BEARING: A BENCHMARK FOR EGOCENTRIC–ALLOCENTRIC CORRESPONDENCE IN MAP-GUIDED INDOOR NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Map-guided navigation draws on the spatial intelligence of multimodal models, since a model must relate an allocentric map to its egocentric view and act on the result. Although maps tailored to a particular system improve its navigation, they are judged only by that navigation, so a general model's ability to relate map and view has received little systematic study. We introduce BEARING, a benchmark that tests this ability through scale, pose and identity correspondences and pairs each test with navigation from the same pose. Its 5 task categories share 60 routes in 9 Matterport3D buildings, from closed-loop navigation of a route to open-loop planning of its segments and static questions at their starts. We evaluate 9 closed-source, open-source and embodied models and find that all remain well below a human reference, with the widest gap in self-localisation. They lose directions rather than distances, in estimating their own heading, reading bearings on the map or converting them into their own frame, and only the strongest models read the map's compass and agent marker as intended. Relating the tasks to one another, we find that identity tests predict the progress of the segment that starts at their pose but not of other segments on the same route, so a navigation failure can be diagnosed with tests at the pose where it occurs.

1 INTRODUCTION

As multimodal models are built into embodied agents, spatial intelligence has become a central capability to evaluate (BAAI RoboBrain Team et al., 2025; NVIDIA, 2026; Yang et al., 2025; Yin et al., 2025). Map-guided navigation is a common application that relies on a specific form of this capability, relating an allocentric map to the egocentric view, and requires acting on the result in an unfamiliar environment (Zhang et al., 2024; Chen et al., 2024b; Zhang et al., 2025; Chen et al., 2026). Although maps tailored to a particular system improve its navigation (An et al., 2022; Georgakis et al., 2022; Wang et al., 2025b), they are judged only by that navigation, so a general model's ability to relate map and view has received little systematic study.

Following classic spatial cognition theory (Klatzky, 1998), a map is defined as an allocentric representation that relates things to one another, and a view as an egocentric representation that relates them to the observer. Human navigation with a map relies on scale, pose and identity correspondences between the two, which convert distances, fix the navigator's pose on the map and match shared objects. Current evaluations of multimodal models seldom take this perspective, so these correspondences remain largely unexamined. Spatial reasoning benchmarks mostly evaluate map understanding (Li et al., 2024; Xing et al., 2025; Rodionov et al., 2025) and view understanding (Yang et al., 2025; Li et al., 2025a; Zhu et al., 2026) in isolation, and those that give both test the correspondences only in static questions (Zhou et al., 2025; Ramakrishnan et al., 2025; Shen et al., 2026; Shin et al., 2026). Map-guided navigation benchmarks give the model both a map and a view (Li et al., 2025b; Chen et al., 2026), but they score only the outcome of each episode and at most isolate localisation by substituting the true pose. They thus take the correspondences largely for granted rather than test them, so a failed episode cannot show which correspondence broke. To our knowledge, no benchmark tests each correspondence with map and view together and links it to navigation at the poses where it is used (Table 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with related benchmarks. #Items counts questions or episodes. SPACE includes text-only variants, and for Video2Mental we take the 2,650 test items of its Section 3.1 not the 2,300 of its introduction. View gives the egocentric input per question or navigation step. Correspondence marks the map–view correspondences tested on their own, where layout matches views to a whole map and implicit marks correspondences that are not tested on their own and enter only through the navigation outcome. Navigation marks routes executed in a simulator, and Paired marks correspondence tests at the poses of the navigation episodes. Abs. = abstract, Photo. = photometric.

Benchmark	#Items	Input		Correspondence			Navigation		Paired
		Map	View	Scale	Pose	Identity	Closed	Open	
Spatial reasoning benchmarks									
TopViewRS (Li et al., 2024)	11.4K	Photo., abs.	–	–	–	–	–	–	–
MapBench (Xing et al., 2025)	1.6K	Abs.	–	–	–	–	–	–	–
VSI-Bench (Yang et al., 2025)	5.1K	–	Vid.	–	–	–	–	–	–
Video2Mental (Zhu et al., 2026)	2.7K	–	Vid.	–	–	–	–	✓	–
UrBench (Zhou et al., 2025)	10.5K	Photo.	Img.	–	✓	✓	–	–	–
SPACE (Ramakrishnan et al., 2025)	2.4K	Abs.	Vid., img.		layout		✓	–	–
ArchSIBench (Shen et al., 2026)	3.0K	Abs.	Img.	–	✓	–	–	–	–
m2sv (Shin et al., 2026)	10.0K	Photo.	Img.	–	✓	–	–	–	–
Map-guided navigation benchmarks									
FloNa (Li et al., 2025b)	0.5K	Abs.	Img.		implicit		✓	–	–
FloorPlan-VLN (Chen et al., 2026)	6.7K	Abs.	Img.		implicit		✓	–	–
BEARING (ours)	1.6K	Photo.	Img.	✓	✓	✓	✓	✓	✓

To fill this gap, we introduce **BEARING** (a **B**enchmark for **E**gocentric–**A**llocentric **R**easoning in **I**ndoor **Navig**ation), 9 tasks in 5 categories built on 60 annotated routes in 9 Matterport3D buildings (Figure 1). In BEARING, every egocentric view is paired with a top-down map of the building that carries room labels, a scale bar and a compass. The 5 categories are layered to evaluate this ability systematically, from single correspondences up to full navigation. *Scale*, *Pose* and *Identity* are static correspondence tests, whose questions each probe one correspondence at the pose where a route segment starts. *Transition* then tests an intermediate stage of navigation, in which the model plans all actions of a segment from its start in a single answer. *Navigation* finally follows the whole route and updates the map and view after each set of actions. The shared routes pair each of these questions with the Transition segment that starts at its pose and each segment with the route that contains it, and the construction extends to any annotated route.

We evaluate 9 closed-source, open-source and embodied models, and all remain well below a human reference, most of all on Pose, where humans reach 97.50 and the best model 72.75. First, models lose directions rather than distances, at 3 distinct steps: estimating their own heading, reading a bearing on the map, and converting it into their own frame. Second, identity tests predict the progress of the segment that starts at their pose but not of other segments on the same route (mean $r = 0.19$ against 0.00). Third, a text map that states directions relative to the agent lets 4 weaker models progress 10 to 20 points further along a segment, while only the 2 strongest of the 8 models tested change their answers as edits to the scale bar, compass and agent marker require.

Our contributions are as follows.

- **BEARING**, a benchmark that tests the scale, pose and identity correspondences between map and view one by one at the poses where navigation uses them, and an evaluation of 9 models on it against a human reference (Sections 3, 4.1 and 4.2).
- A diagnosis of what models read from a map and how they relate it to the view, showing that models lose directions rather than distances, at steps that differ across models, which a map given in their own frame partly remedies, and that only the strongest make full use of the scale bar, compass and agent marker (Sections 4.3 and 4.5).
- A pose-level analysis showing that identity tests predict the progress of the segment that starts at their pose but not of other segments on the same route, so a navigation failure can be diagnosed with tests at the pose where it occurs (Section 4.4).

2 Related Work

2.1 Spatial Reasoning and Map-to-View Benchmarks

Table 1 compares BEARING with representative spatial reasoning and map-guided navigation benchmarks. Most spatial reasoning benchmarks give a single representation, either a map (Li et al., 2024; Xing et al., 2025; Rodionov et al., 2025) or egocentric observations (Fu et al., 2024; Yang et al., 2025; Li et al., 2025a; Zhu et al., 2026), and so leave the relation between the two untested. Those built on egocentric views further show that models are brittle when a question requires reasoning in a frame other than the camera's (Li et al., 2025a; Wang et al., 2026a). Benchmarks that give both a map and a view relate them only in static questions, each covering one or 2 correspondences. UrBench and related cross-view benchmarks ask for orientation, localisation and object matching between satellite and ground or aerial views (Zhou et al., 2025; Zheng et al., 2026; Liu et al., 2026), SPACE (Ramakrishnan et al., 2025) asks which top-down sketch matches a video walkthrough, ArchSIBench (Shen et al., 2026) matches a viewpoint on a floor plan to a photograph, and m2sv (Shin et al., 2026) asks for the camera's heading on an overhead map. Only SPACE also lets the model navigate, and it does so from either the view or a local map rather than both, while Video2Mental executes a plan written from video without a map (Ramakrishnan et al., 2025; Zhu et al., 2026). BEARING tests the scale, pose and identity correspondences on a photometric map of the building the agent is in and pairs each test with navigation that uses map and view together.

2.2 Map-Guided Navigation

Map-based VLN systems mostly build their map from the agent's own observations, as topological graphs (Chen et al., 2024b; An et al., 2022) or semantic and bird's-eye-view maps (Georgakis et al., 2022; Zhang et al., 2025; Lian et al., 2026), so the map arrives registered to the agent's pose. Many also convert it for the model. CM2 predicts its map with the agent at the centre facing up (Georgakis et al., 2022), SeeNav-Agent rotates its bird's-eye view to the agent's heading (Wang et al., 2025b), MapNav labels key regions in text (Zhang et al., 2025), and MapGPT names each candidate viewpoint by its egocentric direction (Chen et al., 2024b). Floor-plan navigation instead supplies a map made in advance (Li et al., 2025b; Huang et al., 2026; Gao et al., 2026; Chen et al., 2026). FloorPlan-VLN still draws the agent's ground-truth pose on the plan and notes that current agents lack the ability to comprehend floor plans (Chen et al., 2026), and floor-plan localisation methods recover the pose with dedicated models (Min et al., 2022; Chen et al., 2024a; Grader & Averbuch-Elor, 2025). These systems and their benchmarks judge a map by the navigation it enables, with ablations that remove, mask or perturb it, so the correspondences stay implicit (Table 1). BEARING instead asks how well general models establish the scale, pose and identity correspondences themselves, at the poses where navigation uses them.

3 Benchmark

BEARING pairs every egocentric view with an allocentric top-down map of the building, which carries room labels, a scale bar and a compass (Figure 1). Outside the Pose category, the map also marks the navigator's position and the prompt states its heading, so no other task requires self-localisation. Appendix A lists the prompts and answer formats of all 9 tasks.

3.1 Task Categories

Navigation evaluates the integrated ability to follow a whole route with the map. It includes the *Nav-Text* task, in which the model follows a route instruction that names the rooms or landmarks to pass and the place to stop. The model interacts over several rounds, each giving the current view and the map with its position and heading and asking for a few 0.25 m moves and 15° turns. Since the view and map are updated after every round, the model navigates in a closed loop and can correct its course until it stops or runs out of budget.

Transition evaluates the ability to plan one segment of a route from the map and view at its start. It includes the *Transition* task, which asks the model to follow an instruction for the segment (Figure 1) and to give all its actions in one answer. Transition thus bridges the static categories and Navigation,

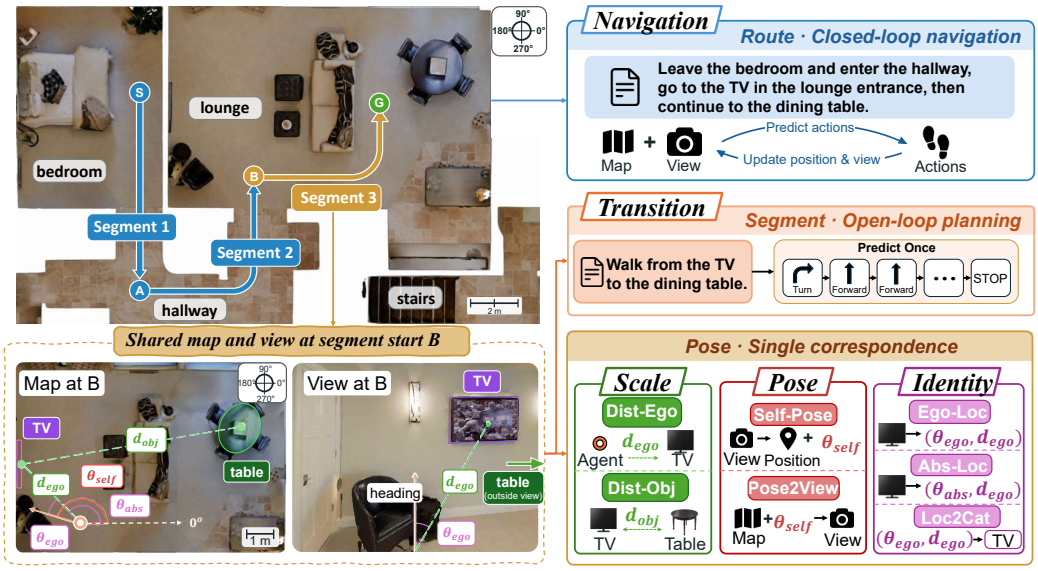


Figure 1: **Overview of BEARING benchmark.** Given a route from start S to goal G, we split it into 3 segments at turning point A and entry point B. The 5 task categories are then constructed at different spatial scales: Navigation follows the whole route, while Transition focuses on a single route segment and the other 3 static task categories are defined at the start of an individual segment. These tasks therefore share the same local state, including the map and egocentric view. Here, at B, the agent is d_{ego} from the TV, seen only in the egocentric view, and the TV is d_{obj} from the table, seen only on the allocentric map, so relating them requires reasoning over both modalities. The agent's heading is θ_{self}, and the TV has a relative bearing θ_{ego} with respect to the agent's heading and an absolute bearing θ_{abs} on the global compass.

since it is answered open loop from a single pose like a static question but requires actions toward a goal like Navigation.

Scale evaluates whether the model can carry metric distances between map and view, using the scale bar of the map and the stated camera model of the view. It includes the *Dist-Ego* task, which asks for the distance in metres from the navigator to a named object, and the *Dist-Obj* task, which asks for the distance between 2 named objects.

Pose evaluates self-localisation, relating the view to a position and heading on the map. It includes the *Self-Pose* task, which asks from which of 6 nearby candidate positions on the map the view was taken and which heading it faced, and the *Pose2View* task, which marks a position and heading on the map and asks which of 6 views was taken there.

Identity evaluates object matching, finding the same object on the map and in the view. It includes the *Ego-Loc* and *Abs-Loc* tasks, which ask for the bearing and distance of a named object relative to the navigator's heading and on the map, respectively, and the *Loc2Cat* task, which gives a bearing and distance and asks which category of object lies there.

3.2 Benchmark Construction

Route Construction. To explore the relationship between different task categories, all task samples in BEARING are built on 60 carefully designed routes in 9 of the 11 Matterport3D buildings (Chang et al., 2017) of the R2R val-unseen split (Anderson et al., 2018), whose annotations span various distinct rooms and object categories (detailed in Appendix A.1). We chose the start and goal of each route, and the shortest path between them is generated automatically. Each path either crosses into another room or passes a landmark that an instruction can name.

Task Generation. We generate all five task categories from a single traversal of each route. Specifically, each route is first partitioned into segments according to structural cues, including landmarks, turns, and room boundaries. Instructions are then automatically generated for both the full route and individual segments based on their start and end locations, as well as the landmarks encountered.

For each segment, we recover its start state by replaying the reference actions from the beginning of the route, ensuring that every segment starts from a pose actually visited during navigation. We employ Habitat (Savva et al., 2019) simulator to execute actions and render egocentric observations and Open3D (Zhou et al., 2018) to render top-down maps augmented with room labels from the Matterport3D region annotations. At each segment start pose, target objects are selected and described using their semantic category and room annotation, and task-specific answers are computed directly from the underlying metric geometry. This generation pipeline ensures that all static tasks associated with a segment share the same pose, visual observation, and map context. As a result, the Scale, Pose and Identity tasks can be naturally paired with Transition and Navigation at exactly the same physical state, enabling consistent cross-task comparison.

Quality Validation. To improve the reliability of the benchmark, we adopt a three-stage quality control procedure. Firstly, we apply automatic filtering to remove ambiguous samples with multiple valid answers and to discard low-quality egocentric views dominated by nearby walls or other large surfaces. Then, all generated instructions are then manually inspected and revised when necessary. Finally, we manually verify the generated samples together with their ground-truth targets, removing 20 static samples with incorrect annotations, most of which correspond to doorless passages mistakenly labeled as doors. This process results in 1,418 validated static samples (Appendix A.2).

3.3 Evaluation Protocol

Evaluation Metrics. Every item is scored in $[0, 1]$, and an unparsable answer scores 0 (Appendix B.3). Nav-Text is scored by sDTW (Ilharco et al., 2019) with a 3 m success radius (Anderson et al., 2018), and Transition by goal progress (Datta et al., 2021), the fraction of the geodesic distance to the goal that the actions remove. Numeric answers are scored by mean relative accuracy (MRA) (Yang et al., 2025), with the bearing and distance of an Ego-Loc or Abs-Loc answer scored as the position they give. Choices are scored by accuracy, with Self-Pose also requiring a heading within 30° and Loc2Cat balanced over 6 category groups. A task score is the mean over its items, and a category score is the unweighted mean of its tasks (Appendix B.1).

Normalization. Each category score is normalised to put its chance floor at 0 and a perfect score at 100, with a score below chance counted as 0. The floor of a static task is the expected score of a uniform guess over its answer space, and that of Nav-Text and Transition is the score of a prompt-blind random walk from the true start (Appendix B.2). A category's floor is the mean of its task floors, and **Overall** is the unweighted mean of the 5 normalised category scores.

4 Experiments

4.1 Setup

Models. We evaluate 9 multimodal large language models (MLLMs) in 3 groups. *Closed-source general* models are Gemini-3.8-Flash, GPT-5.6 Sol and Claude-Opus-5 (Google DeepMind, 2026; OpenAI, 2026; Anthropic, 2026), run with reasoning at each vendor's medium setting. *Open-source general* models are Qwen3.8-27B, Gemma-4-31B and InternVL3.5-38B (Qwen Team, 2026; Team et al., 2026; Wang et al., 2025a). *Embodied foundation models* are Hy-Embodied-A3B, Cosmos-Reason2-32B and RoboBrain2.0-32B (Wang et al., 2026b; NVIDIA, 2026; BAAI RoboBrain Team et al., 2025). The 6 open models are served locally with their official reasoning settings. Hy-Embodied-A3B is the only model that documents training on route-following vision-and-language navigation, collected on R2R-CE training scenes that do not overlap our val-unseen buildings. Cosmos-Reason2-32B is post-trained from Qwen3-VL-32B-Instruct (Bai et al., 2025) on physical-reasoning video, and RoboBrain2.0-32B targets general embodied reasoning, including simulated object search with high-level navigation actions. No model documents training on top-down maps of indoor scenes, although the RoboBrain2.0 report shows the model pointing to a navigation goal on such a map. Every model is run once on the same 1,418 static items, 60 Nav-Text episodes and 119 Transition segments, with the settings listed in Appendix B.4.

Baselines. Every score is reported against a chance floor and a human ceiling. The floor, used for normalisation, is the expected score of a uniform guess over the answer space for static tasks and of a prompt-blind random walk for Nav-Text and Transition (Section 3.3, Appendix B.2). The ceiling comes from 5 annotators unfamiliar with the benchmark, who answer 30 Nav-Text episodes

Table 2: Results with the map and the egocentric view, in %. Category columns are raw scores, and **Overall** is the mean of the 5 category scores after chance normalisation (Section 3.3). Best model in **bold**, second underlined. Human is the mean score of 5 annotators on a 430-item subset, and Chance is the per-task floor (Section 4.1).

	Correspondence			Navigation		
Model	**Scale**	**Pose**	**Identity**	**Transition**	**Nav**	**Overall**
Closed-source general MLLMs						
Gemini-3.8-Flash	**64.75**	**72.75**	**55.98**	**64.52**	**68.75**	**60.87**
GPT-5.6 Sol	63.19	52.18	55.25	59.52	45.76	49.87
Claude-Opus-5	52.11	61.65	44.88	49.68	58.10	47.12
Open-source general MLLMs						
Qwen3.8-27B	59.47	47.09	43.57	48.26	43.31	42.20
Gemma-4-31B	47.35	32.28	28.11	41.00	33.27	28.80
InternVL3.5-38B	41.25	14.81	14.04	15.69	1.25	8.15
Embodied foundation models						
Hy-Embodied-A3B	46.71	18.93	18.22	20.63	11.62	14.37
Cosmos-Reason2-32B	39.99	13.83	11.63	13.87	2.69	7.16
RoboBrain2.0-32B	32.55	12.38	11.35	19.75	4.09	6.05
Human	76.75	97.50	74.22	70.97	84.14	77.92
Chance	23.93	9.51	2.15	15.81	2.46	0.00

and 50 items from each other task, 430 in total, with the same inputs and scoring pipeline as the models. Each item is answered by at least 3 annotators, whose scores are averaged. On this subset, the Overall scores of the models differ from their full-set scores by at most 5.6 points and are lower on average, so the subset does not inflate the gap to the human (Appendix B.5).

4.2 Main Results

Table 2 summarises model performance on BEARING and reveals 3 key findings.

(i) There is a clear performance hierarchy across model groups. Closed-source general MLLMs perform best and take the top 2 places in every category, followed by open-source general MLLMs, while the embodied foundation models remain substantially weaker across nearly all categories, although InternVL3.5-38B scores below Hy-Embodied-A3B. This indicates that specialised training for embodied tasks does not carry over to map-guided navigation in our setting, and the embodied models reach similar Nav-Text scores under their own inference settings (Appendix B.6).

(ii) Pose exposes a particularly large gap between humans and models. Unlike Scale and Identity, which test individual spatial relations, Pose requires the model to align the egocentric view with both a position and a heading on the allocentric map. Humans nearly saturate this category at 97.50%, whereas the strongest model reaches 72.75%. This suggests that robust self-localisation from cross-modal spatial evidence remains challenging.

(iii) The gap between spatial correspondence and effective navigation widens for weaker models. Several weaker models, such as InternVL3.5-38B and Cosmos-Reason2-32B, retain non-trivial performance on Scale and Identity but fall close to chance on Transition and Navigation. In contrast, stronger models do not collapse from Transition to Navigation, and Gemini-3.8-Flash and Claude-Opus-5 even score higher on Nav than on Transition. These results suggest that current models differ not only in whether they can recover spatial relations from the allocentric map and the egocentric view, but also in whether they can reliably turn those relations into sequential actions.

4.3 Error Analysis

We trace the failures of Section 4.2 to the step at which the map–view correspondence breaks, with 1 diagnostic per category (Table 3). (1) Distances survive the move from the view to the map, as most models lose fewer than 10 points when the target is only on the map (Table 3a). (2) Headings are picked rather than estimated, as all models except Gemini and Claude give 67–100% of their Self-Pose headings on the 4 map axes against 5% of the true headings (Table 3b), which explains their

Table 3: Where the map–view correspondence breaks. View marks targets covering at least 300 pixels of the view and map targets absent from it. (a) Dist-Ego score. (b) Share of Self-Pose headings on the 4 map axes, against 5% of the true headings. (c) Median bearing errors. Abs-Loc reads the bearing on the map, and Ego-Loc also converts it into the agent's frame, so for map targets the gap between the 2 estimates the error added by conversion. †Flipping the sign of each answer lowers the error to 27° and 45°, that is, left and right are mirrored. The best model per column is in **bold**. The columns use 109, 31, 206, 33, 38 and 103 items, the same for every model.

	(a) Scale		(b) Pose	(c) Identity		
	Dist-Ego score		Self-Pose heading	Abs-Loc	Ego-Loc	
Model	view (%) ↑	map (%) ↑	on axes (%) ↓	map (°) ↓	map (°) ↓	view (°) ↓
Closed-source general MLLMs						
Gemini-3.8-Flash	69.9	**62.9**	42	**6**	20	**4**
GPT-5.6 Sol	**70.0**	56.1	79	7	**12**	6
Claude-Opus-5	56.1	39.4	**31**	28	29	6
Open-source general MLLMs						
Qwen3.8-27B	66.5	60.6	89	19	41	8
Gemma-4-31B	53.0	53.5	92	15	78†	18
InternVL3.5-38B	41.0	53.5	100	63	70	50
Embodied foundation models						
Hy-Embodied-A3B	45.9	45.2	84	29	92†	17
Cosmos-Reason2-32B	36.2	38.1	99	71	82	32
RoboBrain2.0-32B	32.9	28.7	67	66	69	37

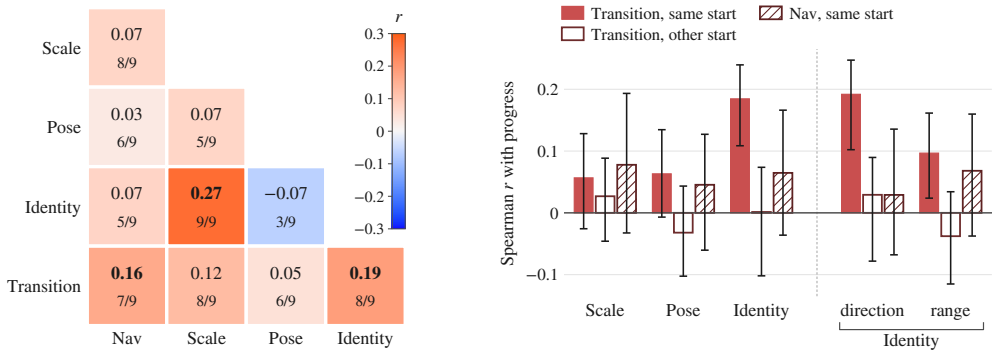
low Pose scores. (3) Bearings are lost in reading the map or in converting it into the agent's frame. For map-only targets, Gemma-4-31B, Hy-Embodied-A3B and Qwen3.8-27B read the map bearing to within 15–29° (Abs-Loc) but are off by 41–92° from the agent (Ego-Loc), and Gemma-4-31B and Hy-Embodied-A3B mirror left and right (Appendix D.2), whereas InternVL3.5-38B, Cosmos-Reason2-32B and RoboBrain2.0-32B already misread the map by 63–71° (Table 3c). The same steps separate the models of Section 4.2. Only the closed-source models keep bearings from the agent within 12 to 29°, and the embodied models fail at the same steps as the open ones. GPT-5.6 Sol owes its weak Pose to picked headings (79%), and Claude-Opus-5 its weak Scale and Identity to misreading the map. **Distances thus survive the move to the map, but directions are lost, in estimating the agent's own heading, in reading a bearing on the map, or in converting it into the agent's frame.**

4.4 Correlation Analysis

We ask whether the correspondence tests predict navigation. Within each model, we correlate category scores across its 60 routes and average the correlations over the 9 models (Figure 2a). (1) The static tests relate to navigation only through Transition. Transition is the only category that predicts Nav-Text ($r = 0.16$ [0.04, 0.26], positive for 7 of 9 models), and Identity is the only correspondence test that predicts Transition (0.19, 8 of 9). Scale, Pose and Identity do not predict Nav-Text directly (0.03 to 0.07, all CIs include 0). Section 4.5 therefore uses Transition as the test bed for the map. (2) The link holds only at the pose where the test is asked, and runs through direction (Figure 2b). Identity items asked at a segment's own start correlate with its Transition progress (0.19 [0.11, 0.24], positive for all 9 models), but items asked at another start on the same route do not (0.00). The link runs mainly through the object's direction rather than its range (0.17 against 0.09, difference 0.08 [0.00, 0.15]), the component that Section 4.3 found to break. Identity items at the start of a Nav-Text episode likewise fail to predict it, as the episode passes through later poses. **Identity tests thus predict navigation only from the pose where they are asked, so a navigation failure can be diagnosed with tests at its own poses.**

4.5 Map Understanding Analysis

We ask in which form the map helps a model and which of its elements the model reads.

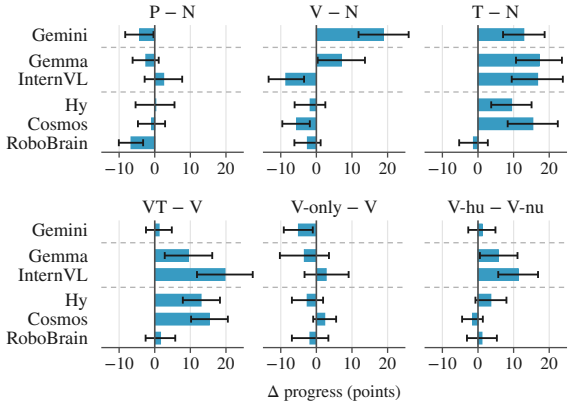


(a) Across routes, within each model (b) Same start vs. another start

Figure 2: Where correspondence scores predict navigation. (a) Spearman correlation between category scores over each model's routes, averaged over the 9 models, with navigation as goal progress. Below each value, k/9 counts the models with a positive correlation, and bold marks a 95% CI that excludes 0. (b) Correlation of correspondence scores with Transition progress for items asked at the segment's own start or at another start on the route, and with Nav-Text progress for items at the route start. Direction and range split the Ego-Loc and Abs-Loc error into its bearing and distance parts (Appendix E). Error bars are 95% CIs from a route bootstrap.

Config	Map input		
	Form	**Crop**	**Orientation**
Baseline			
N	none		
P	blank		
V	image	route	north-up
V-only†	image	route	north-up
Representation			
T	text		
VT	image+text	route	north-up
Viewport			
V-nu	image	agent	north-up
V-hu	image	agent	heading-up

(a) Map configurations



(b) Paired differences in Transition progress

Figure 3: Map form. (a) Map configurations, each changing only the map that comes with the instruction and the egocentric view. V is the input of Table 2, V-only† also withholds the egocentric view, and T gives the map as text that states the bearing and distance of each passage from the agent. (b) Paired differences in Transition progress over the 119 segments, in points, with 95% CIs from a bootstrap over routes. The other models and the absolute scores are in Appendix F.

Map Form. The weaker models are limited by the form of the map, not its content. Using Transition as the test bed (Section 4.4), we rerun its 119 segments under 8 configurations that change only the map (Figure 3a) and compare them in pairs over the same segments (Figure 3b, which plots 6 models since GPT-5.6 Sol and Qwen3.8-27B behave like Gemini-3.8-Flash; Appendix F). The map image (V−N) raises progress by 7 to 19 points for Gemini-3.8-Flash, GPT-5.6 Sol, Qwen3.8-27B and Gemma-4-31B, but not for the other models, and even lowers it for InternVL3.5-38B and Cosmos-Reason2-32B, whereas a blank image (P−N) does not. The same map given as text, with the bearing and distance of each passage from the agent (T−N), helps every model except RoboBrain2.0-32B, by 9 to 17 points. Three further contrasts show what the weaker models miss. A gain from adding the text to the image (VT−V) marks information that the model fails to extract from the image, and reaches 10 to 20 points for Gemma-4-31B, InternVL3.5-38B, Hy-Embodied-A3B and Cosmos-Reason2-32B but nothing for the 3 strongest models. A gain from rotating the map so that the heading points up (V-hu−V-nu) marks a model that cannot align a north-up map with its own heading, and reaches 11 points for InternVL3.5-38B and 6 for Gemma-4-31B but nothing for Gemini-3.8-

Table 4: Change in each model's answer when one map element changes. The scale bar columns give the median answered distance as a multiple of the original answer, the compass columns the share of items whose answered heading turns with the rotated rose when the rose alone rotates, or with the rotated floor when floor and rose rotate together, within $15°$ and pooled over $\pm 90°$, and the marker columns the median shift of the located object along the marker's move in metres. The first row gives the values that reading each element implies, arrows give the better direction for the compass, and the value closest to reading the element is in **bold**. On a repeat of the unchanged item, 3 to 23% of a model's headings turn by $90°$ by chance (Appendix G). Claude-Opus-5 was not run.

	Scale bar *on Dist-Ego*		Compass (%) *on Self-Pose*		Agent marker (m) *on Ego-Loc*		
Model	×0.5	×2	rose ↑	floor ↓	2 m	4 m	8 m
Reading the element	*×0.50*	*×2.00*	*100*	*0*	*2.0*	*4.0*	*8.0*
Closed-source general MLLMs							
Gemini-3.8-Flash	×0.61	×1.75	**72**	**4**	1.6	3.3	6.3
GPT-5.6 Sol	**×0.50**	**×2.00**	48	17	**1.8**	**3.8**	**7.5**
Open-source general MLLMs							
Qwen3.8-27B	×0.82	×1.10	51	22	1.2	2.7	5.2
Gemma-4-31B	×0.87	×1.47	28	29	1.1	1.1	1.0
InternVL3.5-38B	×0.95	×1.00	15	25	0.0	0.4	0.2
Embodied foundation models							
Hy-Embodied-A3B	×0.88	×1.08	48	38	0.5	0.7	0.2
Cosmos-Reason2-32B	×1.00	×1.00	27	26	0.0	0.0	0.1
RoboBrain2.0-32B	×1.00	×1.02	18	20	0.1	−0.1	0.5

Flash. A loss from removing the view (V-only−V) marks information drawn from the view, and appears only for Gemini-3.8-Flash and GPT-5.6 Sol, at 5 and 9 points. **The weaker models thus fail to relate a map image to their own frame, and the map helps them only when it is given in that frame.**

Map Elements. Only the strongest models read the elements that relate the map to the agent. We edit one element at a time and test whether the answer moves as reading it requires (Table 4): we relabel the scale bar on Dist-Ego with the camera height removed from the prompt, rotate the compass on Self-Pose, and move the agent marker on Ego-Loc. Of the 8 models tested, only Gemini-3.8-Flash and GPT-5.6 Sol follow all 3. For the scale bar, GPT-5.6 Sol scales its distances exactly and Gemini-3.8-Flash most of the way, whereas the other models stay within $\times 0.82$ and $\times 1.47$ of their original answers, so their distances rest little on the bar. For the compass, Gemini-3.8-Flash turns its heading with a rotated rose in 72% of items and GPT-5.6 Sol, Qwen3.8-27B and Hy-Embodied-A3B in about half, against at most 12% on a repeat of the unchanged item, but these 3 also turn with a rotated floor in 17 to 38% of items, so they read part of the heading from the image axes. The other 4 models turn with the rose barely more often than on a repeat. For the agent marker, Gemini-3.8-Flash and GPT-5.6 Sol move the located object by most of the offset and Qwen3.8-27B by about two-thirds, whereas the other models move it by about 1 m or less whatever the offset (Appendix G). **Only the strongest models thus read the map image as intended. The others read its compass and agent marker partly or not at all, consistent with how they lose directions (Section 4.3), and benefit from the map only when it is given in their own frame.**

5 Conclusion

We introduced BEARING, which tests the scale, pose and identity correspondences between an allocentric map and an egocentric view one at a time and pairs each test with the navigation that depends on it. All 9 models we evaluated remain well below a human reference, most of all in self-localisation. They lose directions rather than distances, at steps that differ across models. Identity tests predict the progress of the segment that starts at their pose but not of other segments on the same route, so a navigation failure can be diagnosed at the pose where it occurs. A map helps the weaker models only in a form they can relate to their own frame, and only the strongest models read the map's scale bar, compass and agent marker as intended. Because the generator takes any annotated route, the same tests can be generated for other routes.

AI USE STATEMENT

In this work, we used Claude, Claude Code, Codex and GPT for the following tasks that require disclosure, always under the direction of the authors, who made all final decisions. Claude and GPT gave feedback on the research methodology and experimental design, which the authors weighed against their own plans. Claude Code and Codex generated much of the code for the benchmark and evaluation pipeline from the authors' specifications, and assisted in cleaning and reformatting data such as model outputs. Claude assisted with the qualitative analysis and interpretation of results, by checking computed statistics, grouping failure patterns and suggesting how to state the findings. Every conclusion in the paper was checked by the authors against the underlying data.

We did not use generative AI tools to generate synthetic data. The route and segment instructions and all benchmark questions are generated by a rule-based pipeline from the scene annotations, the rooms each route passes and the objects near it, and were then reviewed and revised by hand (Section 3.2). Nor did we use generative AI tools for translation, to develop theoretical models or conceptual frameworks, or to propose or refine hypotheses. Formulating mathematical claims and writing proofs are not applicable to this work.

Additionally, Claude assisted with drafting parts of the paper and editing its language, Claude Code and Codex with writing, modifying and debugging general code, GPT with creating and editing figures, and GPT and Claude with searching and summarising literature. We have reviewed all AI-assisted work. Two authors reviewed all AI-assisted text sentence by sentence, checked every citation against the original source, inspected data samples, tested the code, and reran the experiments whose results are reported. We take responsibility for the final content of this work, including text, claims, code and other artifacts produced with the aid of generative AI.

REFERENCES

Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbert: Multimodal map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385*, 2022.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674–3683, 2018.

Anthropic. System Card: Claude Opus 5. `https://www.anthropic.com/claude-opus-5-system-card`, July 2026.

BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Changan Chen, Rui Wang, Christoph Vogel, and Marc Pollefeys. F 3 loc: fusion and filtering for floorplan localization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18029–18038. IEEE, 2024a.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9796–9810, 2024b.

Kehan Chen, Yan Huang, Dong An, Jiawei He, Yifei Su, Jing Liu, Nianfeng Liu, and Liang Wang. Floorplan-vln: A new paradigm for floor plan guided vision-language navigation. *arXiv preprint arXiv:2603.17437*, 2026.

Samyak Datta, Oleksandr Maksymets, Judy Hoffman, Stefan Lee, Dhruv Batra, and Devi Parikh. Integrating egocentric localization for more realistic point-goal navigation agents. In *Conference on Robot Learning*, pp. 313–328. PMLR, 2021.

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. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Jiang Gao, Xiangyu Dong, Haozhou Li, Haoran Zhao, Yaoming Zhou, and Xiaoguang Ma. Pm-nav: Priori-map guided embodied navigation in functional buildings. *arXiv preprint arXiv:2603.09113*, 2026.

Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15460–15470, 2022.

Google DeepMind. Gemini 3.8 Flash Model Card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-8-Flash-Model-Card.pdf`, September 2026.

Yuval Grader and Hadar Averbuch-Elor. Supercharging floorplan localization with semantic rays. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27116–27125. IEEE, 2025.

Weiqi Huang, Shuangyi Dong, Jiaxin Li, Yifei Guo, Zan Wang, and Wei Liang. Floverse: Floor plan-guided multi-modal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15156–15165, 2026.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*, 2019.

Roberta L Klatzky. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge*, pp. 1–17. Springer, 1998.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1786–1807, 2024.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025a.

Jiaxin Li, Weiqi Huang, Zan Wang, Wei Liang, Huijun Di, and Feng Liu. Flona: Floor plan guided embodied visual navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14610–14618, 2025b.

Guoxin Lian, Shuo Wang, Yucheng Wang, Yongcai Wang, Maiyue Chen, Kaihui Wang, Bo Zhang, Zhizhong Su, Deying Li, and Zhaoxin Fan. Mapdream: Task-driven map learning for vision-language navigation. *arXiv preprint arXiv:2602.00222*, 2026.

Dian Liu, Jie Feng, Di Li, Yuhui Zheng, Guanbin Li, Weisheng Dong, and Guangming Shi. Are vlms lost between sky and space? links ̈2 bench for uav-satellite dynamic cross-view spatial intelligence. *arXiv preprint arXiv:2604.02020*, 2026.

Zhixiang Min, Naji Khosravan, Zachary Bessinger, Manjunath Narayana, Sing Bing Kang, Enrique Dunn, and Ivaylo Boyadzhiev. Laser: Latent space rendering for 2d visual localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11112–11121. IEEE, 2022.

NVIDIA. Cosmos-Reason2: An open reasoning vision-language model for physical AI and robotics. `https://docs.nvidia.com/cosmos/latest/reason2/index.html`, 2026.

OpenAI. GPT-5.6 System Card. `https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf`, July 2026.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? In *International Conference on Learning Representations*, volume 2025, pp. 101439–101498, 2025.

Fedor Rodionov, Abdelrahman Eldesokey, Michael Birsak, John Femiani, Bernard Ghanem, and Peter Wonka. Floorplanqa: A benchmark for spatial reasoning in llms using structured representations. *arXiv preprint arXiv:2507.07644*, 2025.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9339–9347, 2019.

Qirui Shen, Wenda Wang, Jiachen Lu, Zilong Huang, Jin Bai, Lei He, Hongxuan Chen, and Weixin Huang. Archsibench: Benchmarking the architectural spatial intelligence of vision-language models. *arXiv preprint arXiv:2605.20837*, 2026.

Yosub Shin, Michael Buriek, and Igor Molybog. m2sv: A scalable benchmark for map-to-street-view spatial reasoning. *arXiv preprint arXiv:2601.19099*, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Hengyi Wang, Ruiqiang Zhang, Chang Liu, Guanjie Wang, Zehua Ma, Han Fang, and Weiming Zhang. Allocentric perceiver: Disentangling allocentric reasoning from egocentric visual priors via frame instantiation. *arXiv preprint arXiv:2602.05789*, 2026a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Zhengcheng Wang, Zichuan Lin, Yijun Yang, Haobo Fu, and Deheng Ye. Seenav-agent: Enhancing vision-language navigation with visual prompt and step-level policy optimization. *arXiv preprint arXiv:2512.02631*, 2025b.

Ziyi Wang, Xumin Yu, Yongming Rao, Yonggen Ling, Yunheng Li, Oran Wang, Mingqi Gao, Yuchen Zhou, Yves Liang, Zuyan Liu, et al. Hy-embodied-vlm-1.0: Efficient physical-world agents. *arXiv preprint arXiv:2607.12894*, 2026b.

Shuo Xing, Zezhou Sun, Shuangyu Xie, Kaiyuan Chen, Yanjia Huang, Yuping Wang, Jiachen Li, Dezhen Song, and Zhengzhong Tu. Can large vision language models read maps like a human? *arXiv preprint arXiv:2503.14607*, 2025.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13032–13056, 2025.

Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035*, 2024.

Yushuo Zheng, Jiangyong Ying, Huiyu Duan, Chunyi Li, Zicheng Zhang, Jing Liu, Xiaohong Liu, and Guangtao Zhai. Geox-bench: Benchmarking cross-view geo-localization and pose estimation capabilities of large multimodal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13485–13493, 2026.

Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10707–10715, 2025.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018.

Qihui Zhu, Shouwei Ruan, Xiao Yang, Hao Jiang, Yao Huang, Shiji Zhao, Hanwei Fan, Hang Su, and Xingxing Wei. Mind over space: Can multimodal large language models mentally navigate? *arXiv preprint arXiv:2603.21577*, 2026.

A Benchmark Details

Table 5 summarises the 9 tasks of Section 3.1.

Table 5: The 9 tasks. Static questions give the egocentric view at the start pose and the route's map with the agent's position marked, except that Self-Pose marks 6 candidate positions and Pose2View offers 6 candidate views. The prompt states the camera model, the map's conventions and, except in Self-Pose, the agent's heading.

Category	Task	Given	Answer	Metric
Navigation	Nav-Text	route instruction; view and map at each turn	actions, turn by turn	sDTW
Transition	Transition	segment instruction; start view and map	the whole action sequence	progress
Scale	Dist-Ego	a named object	its distance	MRA
	Dist-Obj	2 named objects	their distance	MRA
Pose	Self-Pose	a view; 6 positions on the map	position and heading	accuracy
	Pose2View	a marked pose; 6 candidate views	the matching view	accuracy
Identity	Ego-Loc	a named object	bearing from heading, distance	MRA
	Abs-Loc	a named object	map bearing, distance	MRA
	Loc2Cat	a bearing and distance	the object's category	accuracy

A.1 Routes, Segments and Questions

The R2R val-unseen split covers 11 buildings, of which we leave out 2. Counting MPcat40 objects without building elements such as walls, doors and windows, pLe4wQe7qrG has 9 objects, and

both of its 2 regions are labelled as other rooms. Z6MFQCViBuw has 12 object categories by this count, against 15 to 26 in the retained buildings. Three quarters of its objects are pictures, chairs and curtains, and 17 of its 22 regions are hallways or lounges. Start and goal were chosen by hand rather than sampled, because some randomly sampled endpoints admit no natural instruction. The 60 routes lie in 9 buildings (3 to 13 per building). Their shortest paths from start to goal have a median of 6.3 m (interquartile range 5.0–8.2 m, range 1.7–12.2 m), and on every route the reference actions stay within 3% of this length. Of the 60 routes, 51 cross into another room, and the 9 that stay in one room pass a landmark such as a fireplace, a staircase or a door. The 119 Transition segments come from 59 routes, 48 of which are split into 2 segments, 6 into 3 and 5 left whole; a segment's shortest path has a median of 3.0 m (2.3–4.0 m, range 1.0–9.0 m) and its reference action sequence a median of 19 actions (14–25, range 7–51), within the 60-action limit of the task. The 1,418 static questions comprise 249 Dist-Ego, 239 Dist-Obj, 206 Self-Pose, 206 Pose2View, 234 Ego-Loc, 184 Abs-Loc and 100 Loc2Cat questions. Every static question is anchored at the start pose of a segment; a start pose carries 1 to 7 questions of a task, for different target objects or candidate sets. A further 3 segment starts carry static questions but are not among the 119 scored Transition segments, and 2 of the 119 carry none.

A.2 Question Generation and Validation

Referents are object instances of the Matterport3D semantic annotation, named by object category and room and, where a room holds several of a kind, by a neighbouring object ("the door next to the chair in the living room"). For each question the generator records whether its target is visible in the anchor view (at least 300 pixels) or appears only on the map, which Section 4.3 uses. Relative bearings are positive to the right in every task, and a release check verifies the stated convention against the camera's right vector. A question with more than one correct answer, such as a name that fits 2 instances, is removed at generation, and anchor views that face a wall or a large surface at close range are rejected. A manual review of the questions and their targets removed 20 static questions (14 Loc2Cat, 3 Ego-Loc, and 1 each of Dist-Ego, Dist-Obj and Abs-Loc) whose ground truth was wrong. Of these, 13 were Loc2Cat questions whose answer "door" marked a doorless passage, and 7 referred to a target on the storey above the agent. The uniqueness check at generation compares displayed names, so a stricter audit also counts objects of the same MPcat40 class under another name as rivals. It flags 318 of the remaining static questions as possibly ambiguous, and removing them changes no ranking (Appendix C.3).

A.3 Map Rendering

Maps are photometric top-down renderings of the building mesh made with Open3D (Zhou et al., 2018). Every map carries the room labels on an 80% opaque white background, a ring at the agent's position and a scale bar in a footer below the map, and north-up maps also carry a compass rose in the top-right corner. The scale bar represents 1 or 2 m, whichever is closer to a fifth of the map's width, and the prompt states its length. It is 2 m on 59 of the 60 route maps. The ring carries no heading cue, and the prompt states the heading as a map angle. Self-Pose replaces the ring with 6 lettered candidate positions without a heading cue, and Pose2View marks the queried position and offers 6 views, the correct one, views facing the queried heading from nearby positions, and views of similar objects elsewhere in the building. The V-nu and V-hu configurations of Section 4.5 crop a square around the agent, and V-hu rotates it so that the agent's heading points up and draws no compass. On 101 of the 119 crops in each configuration the scale bar is 2 m.

A.4 Prompts and Answer Formats

Every prompt describes each image by its role. The egocentric view comes with its camera model (pinhole, 640×480 pixels, 90° horizontal field of view, focal length 320 pixels, stated height 1.3 m, level optical axis), and the map with its conventions: map angles are 0° to the right, 90° up and increase counter-clockwise, and the agent's current heading is given in text. Grounding definitions follow (relative bearings in $[-180^\circ, 180^\circ)$, positive to the right; horizontal distance on the ground plane to the object's centre or a doorway's middle), then the goal and the output format, a single answer line (Table 6). Transition additionally states the 60-action limit and, for the start pose, how many forward steps are guaranteed free after a given number of turns; Nav-Text allows up to 16

actions per turn within an 80-action budget, gives in every turn the rooms visited so far, the actions already executed and the remaining budget, and ends a turn early when a forward step is blocked; a turn that returns stop ends the episode. Both action tasks also state the room the navigator is in.

Table 6: Answer formats. Actions are `move_forward`, `turn_left`, `turn_right` and `stop`.

Task	Answer line
Dist-Ego	`Answer: <landmark_distance_m>`
Dist-Obj	`Answer: <object_distance_m>`
Self-Pose	`Answer: <A|B|C|D|E|F>, <absolute_heading_deg>`
Pose2View	`Answer: <A|B|C|D|E|F>`
Ego-Loc	`Answer: <relative_bearing_deg>, <landmark_distance_m>`
Abs-Loc	`Answer: <absolute_bearing_deg>, <landmark_distance_m>`
Loc2Cat	`Answer: <category>`
Transition, Nav-Text	`Answer: <action_1>, <action_2>, ..., <action_n>`

B Evaluation Protocol Details

B.1 Metrics

Numeric answers are scored by

$$\mathrm{MRA}(e) = \frac{1}{10}\sum_{k=1}^{10} \mathbf{1}\left[e < 0.05k\right],$$

the mean relative accuracy of Yang et al. (2025) with thresholds $\theta = 1 - 0.05k$, so an error of 0.5 or more scores 0. The relative error is $e = |\hat{d} - d|/d$ for Dist-Ego and Dist-Obj, where $\hat{d}$ is the answered and d the true distance, and $e = \|\hat{\mathbf{p}} - \mathbf{p}\|/d$ for Ego-Loc and Abs-Loc, where $\hat{\mathbf{p}}$ is the position given by the answered bearing and distance and $\mathbf{p}$ that of the target, both in the frame of the question. Transition scores the goal progress $\max(0, 1 - E/L)$, where E is the geodesic distance from the stopping point to the goal and L that from the start. Nav-Text scores sDTW, the product of success within 3 m and nDTW (Ilharco et al., 2019). Self-Pose is correct when the chosen candidate is the true one and the answered heading is within $30°$. Loc2Cat averages the accuracies of 6 groups, namely door, chair, table, sofa, cabinet and all remaining object categories.

B.2 Chance Floors

Static floors are the expected score of a guess drawn uniformly without reading the input, over distances in $[0.5, 8]$ m, bearings in $[0°, 360°)$, the 40 object categories of Loc2Cat or the 6 choices, with 200 draws per question. Transition and Nav-Text floors replay a prompt-blind random walk from the true start through the same executor and scorer. A category's floor c is the mean of its task floors, and with the category score s and c in percent the normalised score is $\mathrm{clip}((s - c)/(100 - c),\, 0,\, 1) \times 100$. Per-task floors are the Chance row of Table 9.

B.3 Parsing and Truncated Answers

All answers pass through one frozen parser. Markdown emphasis around the answer line is removed before parsing, which changes 2 parsed answers. A reply cut off by the length limit counts only if, after its reasoning, it had written complete answer lines that all parse and agree; any other unparsable answer scores 0. For RoboBrain2.0-32B the answer is taken from the text after its reasoning, then from the last answer line, then from a trailing answer tag.

B.4 Model Settings and Protocol Differences

Cosmos-Reason2-32B is post-trained from Qwen3-VL-32B-Instruct (Bai et al., 2025) on physical-reasoning video, and RoboBrain2.0-32B targets general embodied reasoning, including simulated object search with high-level navigation actions.

Table 7: Inference settings. Open models run on vLLM with each model's chat template, output capped at 32,768 tokens or the context left by the input. For every model, a Nav-Text turn that moves and then stops ends the episode. The prompts state a camera height of 1.3 m, while the rendered floor lies about 1.40 m below the camera (median over start poses); ground truth is computed from the rendered geometry.

Model	Reasoning	Decoding	Notes
Closed-source general MLLMs			
Gemini-3.8-Flash	thinking level medium	vendor default	prompt wording v2; text before images; 2 static questions and 2 Transition segments lost to API errors and left out of its means
GPT-5.6 Sol	reasoning effort medium	vendor default	via the Codex CLI; images before text; Pose2View file names reveal the correct view (see text)
Claude-Opus-5	thinking, effort medium	vendor default	prompt wording v2; images before text
Open-source general MLLMs			
Qwen3.8-27B	thinking, effort xhigh	T=1.0, p=0.95, k=20	official recommended setting
Gemma-4-31B	thinking	T=1.0, p=0.95, k=64	1,120 visual tokens
InternVL3.5-38B	R1 thinking prompt	T=0.6, p=0.95, k=50	up to 12 image tiles
Embodied foundation models			
Hy-Embodied-A3B	thinking	T=0.7	mixture of experts, about 30B parameters with 3B active; official vLLM plugin
Cosmos-Reason2-32B	official answer format	T=0.7, p=0.8, k=20	
RoboBrain2.0-32B	thinking (general mode)	T=0.7, p=1.0, k=50	repetition penalty 1.05; answer extraction as above

Prompt wording v2 changes only phrasing and leaves every image and answer unchanged. In all 1,006 static prompts of Dist-Ego, Dist-Obj, Ego-Loc, Abs-Loc and Loc2Cat it rewords the distance definition, the reference point of doorways and the closing format line, and in Transition it removes a sentence on sliding along obstacles and merges repeated sentences. The Codex CLI passes each image together with its file path, and in Pose2View the file name of the correct view differs from those of the distractors. The reasoning summaries of GPT-5.6 Sol never mention file names, and its Pose2View accuracy (68.9) lies between those of Claude-Opus-5 (63.1) and Gemini-3.8-Flash (73.7), which did not see them.

B.5 Human Reference

The human reference comes from 5 annotators who had not seen the benchmark before. We select 30 of the 60 Nav-Text episodes and 50 items from each of the other 8 tasks, that is, 50 of the 119 Transition segments and 50 questions of each static task, 430 items in total. Each item is answered by at least 3 annotators, and its human score is the mean of their scores. The annotators receive the same images, prompts and answer formats as the models, and their answers are scored with the same pipeline. Table 8 scores every model on the same 430 items. Overall scores differ from the full-set scores by at most 5.6 points (mean 2.8) and single categories by at most 12.3 points (mean 3.7). The models score 1.9 points lower Overall on average, and the Overall ranking changes only in 2 neighbouring pairs, so the gap to the human in Table 2 does not come from the subset.

Table 8: Model scores ($\times 100$) on the 430-item subset answered by the annotators, with full-set scores in parentheses. Overall uses the full-set chance floors, as the Human row does. Gemini-3.8-Flash lacks 1 Abs-Loc and 1 Pose2View question of the subset.

Model	Scale	Pose	Identity	Transition	Nav	Overall
Gemini-3.8-Flash	61.2 (64.8)	71.7 (72.8)	52.6 (56.0)	57.6 (64.5)	71.9 (68.7)	58.0 (60.9)
GPT-5.6 Sol	55.2 (63.2)	47.5 (52.2)	60.5 (55.3)	59.7 (59.5)	36.4 (45.8)	45.9 (49.9)
Claude-Opus-5	50.5 (52.1)	62.5 (61.7)	47.2 (44.9)	52.8 (49.7)	53.9 (58.1)	47.2 (47.1)
Qwen3.8-27B	51.2 (59.5)	40.0 (47.1)	43.4 (43.6)	45.9 (48.3)	37.3 (43.3)	36.6 (42.2)
Gemma-4-31B	46.2 (47.4)	32.5 (32.3)	29.5 (28.1)	41.1 (41.0)	23.4 (33.3)	26.8 (28.8)
InternVL3.5-38B	49.3 (41.2)	15.0 (14.8)	22.0 (14.0)	17.3 (15.7)	0.0 (1.3)	12.3 (8.2)
Hy-Embodied-A3B	42.7 (46.7)	20.0 (18.9)	16.8 (18.2)	14.1 (20.6)	11.9 (11.6)	12.2 (14.4)
Cosmos-Reason2-32B	32.5 (40.0)	15.0 (13.8)	9.2 (11.6)	13.0 (13.9)	0.0 (2.7)	4.9 (7.2)
RoboBrain2.0-32B	20.2 (32.6)	10.0 (12.4)	17.2 (11.4)	18.3 (19.8)	0.0 (4.1)	3.8 (6.0)
Max $\lvert\Delta\rvert$	12.3	7.1	8.0	6.9	9.9	5.6
Human	76.8	97.5	74.2	71.0	84.1	77.9

B.6 Navigation Protocol Checks

Harness. To test whether our closed-loop harness depresses navigation scores, we ran 2 VLN-CE specialists, NaVILA-8B and Uni-NaVid-7B, through it on 100 R2R-CE val-unseen episodes with their official inference code under VLN-CE rules, with sliding along walls, 500 actions and success within 3 m. The harness reproduces their published success rates, 54.0 [44, 64] against 54.0 for NaVILA-8B and 60.0 [50, 69] against 51.8 for the released Uni-NaVid-7B checkpoint, with 95% CIs from a bootstrap over episodes.

Hy-Embodied-A3B. Hy-Embodied-A3B does not reproduce its reported 57.9 under the same rules (1.0 [0, 3]), because it stops in only 3 of the 100 episodes, although it comes within 3 m of the goal in 35. Its navigation prompt and termination procedure are not public. Replaying its official demonstration in our simulator reproduces the demonstration's distances to the goal within 5 mm. At the 21 decision states of the demonstration, our prompt with the released weights reproduces 16 of its first actions, but none of the 11 action prompts we tried outputs stop, and our stop question already answers yes 10 decisions before the demonstration stops. With the demonstration's loop, which executes 2 of every 4 predicted actions, Hy-Embodied-A3B still stops in only 1 of the 100 episodes.

Specialists on BEARING. The 2 specialists do not take a map. Given only the route instruction and their own frame history, they reach sDTW 19.0 and 36.4 on the 60 BEARING routes under our rules and 24.0 and 34.2 under VLN-CE rules, whereas with our map-and-view prompt they score 1.3 and 2.6. Under our rules Uni-NaVid-7B ends within 3 m of the goal on 27 of the 60 routes, so the route instructions can be followed without the map.

Protocol sensitivity of the embodied models. We reconstructed the inference setting of each embodied model from its release and ran it closed loop under our action rules. In these settings the Nav-Text sDTW of Hy-Embodied-A3B, Cosmos-Reason2-32B and RoboBrain2.0-32B is 11.1, 5.2 and 6.3, against 11.6, 2.7 and 4.1 under our protocol, although we chose each setting on 12 of the 60 routes. Our action rules are stricter than VLN-CE, with 80 primitive actions, no sliding along walls, and an episode that ends after 3 consecutive replies without movement. Relaxing them to VLN-CE rules raises the number of routes on which Hy-Embodied-A3B ends within 3 m of the goal from 11 to 26 of 60, a paired difference of 25.0 [8.3, 40.0] points. Its sDTW rises less, from 11.6 to 19.5 with a paired difference of 7.9 [-2.0, 17.7], and it uses a median of 158 actions. Ignoring the stop of Cosmos-Reason2-32B and InternVL3.5-38B changes little. They end within 3 m of the goal on 4 and 3 of the 60 routes, against 2 and 1 under our protocol, and Cosmos-Reason2-32B ends 36 episodes after 3 consecutive replies whose first move is blocked. Closed-loop execution matters more for Transition, which we score open loop. When Transition runs closed loop with our prompt, a new view and a redrawn map after each reply, the progress of Hy-Embodied-A3B rises from 20.6 to 35.9 and that of RoboBrain2.0-32B from 19.8 to 24.1, while that of Gemma-4-31B falls from

41.0 to 37.7. In their own settings, also run closed loop, Hy-Embodied-A3B, Cosmos-Reason2-32B and RoboBrain2.0-32B reach 25.8, 29.1 and 20.0.

C Full Results

C.1 Per-Task Scores

Table 9: Per-task scores ($\times 100$) under the full input; each category score in Table 2 is the mean of its tasks. Chance: expected score of a uniform guess over the answer space, fixed without reading the input. Human: mean of at least 3 of 5 annotators on 50 questions per task (Appendix B.5).

	Scale		Pose		Identity		
Model	Dist-Ego	Dist-Obj	Self-Pose	Pose2View	Ego-Loc	Abs-Loc	Loc2Cat
Closed-source general MLLMs							
Gemini-3.8-Flash	66.9	62.6	71.8	73.7	44.7	52.6	70.6
GPT-5.6 Sol	68.5	57.9	35.4	68.9	45.2	54.3	66.3
Claude-Opus-5	52.2	52.0	60.2	63.1	36.1	38.0	60.6
Open-source general MLLMs							
Qwen3.8-27B	65.1	53.8	44.2	50.0	33.5	41.1	56.0
Gemma-4-31B	51.6	43.1	20.4	44.2	15.6	27.9	40.8
InternVL3.5-38B	43.3	39.2	7.3	22.3	8.8	9.0	24.3
Embodied foundation models							
Hy-Embodied-A3B	46.3	47.1	10.2	27.7	13.8	15.4	25.5
Cosmos-Reason2-32B	35.3	44.7	5.3	22.3	7.6	9.8	17.5
RoboBrain2.0-32B	32.9	32.2	4.9	19.9	7.7	6.3	20.0
Human	86.0	67.5	100.0	95.0	65.5	65.5	91.7
Chance	22.8	25.1	2.8	16.2	1.8	2.0	2.7

C.2 Navigation Details

Table 10: Nav-Text behaviour over the 60 episodes. Success: stopped within 3 m of the goal; within 3 m: came within 3 m at any point. An episode ends when the model outputs stop, when repeated moves are blocked by obstacles, or when the reply reaches the length limit.

			Median per episode		Episode ended by		
Model	Success	Within 3 m	Turns	Primitives	stop	blocked	length
Closed-source general MLLMs							
Gemini-3.8-Flash	51	54	6	48	60	0	0
GPT-5.6 Sol	38	42	8	70	58	2	0
Claude-Opus-5	45	52	6	53	60	0	0
Open-source general MLLMs							
Qwen3.8-27B	33	40	8	53	57	3	0
Gemma-4-31B	28	34	8	75	51	9	0
InternVL3.5-38B	1	5	1	12	49	0	11
Embodied foundation models							
Hy-Embodied-A3B	11	17	17	80	48	11	1
Cosmos-Reason2-32B	2	5	1	5	58	0	2
RoboBrain2.0-32B	4	10	14	36	32	28	0

Gemini-3.8-Flash and Claude-Opus-5 end every episode by stopping. RoboBrain2.0-32B ends 28 episodes blocked against obstacles, InternVL3.5-38B ends 11 with reasoning that runs to the length limit, and Cosmos-Reason2-32B stops after a median of 1 turn and 5 primitives.

C.3 Robustness

Table 11: Overall on 2 subsets of the static questions. Without ambiguous targets: the 318 static questions flagged in the target-ambiguity review are removed. View-required subset: static questions whose answer needs the egocentric view (Dist-Ego 66/249, Dist-Obj 105/239, Loc2Cat 20/100, Abs-Loc 42/184, Ego-Loc 71/234; Self-Pose and Pose2View all).

	All questions	Without ambiguous targets		View-required subset	
Model	Overall	Overall	rank	Overall	rank
Gemini-3.8-Flash	60.87	61.68	1	58.39	1
GPT-5.6 Sol	49.87	50.72	2	47.92	3
Claude-Opus-5	47.12	46.13	3	49.15	2
Qwen3.8-27B	42.20	41.66	4	41.62	4
Gemma-4-31B	28.80	29.49	5	27.15	5
InternVL3.5-38B	8.15	7.42	7	6.93	7
Hy-Embodied-A3B	14.37	14.83	6	13.10	6
Cosmos-Reason2-32B	7.16	7.00	8	6.68	8
RoboBrain2.0-32B	6.05	6.01	9	4.72	9

Every static question needs the map. The Pose questions also need the view, and the Scale and Identity questions need it whenever the named object must be recognised in the scene. Removing the ambiguous targets moves Overall by at most 1.0 point and leaves every rank unchanged. On the view-required subset GPT-5.6 Sol and Claude-Opus-5 exchange second and third place; the other ranks are unchanged.

D Error Analysis Details

D.1 Per-Model Profiles

Bearing errors are the medians of Table 3.

Gemini-3.8-Flash has the smallest Abs-Loc error ($6°$), gives 42% of its Self-Pose headings on the 4 map axes and succeeds in 51 of 60 Nav-Text episodes. Its heading turns with a rotated compass rose in 72% of questions and with a rotated floor in 4%, and its distances scale by $\times 0.61$ and $\times 1.75$ with the relabelled scale bar.

GPT-5.6 Sol matches the relabelled scale bar ($\times 0.50$, $\times 2.00$) and shifts a located object by 1.8 to 7.5 m for marker moves of 2 to 8 m. It scores 68.9 on Pose2View, which states the heading, but 35.4 on Self-Pose, which asks for it, gives 79% of its headings on the axes and succeeds in 38 Nav-Text episodes.

Claude-Opus-5 has an Ego-Loc error of $6°$ for view targets and gives the fewest headings on the axes (31%). Its Abs-Loc error is $28°$, its Dist-Ego score falls from 56.1 for view to 39.4 for map targets, and it succeeds in 45 Nav-Text episodes.

Qwen3.8-27B has Ego-Loc errors of $8°$ for view and $41°$ for map targets, against $19°$ on Abs-Loc, and gives 89% of its headings on the axes. Its heading turns with a rotated rose in 51% and a located object shifts by 1.2 to 5.2 m for marker moves of 2 to 8 m, whereas its distances scale by $\times 0.82$ and $\times 1.10$.

Gemma-4-31B has Ego-Loc errors of $18°$ for view and $78°$ for map targets, or $27°$ with the sign flipped, and gives 92% of its headings on the axes. Its heading turns with a rotated rose in 28%, compared with 17 and 30% on a repeat, and a located object shifts by 1.0 to 1.1 m whatever the marker's move.

InternVL3.5-38B has an Ego-Loc error of $50°$ for view targets, gives every heading on the axes, most often $0°$, and favours option B in Pose2View. The map image lowers its Transition progress by 8.6 points, whereas turning the map heading-up raises it by 11.3 (V-hu$-$V-nu). It succeeds in 1 Nav-Text episode and ends 11 at the length limit.

Hy-Embodied-A3B has Ego-Loc errors of $17°$ for view and $92°$ for map targets, or $45°$ with the sign flipped, and gives 84% of its headings on the axes, most often $90°$. Its heading turns with a rotated rose in 48% and with a rotated floor in 38%, compared with 16 and 8% on a repeat. It succeeds in 11 Nav-Text episodes and ends 11 blocked by obstacles.

Cosmos-Reason2-32B gives 99% of its headings on the axes and chooses Pose2View options more often the earlier they are listed, although the correct option is balanced across positions. The map image lowers its Transition progress by 5.6 points and the map as text raises it by 15.4. It succeeds in 2 Nav-Text episodes, stopping after a median of 1 turn.

RoboBrain2.0-32B has Ego-Loc errors of $37°$ for view and $69°$ for map targets against $66°$ on Abs-Loc. Its Transition progress changes by -2.5 and -1.3 points with the map image and its text form, with 95% CIs that include 0, but falls by 6.7 with a blank image. It succeeds in 4 Nav-Text episodes and ends 28 blocked by obstacles.

D.2 Mirrored Bearings in the Reasoning

For targets known only from the map, Gemma-4-31B and Hy-Embodied-A3B subtract their heading from the target's map angle. Map angles grow counter-clockwise, so the result is positive to the left, while the task defines right as positive. Below are 2 excerpts for a question whose true bearing is $+129°$ (behind, to the right):

> Gemma-4-31B: "$\theta_{rel} = 285 - 49.4 = 235.6 \rightarrow -124.4°$ … To my right is $49.4° + 90° = 139.4°$"
> Hy-Embodied-A3B: "angle = atan2(−300, 240) = -51.34 deg … heading = 49.4 deg … relative = $-51.34 - 49.4 = -100.74$ deg … Answer: -101"

E Correlation Analysis Details

E.1 Method

For each model and each pair of categories, the Spearman correlation (average ranks) is computed over the routes of each scene and averaged over scenes weighted by the number of routes; Figure 2a averages the 9 models. Confidence intervals come from 1,000 bootstrap draws that resample routes within scenes, with the same draw applied to every model.

E.2 Metric Sensitivity

Table 12: Route-level correlation between navigation and each category (as Figure 2a) with navigation scored by sDTW, nDTW or goal progress; mean over the 9 models, 95% CI from the joint route bootstrap; bold: CI excludes 0. sDTW is zero for every failed episode, so for the weaker models it varies on only a few routes.

Pair	sDTW	nDTW	progress
Nav–Scale	-0.13 [-0.25, $+0.07$]	$+0.06$ [-0.07, $+0.17$]	$+0.07$ [-0.05, $+0.18$]
Nav–Pose	$+0.01$ [-0.15, $+0.26$]	$+0.03$ [-0.07, $+0.13$]	$+0.03$ [-0.07, $+0.11$]
Nav–Identity	$+0.18$ [-0.06, $+0.33$]	$+0.06$ [-0.07, $+0.16$]	$+0.07$ [-0.03, $+0.17$]
Nav–Transition	$+0.13$ [-0.02, $+0.28$]	$+0.08$ [-0.02, $+0.15$]	**$+0.16$ [$+0.04$, $+0.26$]**

Scoring Pose continuously, by the chosen candidate's distance from the true position and the heading error, leaves its correlation with navigation between -0.02 and 0.03, and no Pose correlation with Nav-Text or Transition excludes zero under any of these scorings.

E.3 Direction and Range of Identity

The Ego-Loc and Abs-Loc score is the MRA of the located point's error, $\sqrt{r^2 + 1 - 2r\cos b}$ for range ratio r and bearing error b. Of the 3,761 Ego/Abs-Loc answers of the 9 models, 1,552 have a bearing error of $30°$ or more and score 0 whatever the range, because the error is then at least

$\sin 30° = 0.5$. A range error lowers the located point's error in 766 answers, but raises the score in only 16. Figure 2b scores both parts on this common scale. Under the other scorings we tried (bearing error linear and range by MRA, or both on the common scale scored linearly, each with and without Loc2Cat), direction correlates with same-start Transition progress more strongly than range by 0.08 to 0.12.

E.4 Model-Level Correlation

Across the 9 models, the category scores of Table 2 are rank-correlated at 0.88 to 0.98 for every pair of categories.

F Map Utility Details

Table 13: Map utility as Transition progress ($\times 100$) of every model under each map configuration of Figure 3a, on the same 119 segments. V is the Transition column of Table 2. Claude-Opus-5 ran V only, and a dash marks a configuration that was not run.

Config	Gemini	GPT	Claude	Qwen	Gemma	InternVL	Hy	Cosmos	RoboBrain
N	45.9	40.2	–	36.2	34.0	24.3	22.4	19.4	22.3
P	41.6	41.4	–	38.3	31.5	26.8	22.7	18.5	15.6
V	64.5	59.5	49.7	48.3	41.0	15.7	20.6	13.9	19.8
V-only	59.5	50.3	–	44.6	37.6	18.4	18.1	16.2	18.0
T	58.7	55.6	–	45.6	51.1	41.0	31.8	34.8	21.0
VT	65.4	60.9	–	51.5	50.5	35.4	33.6	29.2	21.4
V-nu	64.5	57.6	–	45.6	38.7	21.8	19.3	16.9	19.8
V-hu	65.6	57.2	–	48.8	44.4	33.1	22.9	15.4	20.9

Table 14: Map utility as the mean paired difference in Transition progress (points), with its 95% CI from a bootstrap over routes (20,000 draws), for all models that ran the configurations. The first 6 rows are the panels of Figure 3b, and V−T, VT−T and V-nu−V, the square crop alone, complete the set.

Contrast	Gemini	GPT	Qwen	Gemma	InternVL	Hy	Cosmos	RoboBrain
P−N	−4.2 [−8.3, −0.3]	+1.2 [−3.4, 5.7]	+2.1 [−2.1, 6.4]	−2.5 [−6.2, 1.1]	+2.5 [−2.8, 7.7]	+0.3 [−5.4, 5.5]	−0.9 [−4.7, 2.9]	−6.7 [−10.1, −3.3]
V−N	+18.8 [11.8, 25.9]	+19.4 [13.0, 26.0]	+12.1 [5.5, 18.8]	+7.0 [0.4, 13.6]	−8.6 [−13.4, −3.5]	−1.7 [−6.1, 2.5]	−5.6 [−9.6, −1.8]	−2.5 [−6.2, 1.2]
T−N	+12.9 [7.0, 18.7]	+15.4 [8.7, 22.0]	+9.4 [3.7, 15.2]	+17.1 [10.7, 23.6]	+16.7 [9.5, 23.8]	+9.4 [3.7, 15.0]	+15.4 [8.3, 22.4]	−1.3 [−5.3, 2.8]
VT−V	+1.2 [−2.5, 4.8]	+1.4 [−2.9, 5.8]	+3.3 [−2.2, 8.7]	+9.5 [2.8, 16.1]	+19.7 [11.9, 27.5]	+13.0 [7.9, 18.3]	+15.3 [10.2, 20.5]	+1.6 [−2.6, 5.8]
V-only−V	−5.0 [−9.1, −1.0]	−9.3 [−15.3, −3.0]	−3.6 [−10.0, 2.6]	−3.4 [−10.2, 3.5]	+2.8 [−3.3, 9.1]	−2.6 [−6.9, 1.9]	+2.3 [−0.9, 5.6]	−1.8 [−6.9, 3.4]
V-hu−V-nu	+1.1 [−2.7, 4.9]	−0.3 [−5.8, 5.2]	+3.2 [−3.1, 9.5]	+5.7 [0.5, 11.1]	+11.3 [5.7, 16.9]	+3.6 [−0.7, 8.0]	−1.5 [−4.5, 1.4]	+1.1 [−3.1, 5.3]
V−T	+5.1 [−0.4, 11.1]	+4.0 [−1.5, 9.8]	+2.7 [−2.5, 8.0]	−10.1 [−16.6, −3.6]	−25.3 [−32.4, −18.2]	−11.2 [−16.1, −6.3]	−21.0 [−26.6, −15.3]	−1.2 [−5.7, 3.2]
VT−T	+6.6 [2.2, 11.3]	+5.4 [0.8, 10.1]	+5.9 [1.4, 10.4]	−0.6 [−4.2, 3.0]	−5.6 [−10.9, −0.4]	+1.8 [−3.6, 7.4]	−5.6 [−10.3, −0.9]	+0.4 [−3.1, 3.8]
V-nu−V	+0.3 [−3.8, 4.3]	−2.0 [−7.4, 3.4]	−2.6 [−8.3, 3.2]	−2.3 [−7.7, 3.1]	+6.1 [1.1, 11.2]	−1.3 [−5.0, 2.4]	+3.1 [−0.1, 6.2]	+0.1 [−4.1, 4.1]

GPT-5.6 Sol and Qwen3.8-27B follow the pattern of Gemini-3.8-Flash, except that a blank image (P−N) costs Gemini-3.8-Flash a few points but not them, and removing the view (V-only−V) costs Qwen3.8-27B 3.6 points with a CI that includes 0.

G Map Element Sensitivity Details

Table 15: Map elements with uncertainty (Table 4). Scale bar and agent marker: median over questions with its 95% bootstrap CI. Compass: share of questions whose answered heading turns with the rotated rose (rose rotated alone) or with the rotated floor (floor and rose rotated together), within 15°, and below it the share that turns by the same angle on a repeat of the unchanged question. Repeat noise: the median ratio between 2 answers to the same unchanged Dist-Ego question.

	Scale bar		Rose rotated		Floor + rose rotated		Agent marker (m)			Repeat
Model	$\times 0.5$	$\times 2$	$+90^\circ$	-90°	$+90^\circ$	-90°	2	4	8	noise
Gemini	$\times 0.61$ [$\times 0.55$, $\times 0.68$]	$\times 1.75$ [$\times 1.64$, $\times 1.89$]	68% rep. 3%	75% rep. 3%	3% rep. 3%	4% rep. 3%	1.6 [1.3, 1.8]	3.3 [2.8, 3.6]	6.3 [1.4, 7.1]	$\times 1.10$
GPT	$\times 0.50$ [$\times 0.50$, $\times 0.51$]	$\times 2.00$ [$\times 1.97$, $\times 2.04$]	40% rep. 8%	55% rep. 10%	14% rep. 8%	19% rep. 10%	1.8 [1.6, 1.9]	3.8 [3.2, 3.9]	7.5 [6.5, 7.9]	$\times 1.06$
Qwen	$\times 0.82$ [$\times 0.75$, $\times 0.92$]	$\times 1.10$ [$\times 1.00$, $\times 1.20$]	56% rep. 12%	47% rep. 7%	23% rep. 12%	21% rep. 7%	1.2 [0.6, 1.6]	2.7 [2.2, 3.1]	5.2 [2.5, 6.6]	$\times 1.17$
Gemma	$\times 0.87$ [$\times 0.77$, $\times 0.92$]	$\times 1.47$ [$\times 1.23$, $\times 1.65$]	25% rep. 17%	32% rep. 30%	32% rep. 17%	26% rep. 30%	1.1 [0.5, 1.7]	1.1 [0.2, 2.2]	1.0 [-0.4, 2.7]	$\times 1.23$
InternVL	$\times 0.95$ [$\times 0.83$, $\times 1.00$]	$\times 1.00$ [$\times 1.00$, $\times 1.00$]	11% rep. 26%	20% rep. 17%	27% rep. 26%	24% rep. 16%	0.0 [-0.1, 0.5]	0.4 [0.1, 0.9]	0.2 [-0.5, 0.5]	$\times 1.25$
Hy	$\times 0.88$ [$\times 0.81$, $\times 0.95$]	$\times 1.08$ [$\times 1.00$, $\times 1.19$]	45% rep. 16%	50% rep. 8%	40% rep. 16%	37% rep. 8%	0.5 [0.0, 0.8]	0.7 [0.2, 1.1]	0.2 [-0.3, 1.8]	$\times 1.24$
Cosmos	$\times 1.00$ [$\times 1.00$, $\times 1.00$]	$\times 1.00$ [$\times 1.00$, $\times 1.25$]	32% rep. 20%	21% rep. 25%	26% rep. 20%	27% rep. 25%	0.0 [-1.1, 0.0]	0.0 [-0.6, 0.0]	0.1 [-0.9, 1.9]	$\times 1.29$
RoboBrain	$\times 1.00$ [$\times 0.90$, $\times 1.10$]	$\times 1.02$ [$\times 1.00$, $\times 1.17$]	17% rep. 18%	18% rep. 15%	18% rep. 18%	23% rep. 15%	0.1 [-0.3, 0.7]	-0.1 [-0.5, 0.2]	0.5 [-1.7, 0.9]	$\times 1.65$

The scale-bar arms ask Dist-Ego with the camera height removed from the prompt (120 questions per arm), so the bar is the only metric cue. The compass arms ask Self-Pose, and the marker arms ask Ego-Loc; the marker baseline comes from a second sample (seed 10), so that comparison crosses samples. Repeat noise is measured on second samples of the unchanged questions. Claude-Opus-5 was run on the full benchmark but not on these edits, for reasons of API cost and access. InternVL3.5-38B and RoboBrain2.0-32B were each served from 2 machines with the same weights and settings, and their invalid rates and repeat differences do not differ between the machines.

When the rose alone is rotated, Gemini turns 72% of its answered headings with it and GPT-5.6 Sol, Qwen3.8-27B and Hy-Embodied-A3B about half, against at most 12% on a repeat of the unchanged question for these 4 models. When the floor rotates with the rose, Gemini keeps its heading, whereas GPT-5.6 Sol and Qwen3.8-27B turn with the floor in about a fifth of questions and Hy-Embodied-A3B in 38%, so they read part of the heading from the image axes. When the marker moves, GPT-5.6 Sol and Gemini move the located object by most of the offset and Qwen3.8-27B by about two-thirds, whereas the other models move it by about 1 m or less, whatever the offset.