



MOTIVATION

Reasoning about yourself in space

Human perception is inherently *situated* — continuously relating ourselves to the physical world and reasoning over actions from our **own viewpoint, pose, and motion**.

Environment-centric

Relations among objects in a scene — the focus of prior benchmarks.

Observer-centric

Reasoning relative to the agent's viewpoint, position & motion.

THE QUESTION

Do multimodal models understand **themselves in space** — where they are, how they've moved, and what they can do given the physical constraints?

THE BENCHMARK

Six situated-awareness tasks

 **Self-Localization** — where am I?

 **Route Shape** — shape of my path?

 **Spatial Memory** — what changed?

Relative Direction

"From my viewing point at the end of the video, where am I located at the beginning of the video?"

A. Same location
B. Front left
C. Back right
D. Front right

Spatial Affordance

"Can I touch the sun chair to my right using only arm movement, without any body or position change?"

A. Yes
B. No

Route Shape

"What's the shape of my moving trajectory?"

A. L-shape. B. U-shape. C. Circle. D. Rectangle

Localization

"Am I located at the corner, along the side, or near the center of the lawn?"

A. At the center
B. Along the side
C. Near the center

Scene: Courtyard Commons

"Which object changes between earlier and later in the video?"

A. Backpack B. Fire hydrant
C. Sun chair D. Patio umbrella

Reverse Route Plan

"From my viewpoint at the end of the video, how can I go back to my starting point?"

A. Turn around, go straight. Turn *right*, go straight, then turn *right* and continue straight.
B. Turn around, go straight. Turn *right*, go straight, then turn *left* and continue straight.
C. Turn around, go straight. Turn *left*, go straight, then turn *right* and continue straight.
D. Turn around, go straight. Turn *left*, go straight, then turn *left* and continue straight.

 **Relative Direction** — way to where I was?

 **Reverse Route Plan** — how to get back?

 **Spatial Affordance** — can I act now?

THE BIG PICTURE

Camera motion $\neq$ observer motion



INSIDE SAW-BENCH

Real egocentric video, human-annotated

Self-recorded with **Ray-Ban Meta (Gen 2)** smart glasses; audio removed so reasoning is purely visual.



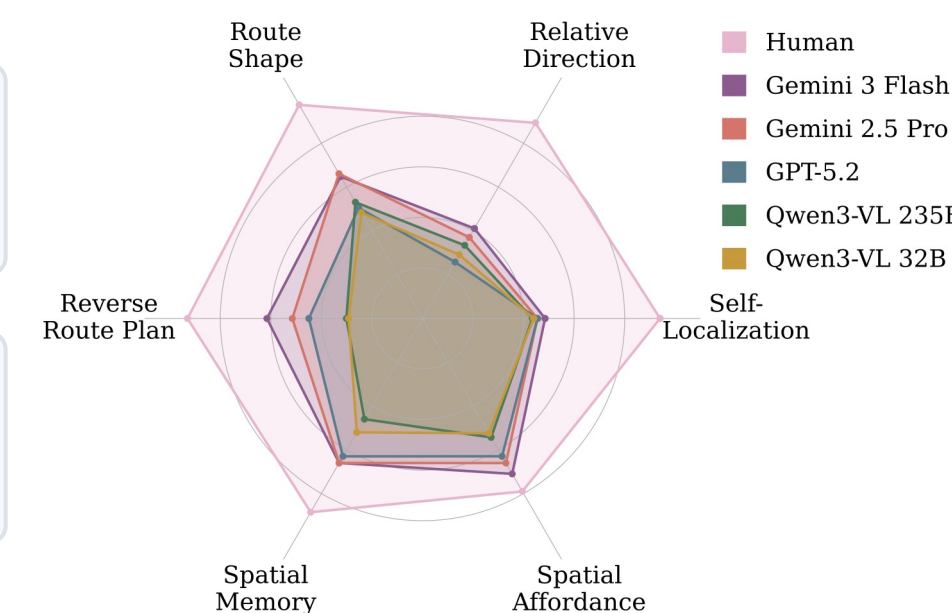
Courtyards · Parking lots · Lawns & plazas · Lecture halls · Classrooms · Rec rooms · Households

786
real-world videos

6
awareness tasks

2,071
human QA pairs

24
models evaluated



FINDING 1 · ROTATION & EGOMOTION

Where you look $\neq$ where you go



60.0%

Gemini 3 Flash
misreads as zigzag

53.3%

Qwen3-VL 235B
misreads as zigzag

FINDING 2 · ERROR ACCUMULATION

Error compounds as the path turns

Models	Straight	Single Turn	Two Turns
Human	100.00	96.67 (-3.33%)	90.00 (-10.00%)
Gemini 3 Flash	73.33	70.69 (-3.60%)	40.61 (-44.63%)
Gemini 3 Pro	63.33	56.90 (-10.16%)	36.46 (-42.44%)
Gemini 2.5 Pro	73.33	55.17 (-24.76%)	33.41 (-54.44%)
GPT-5.2	30.00	39.66 (+32.20%)	22.49 (-25.03%)
Qwen3-VL 235B	90.00	8.62 (-90.42%)	27.85 (-69.06%)
Qwen3-VL 32B	80.00	12.07 (-84.91%)	21.83 (-72.71%)

FINDING 3 · OBJECT PERMANENCE

Out of view is read as gone



Models describe each frame correctly yet treat **non-visibility as non-existence** — they lack a persistent world-state.

FINDING 4 · SCENE OPENNESS

Openness $\neq$ difficulty

