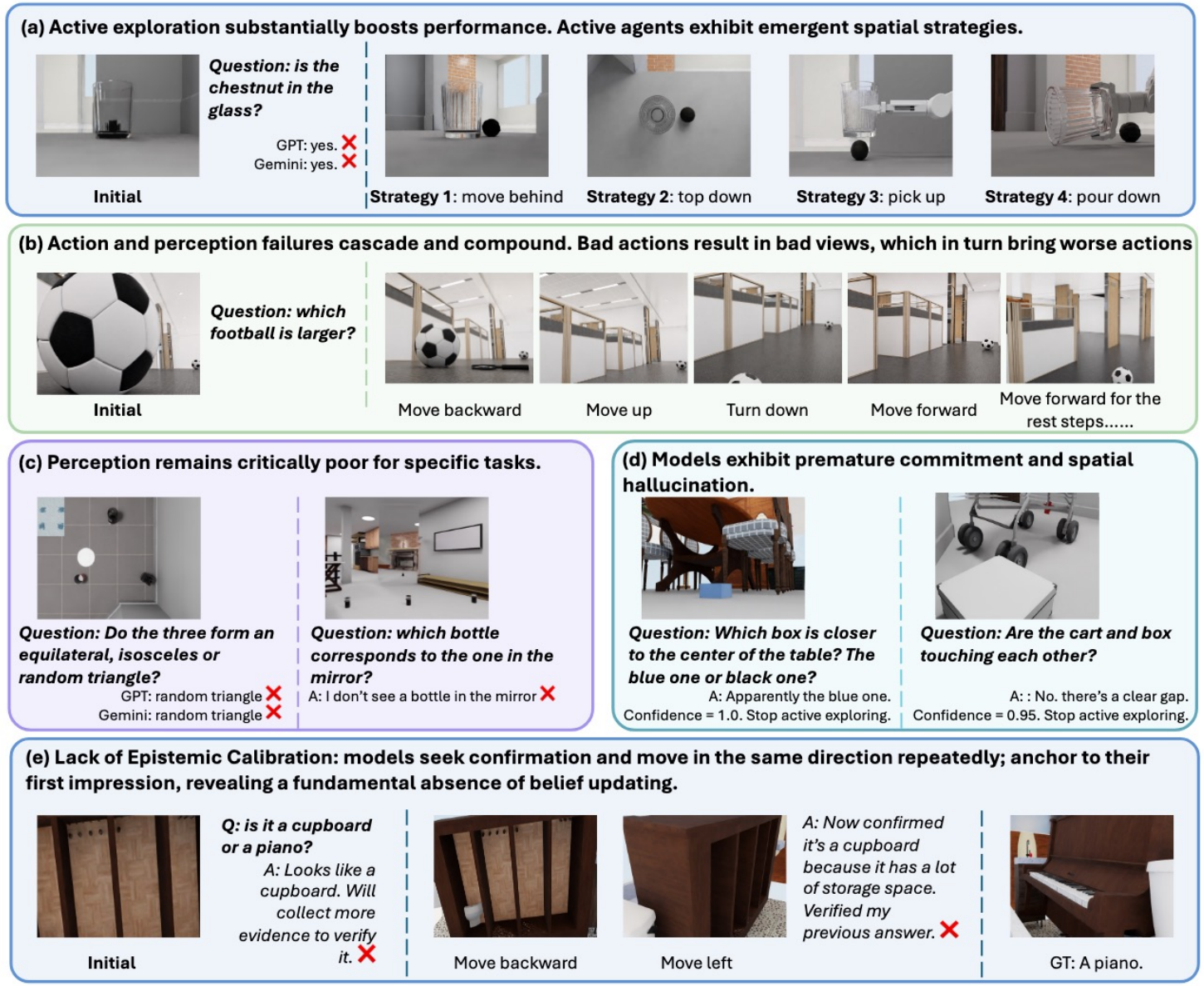
1. Humans far exceed models, even when models are given the right views.

2. Humans show epistemic caution: they gather more evidence, seek viewpoints that could falsify their hypothesis, and lower confidence under ambiguity.

3. Models commit prematurely and overconfidently and repeat the same direction — gathering redundant, not informative, views.

4. The core gap is epistemic calibration — anchoring to first impressions without belief revision — which is distinct from perception and not closed by a better encoder or more steps.

# Qualitative Study



# 04

# Conclusion & Takeaways

# What this means for embodied AI

## 01

### Action for perception

The frontier is not better passive perception, but learning which actions yield information.

## 02

### Metacognition matters

Agents need to know when they don't know — to seek falsifying views and revise beliefs.

## 03

### A testbed, not a fix

ESI-Bench measures the gap; closing it likely needs new training objectives, not just scale.

Benchmarks that **close the perception–action loop** turn ‘spatial reasoning’ into a problem about acting under uncertainty.

# Limitations & open questions

### What ESI-Bench does not settle

- Evaluation is simulation-based (OmniGibson); sim-to-real transfer of these gaps is untested.
- Agents are prompted MLLMs — results may shift with a different agent harnessing.
- “Active” is largely viewpoint selection; long-horizon manipulation is light.

**Open questions**

How do we **train** — not just measure — falsification-seeking and belief revision?

Is benchmarking enough, or do we need intrinsic objectives for calibration?

Where should the “which action is informative?” signal come from — reward, world model, or prior?

# Discussion Points

If a model could choose its next view to **disprove** its current belief — would that alone close most of the gap?

- Where should the ‘which action is informative?’ signal come from — reward, world model, or prior?
- Does explicit 3D help only when near-perfect? What is the break-even reconstruction quality?
- What would a training objective for ‘knowing when you've seen enough’ look like?

# Thank you