EGOPATHBENCH: EVALUATING ZERO-SHOT EGOCENTRIC WAYPOINT DECISION-MAKING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot waypoint navigation requires vision-language models to select, from the current first-person observation, a sequence of spatial actions that is feasible for the agent and reaches the goal, placing joint demands on the integrated spatial intelligence of today's foundation VLMs. Existing spatial-intelligence benchmarks primarily evaluate isolated judgments of relations, directions, or targets and therefore do not directly measure the integrated navigation ability required to combine target recognition, action-consequence assessment, distance estimation, and path planning. To fill this evaluation gap, we introduce EgoPathBench, a dataset and five-task benchmark for first-person waypoint decision-making. Each question presents an egocentric RGB image, a natural-language goal, and numbered visible waypoints; a model returns traversable candidates or an ordered route. Predictions are evaluated for candidate feasibility, adjacent-edge legality, and goal arrival under the task-specific Point or Embodied feasibility rules. EgoPathBench contains 31,852 training, 1,345 validation, and 1,111 benchmark questions and retains at least one geometrically verified reference route for every route question. Across nine VLMs, the highest EgoPath Score is only 28.3. The top-ranked model reaches 35.9% success on Point Path, but only 2.9% and 4.0% on Embodied Path and Intent Path, respectively, showing that current models remain limited in forming complete, goal-consistent routes under embodiment constraints. Beyond the evaluation data, we release the corresponding training resource. Fine-tuning Qwen 3.5 4B on the released training split raises its EgoPath Score from 3.9 to 38.9 and improves all four reported evaluations across three external spatial benchmarks, with gains of 1.4–9.6 points.

1 INTRODUCTION

Foundation vision-language models are increasingly used in embodied navigation to translate first-person observations and task goals into spatial actions. Recent work has explored VLMs or LLMs for explicit navigation reasoning, generalist navigation, visual next-step planning, and zero-shot ObjectNav Zhou et al. (2024); Zheng et al. (2024); Zhang et al. (2024); Cai et al. (2025). Navigation decisions require more than recognizing objects or answering an individual spatial-relation question: a model must interpret the current environment, the intended target, and the available actions to make a goal-directed spatial choice. The integrated spatial intelligence of foundation VLMs is therefore an important basis for zero-shot navigation to generalize across new scenes and goals.

Existing evaluations characterize complementary parts of spatial and embodied reasoning. SpatialVLM and SpatialEval evaluate metric and relational spatial judgments; VSI-Bench studies spatial understanding and memory from video observations; and 3DSRBench evaluates reasoning about 3D structure Chen et al. (2024); Wang et al. (2024); Yang et al. (2025); Ma et al. (2025). EmbSpatial-Bench, EgoThink, and OpenEQA extend evaluation to first-person or embodied observations Du et al. (2024); Cheng et al. (2024); Majumdar et al. (2024). These benchmarks provide useful measurements of component abilities. Navigation evaluations provide a complementary view: NavBench studies navigation comprehension and sequential execution Qiao et al. (2025a), while outcomes in full navigation systems may additionally depend on mapping, localization, memory, control, replanning, and recovery. NaviTrace evaluates two-dimensional navigation traces from a single real-

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.

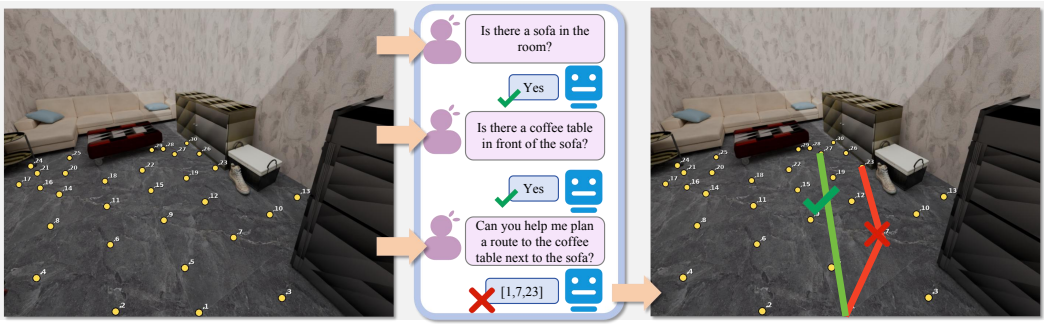

Figure 1: Conceptual motivation for EgoPathBench. On the same first-person scene, familiar recognition and local spatial questions can be answered correctly. The right panel contrasts a complete, goal-reaching route with an incorrect route that violates the route and goal criteria. This contrast motivates evaluating integrated, embodiment-aware route decisions rather than component visual questions alone.

world image against expert demonstrations using a semantic-aware score Windecker et al. (2025). EgoPathBench complements these directions by isolating waypoint selection from the surrounding navigation system and directly evaluating whether a foundation VLM can choose a geometrically feasible, goal-consistent route over currently visible actions.

To fill this evaluation gap, we introduce EgoPathBench, a dataset and five-task benchmark for first-person waypoint decision-making. EgoPathBench uses waypoint selection in the current view as a controlled measurement substrate for integrated spatial intelligence. Each question presents an egocentric RGB image, a natural-language prompt, and numbered visible waypoints; the model returns a JSON list containing traversable candidates or an ordered route. Across five tasks, the benchmark organizes target recognition and grounding, waypoint-to-scene correspondence, geometric action consequences, distance and route efficiency, embodied feasibility, and multi-step route planning within a unified action-selection interface. The resulting construct tests whether these abilities jointly support a first-person decision with explicit action consequences. Figure 1 illustrates this motivation conceptually: the same scene can support familiar recognition and local spatial judgments, while forming a complete, legal, and goal-consistent waypoint route requires these abilities to operate jointly over an action sequence.

Point Traversability asks which visible waypoints are walkable, while Embodied Traversability asks which are traversable by a robot with a diameter of 0.6 m. Explicit Point-Goal Path and Explicit Embodied-Goal Path share the scene, target, and candidate space while asking for a route to an explicit target; the Embodied task additionally specifies the robot dimensions. Intent-Grounded Embodied Path instead describes the target through an intent and a visual cue, requiring target grounding together with embodied route selection. Point and Embodied tasks use the same observation and waypoint vocabulary, while the corresponding geometric feasibility graph defines the consequences of each selected action.

The ordered waypoint route makes these abilities jointly testable in a structured output. A model must select an appropriate immediate action, maintain agent-specific feasibility across every adjacent edge, and terminate at an acceptable goal. EgoPathBench therefore checks candidates, the required start, adjacent edges, the endpoint, and post-success route efficiency separately. A current-view waypoint route provides a verifiable joint spatial-decision output and a local planning representation that can be used within waypoint-based zero-shot navigation systems. Evaluating this representation through a shared observation and action interface enables direct comparison of foundation-VLM spatial decisions independently of other navigation modules.

To align the model observation, displayed actions, and their consequences, we anchor every waypoint to a scene location and construct Point and Embodied feasibility graphs from scene geometry. Targets, visible waypoints, legal edges, acceptable endpoints, and reference routes are fixed before prompt generation. Every route question retains at least one reference route that passes the construction and geometry checks and projects into the current view. The model receives only the RGB

image with numbered waypoints and a natural-language prompt, while its prediction is evaluated using candidate feasibility, adjacent-edge legality, and goal arrival defined in the same scene.

EgoPathBench contains 31,852 training, 1,345 validation, and 1,111 benchmark questions. In addition to the benchmark split for unified model comparison, we release geometry-grounded training answers, legal reference routes, and Spatial CoT supervision, supporting both evaluation and the study of learned first-person spatial decision-making. We evaluate nine foundation VLMs, analyze where their waypoint routes fail, and examine the benchmark's dependence on its paired visual interface and scene-grounded evaluator. We also train a Qwen 3.5 4B model to test whether the released supervision transfers to improved performance on EgoPathBench and external spatial tasks.

The results reveal a substantial gap in current models' integrated spatial decision-making. The highest EgoPath Score across nine VLMs is 28.3. The top-ranked model reaches 35.9% success on Point Path, but only 2.9% and 4.0% on Embodied Path and Intent Path, respectively, showing a marked decline when a route must jointly satisfy embodiment and goal conditions. Fine-tuning Qwen 3.5 4B on the released training split raises its EgoPath Score from 3.9 to 38.9 and improves all four reported evaluations across three external spatial benchmarks, with gains of 1.4–9.6 points. These results show that EgoPathBench differentiates current models' first-person spatial decisions and that the released data provides effective supervision for this ability.

The contributions are:

- We introduce EgoPathBench, which uses five first-person waypoint-decision tasks to evaluate the integrated spatial intelligence of foundation VLMs through controlled, action-valued outputs.
- We construct and release 31,852 training, 1,345 validation, and 1,111 benchmark questions by aligning rendered observations, displayed waypoints, and resolved targets with Point and Embodied feasibility graphs, together with legal reference routes and Spatial CoT supervision.
- We systematically evaluate nine foundation VLMs, identify a major capability gap in embodied route decisions, and evaluate the released training resource on EgoPathBench and three external spatial benchmarks.

2 Related Work

Foundation-model navigation and evaluation. R2R, REVERIE, RxR, and VLN-CE established vision-language navigation in discrete and continuous indoor environments Anderson et al. (2018); Qi et al. (2020); Ku et al. (2020); Krantz et al. (2020). Foundation-model agents such as NavGPT, NaviLLM, and NaVid subsequently used language reasoning, generalist embodied modeling, and video-based planning for navigation Zhou et al. (2024); Zheng et al. (2024); Zhang et al. (2024). Recent zero-shot systems organize spatial decisions through different interfaces. SmartWay combines waypoint prediction with history-aware backtracking; VLFM, InstructNav, and CA-Nav use occupancy or value maps and sub-instruction constraints; AgenticNav and P2DNav expose pixel-level or hierarchical direction-to-grounding actions; and DreamNav predicts trajectories rather than isolated points Shi et al. (2025); Yokoyama et al. (2024); Long et al. (2024); Chen et al. (2025); Li et al. (2026); Sheng et al. (2026); Wang et al. (2025). Open-Nav studies spatio-temporal reasoning with open-source models, Nav-R1 learns structured navigation traces, and LHPR-VLN extends evaluation to decision consistency across long-horizon subtasks Qiao et al. (2025b); Liu et al. (2025); Song et al. (2025). Across full navigation systems, outcomes can depend on different combinations of candidate generation, mapping, memory, control, and replanning over repeated observations. EgoPathBench measures a controlled capability within this process: given the current first-person observation and a shared vocabulary of visible spatial actions, can a foundation VLM form a geometrically feasible and goal-consistent waypoint decision?

NaviTrace is the closest evaluation to our setting. It presents a single real RGB image, a navigation instruction, and an embodiment description, and asks a VLM to produce a continuous trace in image space Windecker et al. (2025). The prediction is scored by Dynamic Time Warping against human-annotated traces, endpoint error, and embodiment-conditioned penalties derived from pixel semantics. NaviTrace therefore measures agreement with expert navigation demonstrations under

a score that combines trace similarity, endpoint accuracy, and embodiment-conditioned semantic penalties.

EgoPathBench differs in the supervision associated with each observation. Each image is registered to an underlying 3D scene representation, from which we derive visible waypoints, agent-specific feasibility, direct traversability between waypoints, and acceptable goal regions. This scene backing allows an arbitrary predicted route to be evaluated by the geometric outcomes of its selected actions: whether its waypoints are feasible, whether every consecutive transition satisfies the specified agent constraints, and whether the route reaches the goal. A reference route certifies that a geometrically valid solution exists, but is not the unique trajectory that a prediction must imitate. Thus, NaviTrace evaluates trajectory agreement defined by expert demonstrations and image semantics, whereas EgoPathBench evaluates waypoint-action outcomes defined by scene geometry.

Spatial intelligence and embodied reasoning. Existing spatial-intelligence benchmarks examine complementary components of foundation VLMs' spatial ability. SpatialVLM and SpatialEval emphasize distance, direction, and spatial relations; VSI-Bench studies video-based spatial understanding and memory; ViewSpatial evaluates multi-perspective spatial localization; and 3DSRBench focuses on 3D structure Chen et al. (2024); Wang et al. (2024); Yang et al. (2025); Li et al. (2025a); Ma et al. (2025). EmbSpatial-Bench, EgoThink, and OpenEQA introduce first-person or embodied observations Du et al. (2024); Cheng et al. (2024); Majumdar et al. (2024). Embodied3DBench further evaluates low-level embodied skills including grounding, affordance, and trajectory prediction; CapNav studies capability-conditioned navigation under agent-specific mobility constraints; and IndustryNav evaluates active planning and collision-aware navigation in dynamic industrial environments Zhang et al. (2026); Su et al. (2026); Li et al. (2025b). Together, these benchmarks show that embodied spatial intelligence involves not only recognizing scene relations, but also determining where an agent can act and what spatial consequences its actions produce.

EgoPathBench organizes these abilities into a joint decision with an explicit action interpretation. A model must relate the target, candidate locations, agent constraints, and route structure in the current observation, then return either a traversable waypoint set or an ordered route. The prediction is evaluated through geometry-defined candidate feasibility, consecutive-edge legality, and goal attainment. EgoPathBench thereby connects low-level embodied spatial cues to ordered first-person waypoint decisions with explicit scene-grounded consequences.

3 EgoPathBench Dataset and Benchmark

3.1 Benchmark Formulation

EgoPathBench represents navigation decisions as selections over a visible waypoint vocabulary. Each example provides an egocentric RGB image with numbered candidate locations and a natural-language task. The model returns a JSON list of display IDs. For the two traversability tasks, the list denotes an unordered set of traversable candidates. For the three path tasks, it denotes an ordered route from a specified start to an acceptable goal.

Formally, an example consists of an egocentric observation o, a displayed waypoint set A, a target specification t, and a task-specific feasibility graph G. Every display ID in A is anchored to a 3D scene location. Given o, A, and t, the model predicts either a subset of A or an ordered sequence over A. Traversability tasks evaluate the feasibility of selected vertices. Route tasks additionally require the specified start, legal consecutive edges in G, and an endpoint in the acceptable goal set.

Point and Embodied tasks share the same observation and waypoint vocabulary but use different feasibility graphs. Predictions are evaluated on the corresponding Point or Embodied graph. The Point graph is constructed without footprint expansion, while the Embodied graph accounts for the 0.6 m robot diameter specified in the task prompt. The same visually selected action can therefore have different consequences under the two evaluation rules.

EgoPathBench measures joint waypoint decision making from limited first-person visual evidence. The model must identify the target, associate displayed IDs with scene locations, and select actions that jointly satisfy traversability, embodiment, route continuity, and goal-reaching requirements. The

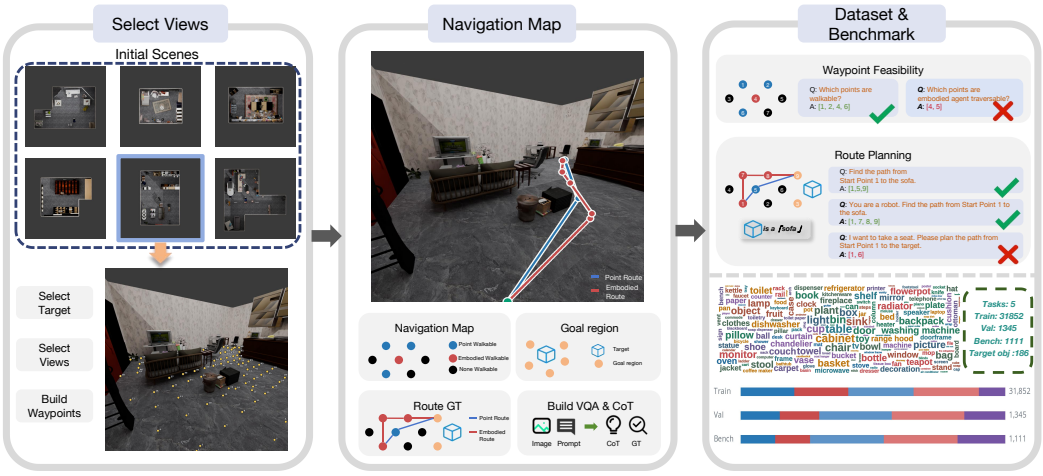

Figure 2: EgoPathBench construction pipeline. We select scenes, first-person views, and targets, then build visible waypoints, point- and embodiment-specific navigation maps, goal regions, and geometry-verified reference routes. These fixed scene annotations support image-question and Spatial CoT generation, yielding two waypoint-feasibility tasks and three route-planning tasks across the training, validation, and benchmark splits.

model acts from the current observation, while the consequences of its choices are evaluated against the registered scene geometry.

3.2 Five-Task Design

EgoPathBench contains two traversability tasks and three route tasks. The five tasks progressively introduce target grounding, embodiment, and route-composition requirements while retaining a common input and output interface. Table S1 in Appendix A summarizes the five tasks.

Point Traversability asks the model to select all walkable candidates. **Embodied Traversability** uses the same scene view and candidate space and specifies a robot diameter of 0.6 m, requiring sufficient clearance at the selected locations. This paired design tests how embodiment changes action feasibility within the same observation.

Point Path provides an explicit target and asks for an ordered route to it. **Embodied Path** holds the target, view, and candidate space fixed, adds the robot dimensions to the prompt, and evaluates the route on the Embodied graph. **Intent Path** replaces the explicit object name with an intent and a visual cue. It therefore requires the model to resolve the intended object before selecting an embodied route.

3.3 Data Construction and Annotation

Figure 2 summarizes the construction process. We first establish scene provenance and first-person scene–view records, then generate visible waypoints and geometric task labels. Natural-language questions and Spatial CoT are produced only after the target, action space, and reference route have been fixed. Quality filtering and challenge-set selection produce the final release.

Scenes, views, and targets. We use unified simulatable indoor assets normalized through InternScenes Zhong et al. (2025). The scenes retained in our release derive from 3RScan Wald et al. (2019), ScanNet Dai et al. (2017), ARKitScenes Baruch et al. (2021), and Matterport3D Chang et al. (2017). We use these assets for scene geometry and rendering, and construct our own first-person observations, waypoint annotations, route labels, and question data. For each scene, we sample first-person cameras in traversable space and remove views whose immediate camera neighborhood is obstructed. Target instances are then filtered by frustum projection, observation distance, projected size, and visible surface evidence, excluding targets that are outside the image, too small, or heavily occluded. Shared camera parameters and scene geometry bind each retained target and RGB obser-

vation to a common scene–view record, which provides the spatial basis for subsequent waypoint and route annotation.

Waypoints and geometric labels. For each retained view, we project two types of markers into the image: ground action candidates and selected visible object- or structural-surface locations used as non-traversable negatives. Depth, ray-visibility, and marker-spacing checks associate every displayed ID with a distinct scene location while reducing marker overlap and foreground occlusion. Route sheets use visible ground waypoints and explicitly insert the waypoints required by retained reference routes. We evaluate ground actions under both point and embodied geometry. Traversability examples therefore mix feasible ground actions, clearance-sensitive ground actions, and visible surface negatives, so the task cannot be solved by selecting every displayed ID. Point and Embodied Traversability share the same view, while their labels follow the corresponding agent-feasibility conditions. Each displayed candidate stores its ID, image projection, and scene coordinate; the numbered RGB overlay and the geometric annotations therefore refer to the same action vocabulary.

Paired route construction. For each route unit, we anchor the start on visible floor near the bottom of the image and generate feasible goal locations around the target footprint. We then search separately through point and embodied free space, convert each continuous path into sparse route waypoints, and project the required waypoints back into the current image. Construction checks route connectivity, task-specific feasibility, arrival in the target region, and the availability of every required waypoint through the displayed action interface. Routes that fail these checks are discarded. Point Path and Embodied Path consequently share the target, start, and scene view while retaining routes validated for their respective agent geometries. Every retained route question has at least one scene-verified reference route to its target.

Question text and Spatial CoT. Target identity, displayed waypoints, goal region, and reference-route identity are fixed before language generation. Explicit-target tasks name the required object directly. For Intent Path, we construct a view-specific candidate universe from visible objects and generate a target description using supported relations, attributes, colors, or distance cues. The resulting description is resolved again against the objects available in the current view. We retain it only when it uniquely identifies the fixed target without unsupported cues. Intent Path is derived from an accepted Embodied Path instance, so changing the linguistic specification does not change its target or route geometry. The training split additionally includes Spatial CoT. GPT-5.5 verbalizes the fixed target, candidates, feasibility labels, legal edges, and reference route into task-specific reasoning text. The exported answer is checked against the formal annotation, keeping language generation downstream of the geometric ground truth.

3.4 Quality Control

Quality control applies geometry and view–action checks before language is attached. Candidate locations, route edges, and target regions share one scene coordinate system; disconnected or colliding routes, routes that miss the target region, and examples without required display waypoints are removed. The RGB image, waypoint overlay, and ID-to-waypoint record come from the same rendered view, and point and embodied routes are validated under their corresponding agent geometries. Prompts may describe only targets and scene facts fixed during construction. Explicit-target questions are checked against the selected target, while intent questions must resolve uniquely among objects visible from the current view. A full visual–language alignment audit then checks targets, prompts, displayed waypoints, reference routes, and answers across the formal release. Ambiguous or unsupported prompts are repaired or removed without changing the established geometric labels or route identities.

3.5 Benchmark Selection and Release

After constructing the full question pool, we select the formal benchmark as a challenging subset emphasizing scene clutter, traversability boundaries, point–embodied feasibility differences, competing targets, and route composition. Selection operates on complete task bundles. Point and Embodied Traversability form a paired view bundle; Point Path and Embodied Path form a paired route bundle with a shared target and candidate space; eligible Intent Path questions accompany their corresponding embodied routes. Bundle-level selection prevents the benchmark from retaining only one side of a paired comparison. We then select scenes while balancing task coverage, geometric difficulty, and

		Point Trav.		Embodied Trav.		Point Path			Embodied Path			Intent Path		
Model	Score	BA	F1	BA	F1	VPR	SR	SPL	VPR	SR	SPL	VPR	SR	SPL
Zero-shot foundation VLMs														
Gemini 3.1 Pro	**28.3**	76.5	**79.0**	72.7	56.2	63.7	**35.9**	**28.7**	9.1	**2.9**	**2.4**	10.4	**4.0**	**3.3**
GPT-5.5	27.3	74.1	77.1	**77.3**	**62.5**	60.5	31.1	24.8	5.5	1.3	1.3	9.0	1.5	1.4
Claude Opus 4.8	25.6	**77.9**	74.3	73.5	59.6	**66.0**	21.0	16.7	**12.0**	1.6	1.5	**14.4**	2.5	2.2
MiniMax M3	21.8	77.0	76.2	68.7	52.8	50.8	13.3	8.7	5.5	1.9	1.7	9.0	2.5	2.3
Qwen 3.6	16.4	67.4	72.1	64.8	49.1	49.2	15.2	10.9	2.6	1.0	0.9	5.0	1.5	1.3
Mistral L3	15.8	65.2	70.8	68.5	52.5	35.6	11.0	5.8	0.3	0.0	0.0	3.5	0.5	0.5
Llama 4	14.9	66.8	66.5	66.0	49.8	36.2	8.7	5.9	1.9	0.0	0.0	2.0	0.0	0.0
Kimi K2.6	9.7	60.2	69.8	57.8	44.2	40.5	11.7	8.2	1.3	0.7	0.5	1.5	0.5	0.4
Grok 4.3	1.4	52.3	58.4	50.0	35.8	18.4	2.6	1.2	5.2	0.0	0.0	3.5	0.0	0.0
Training-resource study														
Qwen3.5-4B base	3.9	54.6	55.6	54.9	33.1	9.1	0.7	0.1	0.3	0.0	0.0	0.0	0.0	0.0
+ EgoPathBench SFT	38.9	89.3	89.2	83.4	71.2	77.0	31.4	28.6	34.9	7.1	6.9	44.8	10.4	10.0

Table 1: EgoPathBench leaderboard (%). The upper block compares zero-shot foundation VLMs; the lower block compares Qwen3.5-4B before and after training on EgoPathBench. Score is the equal-weight five-task macro-average. BA/F1 evaluate traversability, while VPR, SR, and SPL evaluate route validity, goal-reaching success, and efficiency.

scene diversity and limiting concentration within InternScenes. Splits are formed by source group, so the same original scan or related regions do not cross training, validation, and benchmark sets. The final release contains 31,852 training questions, 1,345 validation questions, and 1,111 benchmark questions across the five first-person waypoint-decision tasks.

4 Experiments

4.1 Experimental Setup

The benchmark contains 1,111 questions: 146 Point Traversability, 146 Embodied Traversability, 309 Point Path, 309 Embodied Path, and 201 Intent Path. Traversability uses candidate-level balanced accuracy (BA), the mean recall across the task-feasible and infeasible classes, and F1, the precision–recall harmonic mean for feasible candidates. For route tasks, valid path rate (VPR) measures complete route legality, success rate (SR) additionally requires an acceptable endpoint, and success weighted by path length (SPL) discounts successful routes that are longer than the shortest legal reference while assigning zero to failures. EgoPath Score equally averages chance-adjusted traversability BA and route SR across the five tasks. Formal definitions are provided in the supplementary material.

We evaluate nine foundation VLMs. Every model receives the same image, task prompt, visible waypoint IDs, and JSON output protocol, and every prediction is scored by the same evaluator. Endpoint identifiers, generation settings, training hyperparameters, and control-specific protocols are provided in the supplementary material.

4.2 Zero-Shot VLM Results

Table 1 gives the complete zero-shot leaderboard. Gemini 3.1 Pro ranks first with an EgoPath Score of 28.3, followed by GPT-5.5 at 27.3 and Claude Opus 4.8 at 25.6.

Candidate-level traversability is markedly stronger than complete route construction. The best Point and Embodied Traversability BA values are 77.9% and 77.3%, with best F1 scores of 79.0% and 62.5%. For Point Path, the strongest VPR is 66.0%, but SR falls to 35.9% and SPL to 28.7%. Thus, plausible local actions do not by themselves produce a legal, goal-reaching route. Tasks with embodiment constraints produce the largest drop. On Embodied Path, the best VPR, SR, and SPL are 12.0%, 2.9%, and 2.4%; on Intent Path, they are 14.4%, 4.0%, and 3.3%. The VPR–SR gap

indicates complementary failures in edge legality and endpoint selection, while low SPL largely reflects scarce complete successes.

Figure 4 places the zero-shot leaderboard against a same-question human reference under the identical task interface. The human profile is higher on all five task axes, exceeding the strongest VLM on these questions by 25.6 points in EgoPath Score. The human advantage appears in both candidate-feasibility judgments and all three route tasks, indicating that the gap is not driven by any single task or metric but reflects a broader limitation in first-person spatial decision-making. Figures S1–S10 in Appendix D show same-question human and model outputs across all five tasks.

4.3 Single-Image Geometric Answerability

To assess whether path feasibility can be judged from a single RGB image, we conduct a human audit of model-predicted paths. Four participants independently judge whether each of 40 paths is traversable using only the original egocentric RGB image, numbered waypoints, the candidate path, and the specified embodiment constraints, without access to depth, 3D scene geometry, or validity labels. Participants P01–P04 correctly judge 39, 32, 35, and 34 paths, respectively, totaling 140 correct judgments out of 160. As shown in Table 2, human judgments achieve **91.25%** accuracy on valid paths and **83.75%** on invalid paths, with **87.5%** overall accuracy. These results show strong agreement between image-based human judgments and geometry-derived labels, providing direct empirical support for the visual answerability of our path-feasibility evaluation. Appendix B provides supporting geometric validation and visual-input controls.

Evaluation setting	Valid-path accuracy	Invalid-path accuracy	Overall accuracy
Human, RGB only	91.25%	83.75%	**87.5%**

Table 2: RGB-only human assessment of geometric path validity by four participants (40 judgments each).

4.4 Where Complete Routes Fail

Figure 3 decomposes route predictions into five parallel diagnostics, ordered from output validity and target-endpoint selection to the legality of the initial action, the complete route, and their joint success. Across tasks, 96.3–96.8% of outputs are evaluable routes, so the output protocol is not the main bottleneck. Goal-consistent endpoint rates are much lower at 28.9% for Point Path, 4.8% for Embodied Path, and 5.2% for Intent Path, exposing substantial difficulty in grounding the requested target to a terminal waypoint. A model may recognize where the target object is yet fail to identify a nearby terminal waypoint that is feasible for the specified agent.

Initial action selection is considerably stronger: 96.0%, 90.4%, and 88.0% of predictions take a legal first edge. This local feasibility does not carry through the selected sequence. Full-route legality falls to 46.8% for Point Path, 4.8% for Embodied Path, and 6.5% for Intent Path; among predictions with a legal first edge, 51.3%, 94.7%, and 92.7% contain an illegal later edge. Joint success, which requires both complete-route legality and a goal-consistent endpoint, is only 16.7%, 1.0%, and 1.4%. The diagnostics therefore expose two distinct limitations: selecting the intended terminal waypoint and maintaining agent-specific feasibility from the current position to that endpoint.

This suffix failure persists after controlling for both ends of the decision. Among predictions with a legal first edge and acceptable goal, 42.2% of Point, 75.8% of Embodied, and 69.4% of Intent routes still contain an illegal intermediate edge. On the half of questions with fewer displayed waypoints, rates remain similar at 42.8%, 76.6%, and 67.3%, so the pattern is not confined to dense action overlays. Illegal-edge incidence is already 90.6% and 88.4% on one-edge-reference Embodied and Intent questions, rising to 94.0% and 92.4% for references with at least three edges. The central signal is whether the intermediate sequence remains valid under agent geometry; additional model-level and length-stratified results appear in the supplementary material.

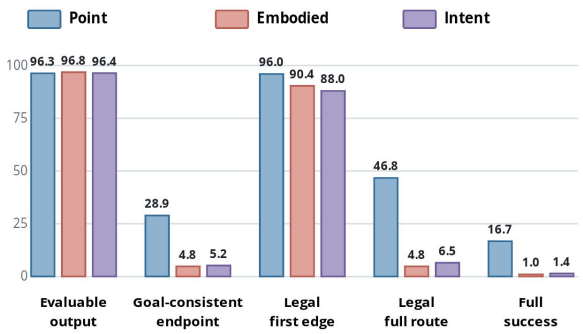

Figure 3: Route diagnostics pooled over nine VLMs (%). Rates over all predictions measure evaluable output, goal-consistent endpoint selection, legal first action, full-route legality, and joint success. Each group compares Point, Embodied, and Intent paths.

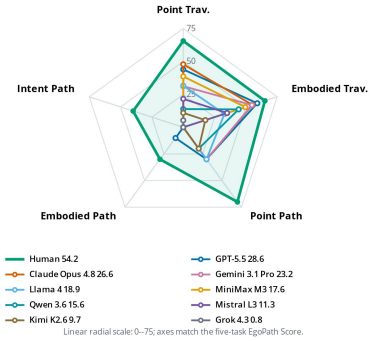

Figure 4: Humans versus nine VLMs on the same questions. Axes show $2\mathrm{BA} - 1$ for traversability and route success rates; the legend reports EgoPath Score.

4.5 Training-Resource Evaluation

We fine-tune Qwen3.5-4B on the released EgoPathBench training split. The base model and selected checkpoint use the same EgoPathBench contract and evaluation items from VSI-Bench Route Planning, SpatialEval-VTQA, and 3DSRBench Yang et al. (2025); Wang et al. (2024); Ma et al. (2025). This evaluates the released training data in- and out-of-domain.

The lower block of Table 1 reports the in-domain comparison. The fine-tuned checkpoint improves every reported metric and reaches an EgoPath Score of 38.9.

Benchmark	Setting	Base	SFT	Δ
VSI-Bench Route Planning	Full	29.38	33.51	+4.13
	Debiased	20.18	24.56	+4.38
SpatialEval-VTQA	Full	61.8	71.4	+9.6
3DSRBench	Full	58.0	59.4	+1.4

Table 3: Performance of Qwen3.5-4B on three spatial benchmarks before and after supervised fine-tuning (SFT) on EgoPathBench (%).

The selected checkpoint improves all reported external evaluations. On VSI-Bench Route Planning, performance rises from 29.38% to 33.51% in the Full setting (+4.13) and from 20.18% to 24.56% in the Debiased setting (+4.38). Under the Full settings, accuracy also increases by 9.6 points on SpatialEval-VTQA and 1.4 points on 3DSRBench, showing transfer to spatial tasks outside EgoPathBench.

5 Conclusion

EgoPathBench evaluates whether VLMs turn first-person observations and goals into feasible, goal-reaching actions across five scene-grounded tasks spanning feasibility, grounding, route legality, goal arrival, and efficiency.

Across nine VLMs, the best EgoPath Score is 28.3; the leader reaches 35.9% Point Path success but only 2.9% and 4.0% on Embodied and Intent Path. Fine-tuning improves Qwen 3.5 4B from 3.9 to 38.9 and yields gains on all four reported evaluations across three external spatial benchmarks, supporting both evaluation and training. The gap between locally plausible actions and complete embodied routes identifies sustained geometric and goal consistency across multi-step decisions as a central direction for future VLM research.

AI USE STATEMENT

Generative AI was used to assist with writing and polishing the manuscript and to generate synthetic dataset content. For the latter, the Data Construction and Annotation section and the Training-resource evaluation paragraph in Appendix A describe the use of GPT-5.5 to verbalize fixed targets, candidates, feasibility labels, legal edges, and reference routes into Spatial CoT supervision. As described there, target identity and geometric annotations were fixed before language generation, and exported answers were checked against the formal annotations. The authors take responsibility for the final content of this work, including AI-assisted text and dataset artifacts.

REPRODUCIBILITY STATEMENT

Detailed implementation information is provided in the appendices. The code and data will be made publicly available upon acceptance of the paper.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, 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*, 2018.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Yuxin Cai, Xiangkun He, Maonan Wang, Hongliang Guo, Wei-Yun Yau, and Chen Lv. Cl-cotnav: Closed-loop hierarchical chain-of-thought for zero-shot object-goal navigation with vision-language models. *arXiv preprint arXiv:2504.09000*, 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. In *Proceedings of the International Conference on 3D Vision*, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Kehan Chen, Dong An, Yan Huang, Rongtao Xu, Yifei Su, Yonggen Ling, Ian Reid, and Liang Wang. Constraint-aware zero-shot vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. *arXiv preprint arXiv:2311.15596*, 2024.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Proceedings of the European Conference on Computer Vision*, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

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

Yifan Li, Lichi Li, Anh Dao, Xinyu Zhou, Wenjun Huang, Tianyi Ma, Yicheng Qiao, Zheda Mai, Daeun Lee, Zichen Chen, Pan Wang, Lehan Yang, Tianlong Wang, Zhen Tan, Sheng Li, Mohit Bansal, Yang Ni, and Yu Kong. Industrynav: Exploring spatial reasoning of embodied agents in dynamic industrial navigation. *arXiv preprint arXiv:2511.17384*, 2025b.

Yijian Li, Changze Li, Hantian Shi, Jiaying Luo, Jiyuan Cai, Ming Yang, and Tong Qin. Agenticnav: Zero-shot vision-and-language navigation as a tool-calling harness. *arXiv preprint arXiv:2606.10577*, 2026.

Qingxiang Liu, Ting Huang, Zeyu Zhang, and Hao Tang. Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884*, 2025.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. In *Proceedings of the Conference on Robot Learning*, 2024.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M. de Melo, Alan Yuille, and Jieneng Chen. 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Arjun Majumdar et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Yanyuan Qiao, Haodong Hong, Wenqi Lyu, Dong An, Siqi Zhang, Yutong Xie, Xinyu Wang, and Qi Wu. Navbench: Probing multimodal large language models for embodied navigation. In *Advances in Neural Information Processing Systems*, 2025a.

Yanyuan Qiao, Wenqi Lyu, Hui Wang, Zixu Wang, Zerui Li, Yuan Zhang, Mingkui Tan, and Qi Wu. Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2025b.

Kai Sheng, Liuyi Wang, Haojie Dai, Jinlong Li, Yongrui Qin, Zongtao He, Chengju Liu, and Qijun Chen. P2DNav: Panorama-to-downview reasoning for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2605.19634*, 2026.

Xiangyu Shi, Zerui Li, Wenqi Lyu, Jiatong Xia, Feras Dayoub, Yanyuan Qiao, and Qi Wu. Smartway: Enhanced waypoint prediction and backtracking for zero-shot vision-and-language navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2025.

Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Xia Su, Ruiqi Chen, Benlin Liu, Jingwei Ma, Zonglin Di, Ranjay Krishna, and Jon Froehlich. Capnav: Benchmarking vision language models on capability-conditioned indoor navigation. *arXiv preprint arXiv:2602.18424*, 2026.

Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. 2024. URL https://arxiv.org/abs/2406.14852.

Yunheng Wang, Yuetong Fang, Taowen Wang, Yixiao Feng, Yawen Tan, Shuning Zhang, Peiran Liu, Yiding Ji, and Renjing Xu. Dreamnav: A trajectory-based imaginative framework for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2509.11197*, 2025.

Tim Windecker, Manthan Patel, Moritz Reuss, Richard Schwarzkopf, Cesar Cadena, Rudolf Lioutikov, Marco Hutter, and Jonas Frey. Navitrace: Evaluating embodied navigation of vision-language models. *arXiv preprint arXiv:2510.26909*, 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. VLFM: Vision-language frontier maps for zero-shot semantic navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2024.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *Robotics: Science and Systems*, 2024.

Jiyao Zhang, Mingxu Zhang, Yitong Peng, Haoxuan Liu, Chenshuo Wang, Yuxing Long, Haoyang Huang, Dongjiang Li, Nan Duan, Hui Shen, and Hao Dong. Embodied3dbench: Benchmarking low-level embodied spatial intelligence of vision language models. *arXiv preprint arXiv:2605.29074*, 2026.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Weipeng Zhong, Peizhou Cao, Yichen Jin, Luo Li, Wenzhe Cai, Jingli Lin, Hanqing Wang, Zhaoyang Lyu, Tai Wang, Bo Dai, Xudong Xu, and Jiangmiao Pang. Internscenes: A large-scale simulatable indoor scene dataset with realistic layouts. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

A Evaluation Protocol and Reproducibility

Task	Goal	Evaluation graph	Output	Scored constraints	N
Point Trav.	–	Point	Set	Traversability	146
Embodied Trav.	–	Embodied	Set	Feasibility	146
Point Path	Explicit	Point	Route	Edges, endpoint	309
Embodied Path	Explicit	Embodied	Route	Footprint, edges, goal	309
Intent Path	Intent	Embodied	Route	Intent, footprint, edges	201

Table S1: Five EgoPathBench tasks. The shared first-person waypoint interface is specialized by goal specification, agent geometry, output structure, evaluated constraints, and benchmark question count.

Table S2 summarizes the observation, action, and scoring contract for each task. Traversability tasks require a set of locally feasible waypoint actions. Route tasks require an ordered sequence beginning at the specified start; every consecutive edge must be legal for the task's agent geometry, and the final waypoint must belong to the acceptable goal set.

Task	Decision input	Output	Task specification	Primary scoring condition
Point Traversability	RGB with point candidates	ID set	Walkable candidates	Candidate-wise traversability classification
Embodied Traversability	RGB with footprint-aware candidates	ID set	0.6 m body; no goal	Candidate-wise body-feasible classification
Point Path	RGB, start ID, explicit target	Ordered route	Explicit goal	Valid IDs, required start, legal edges, acceptable endpoint
Embodied Path	RGB, start ID, explicit target	Ordered route	0.6 m body; explicit goal	Point-Path checks plus footprint-aware edge legality
Intent Path	RGB, start ID, natural-language intent	Ordered route	0.6 m body; resolved goal	Intent-consistent endpoint and embodied route legality

Table S2: Interface and scoring contract for the five EgoPathBench tasks. All outputs use visible display IDs from the current first-person image.

Prompt and output protocol. Every question stores the system prompt, user prompt, image, and visible-waypoint record used for evaluation. Table S3 gives the released prompt templates. Bracketed fields are filled from the fixed question record; the requested answer is always one JSON array of visible display IDs.

Task	System prompt	User-prompt template
Point Traversability	You are a navigation perception assistant.	The image shows numbered candidate points. Output a JSON array of display IDs that are walkable.
Embodied Traversability	You are a navigation perception assistant.	The image shows numbered candidate points. The robot has diameter 0.6m. Output a JSON array of display IDs that are walkable for the robot.
Point Path	You are a navigation planning assistant.	The image shows numbered candidate points. Start at display ID 1. Navigate to the [explicit target]. Output a JSON array of display IDs representing a valid path.
Embodied Path	You are an embodied navigation robot. Plan a collision-free path for your body using the numbered candidate points shown in the image.	The image shows numbered candidate points. You are the robot, and your body diameter is 0.6 m. Start at display ID 1. Navigate to the [explicit target]. Return a JSON array of display IDs representing a valid collision-free path for the robot.
Intent Path	You are an embodied navigation robot. A user gives you a short request that implies the kind of object they want. Infer the target family from the request, resolve the final grounded target in the scene, and plan a collision-free path for your body.	The image shows numbered candidate points. You are the robot, and your body diameter is 0.6 m. Start at display ID 1. A person says: "[intent request]." Return a JSON array of display IDs representing a valid collision-free path to the resolved target.

Table S3: Released prompt templates for the five tasks. Bracketed fields are populated from the fixed question record.

Route evaluator. The released sidecar applies ordered candidate-validity, start, direct-edge, and goal-membership checks; SPL is computed only after a route satisfies this success contract. Table S4 lists the complete check sequence.

Stage	Pass condition	Outcome
Parse	One ordered ID list is recovered.	Format
Candidate	Every ID is visible and allowed.	Invalid ID
Start	The first ID matches the required start.	Wrong start
Edge	Every consecutive pair is a legal direct edge.	Illegal edge
Endpoint	The final ID is an acceptable goal.	Wrong goal
Efficiency	A successful route is compared with shortest references.	SPL

Table S4: Ordered evaluation contract for route-bearing tasks. SPL is applied after route success.

Navigation maps. Navigation maps use a height-filtered 2D footprint approximation. Obstacle geometry intersecting the 0.10–1.40 m height interval is projected onto a grid with 0.05 m cells. The Embodied map uses a 0.30 m radius with no additional margin to expand obstacles and shrink the walkable floor region. When an obstacle mesh is unavailable, its expanded oriented bounding rectangle is used. The Point map uses the same geometric representation without footprint expansion.

Acceptable endpoints. For Embodied Path and Intent Path, we compute the 2D distance from each embodied-feasible candidate waypoint to the target's oriented bounding rectangle. Let $d_{\min}$ be the smallest such distance. The acceptable endpoint set contains all embodied-feasible candidates within $d_{\min} + 0.10$ m of the rectangle. For Point Path, the initial threshold is computed from point-feasible candidates in the same way and is increased, when needed, to include the nearest embodied-feasible candidate. This extension applies to 253 of the 309 Point Path questions. A predicted route reaches the goal when its final waypoint belongs to the corresponding acceptable endpoint set.

Metric definitions. Task-feasible candidates are the positive class for traversability. Let P and R denote precision and recall, and let V_i indicate that route prediction i is parseable, uses valid IDs, begins at the required start, and contains only legal consecutive edges. Let S_i additionally require

an acceptable endpoint. For N questions,

$$\mathrm{BA} = \frac{\mathrm{TPR} + \mathrm{TNR}}{2}, \qquad F1 = \frac{2PR}{P+R},$$
$$\mathrm{VPR} = \frac{1}{N}\sum_i V_i, \qquad \mathrm{SR} = \frac{1}{N}\sum_i S_i,$$
$$\mathrm{SPL} = \frac{1}{N}\sum_i S_i \frac{\ell_i}{\max(\ell_i, p_i)},$$

where ℓ_i is the shortest legal reference length to an acceptable goal and p_i is the predicted route length. Thus, unsuccessful routes receive zero SPL. With component metrics expressed in $[0, 1]$, the reported aggregate is

$$\mathrm{EgoPathScore} = \frac{100}{5}\big[(2\mathrm{BA}_{\mathrm{PT}} - 1) + (2\mathrm{BA}_{\mathrm{ET}} - 1) + \mathrm{SR}_{\mathrm{PP}} + \mathrm{SR}_{\mathrm{EP}} + \mathrm{SR}_{\mathrm{IP}}\big].$$

Foundation VLM evaluation. The nine endpoint identifiers are *claude-opus-4-8*, *gemini-3.1-pro-preview*, *gpt-5.5*, *grok-4.3-fast*, *kimi-k2.6*, *meta/llama-4-maverick-17b-128e-instruct*, *MiniMax-M3*, *mistralai/mistral-large-3-675b-instruct-2512*, and *qwen3.6-plus*. Each run uses an 8,192-token completion budget. Temperature is 0 and top-p is 1 where exposed; provider-native reasoning and unavailable controls retain their defaults. All endpoints receive the same task-specific prompt and JSON contract. Each leaderboard entry is one complete pass over the fixed 1,111 questions.

Training-resource evaluation. We train Qwen3.5-4B with LoRA rank 8, alpha 16, and zero dropout while freezing the vision tower. Stage one uses per-device batch size 1, gradient accumulation 8, a cosine schedule from 10^{-4} with 10% warmup, two epochs, and seed 42. Continuation from the final adapter uses two further epochs at 5×10^{-5} with 5% warmup. Both stages use bfloat16, a 4,096-token cutoff, and maximum image area 262,144 pixels. We evaluate the model saved after 3,000 steps of continued training. EgoPathBench and external evaluation use completion budgets of 8,192 and 4,096 tokens, respectively.

External evaluation uses the VSI-Bench Route Planning subset in both its Full and Debiased settings, together with SpatialEval-VTQA and 3DSRBench. The base model and selected checkpoint are evaluated on the same items and with the same protocol within each setting. VSI-Bench provides the navigation-focused video evaluation, while the other two benchmarks test transfer to complementary spatial tasks.

The SFT export contains one image-grounded conversation for every training question. Its assistant response contains the accepted GPT-written Spatial CoT followed by the geometry-verified JSON answer; format repair may normalize the surrounding tags and final answer line but does not replace the rationale. Table S5 reports the release audit.

Field	Value
Complete training rows	31852 / 31852
Missing image / rationale	0 / 0
Rationale source	GPT-written, non-template
Format and answer audit	Pass

Table S5: Audit of the released EgoPathBench SFT export.

Software, compute, and release. Scene rendering and geometric annotation use Blender 4.4 and Python 3.12. Model training uses Python 3.10.20, PyTorch 2.12.0 with CUDA 13.0, Transformers 5.2.0, PEFT 0.15.1, Accelerate 1.6.0, and LLaMA-Factory on two NVIDIA A100 80GB GPUs. Upon publication, we will release the generated questions, annotations, training supervision, task prompts, construction and evaluation code, model configurations, and fixed prediction records under licenses permitting research use and consistent with the terms of the upstream scene assets.

Aggregate-score sensitivity. The primary EgoPath Score uses the equal-weight task macro-average defined in the main paper. Replacing route SR with SPL, or first averaging within the traversability and route families and then weighting the two families equally, preserves the complete nine-model ordering (Spearman $\rho = 1.0$ for both alternatives).

B Geometric Evaluation Validation

Human-audit sampling. For the human audit in Section 4.3, we select 40 subpaths from model predictions on the Embodied Path task, comprising 20 geometrically valid and 20 invalid subpaths. The sample contains 14 subpaths with two waypoints, 14 with three waypoints, and 12 with four waypoints. All four participants judge the same 40 subpaths using the RGB-only protocol described in that section.

B.1 Reference-Route Admission and Target Visibility

Reference-route admission. Reference routes are checked for projection into the current view and against the visible-obstacle mask before admission. Table S6 reports these checks for all 819 route questions: every dense reference point projects into the view, and both sparse and dense routes pass the mask check. Target bounding-box size and visible-surface statistics characterize the corresponding target observations.

Task	N	Path proj.	Mask pass	BBox px	Surface
Point	309	100.0	100.0	37/82	32.7/56.2
Embodied	309	100.0	100.0	37/82	32.7/56.2
Intent	201	100.0	100.0	49/93	31.2/53.1

Table S6: Reference-route admission checks for 819 route questions. Path projection covers every dense reference point; mask pass checks sparse and dense routes against the visible-obstacle mask. Target columns report P10/median.

Target-visibility analysis. Beyond these reference-route admission checks, we analyze model performance on a subset with stricter target visibility. The strict visibility subset requires a target bounding-box short side of at least 64 pixels and at least 50% visible surface. Table S7 compares the full set with the 118 Point, 118 Embodied, and 83 Intent questions satisfying these conditions. The endpoint-hit versus full-route-success gap remains across all three tasks, ranging from 4.3 to 13.3 percentage points on the strict subsets. Illegal-edge rates are 55.9%, 92.7%, and 92.1% for Point, Embodied, and Intent, respectively.

Task	View	N	End/Full	Gap	Illegal
Point	All	309	28.9/16.7	12.2	49.5
Point	Strict	118	27.7/14.4	13.3	55.9
Embodied	All	309	4.8/1.0	3.7	92.0
Embodied	Strict	118	5.9/1.6	4.3	92.7
Intent	All	201	5.2/1.4	3.8	89.9
Intent	Strict	83	6.3/1.6	4.7	92.1

Table S7: Visibility sensitivity over nine VLMs (%). Strict additionally requires target bbox short side ≥ 64 px and visible surface $\geq 50\%$. End/Full reports endpoint hit/full-route success.

B.2 Scene-Consequence Audit and Evaluator Sensitivity

Scene-consequence audit. We audit one selected illegal edge from each of the 5,563 route predictions containing at least one geometry-defined illegal edge: 1,377 Point, 2,559 Embodied, and 1,627 Intent predictions. Along each selected edge, we sample the swept corridor at 0.05 m intervals using the formal 0.30 m embodied radius and search aligned depth and object-index renders for obstruction evidence. An explicit registered obstruction is recovered for 4,934 cases (88.7%): 4,662 (83.8%) are supported by depth and a further 272 (4.9%) by the object-index render alone.

Evaluator sensitivity. After checking the scene evidence for illegal edges, we examine sensitivity to geometric parameter settings. Geometry sensitivity varies the nominal 0.30 m radius, 0.05 m occupancy grid, and 0.10 m goal-distance tolerance as specified in Table S8. Across these variants, edge-label changes range from 0.0% to 8.5% and success-decision changes from 0.1% to 1.1%; the nine-model ranking remains unchanged (Spearman $\rho = 1.00$). The table also reports the fraction of questions retaining a graph solution under each variant.

Variant	Ref.	Edge flip	Success flip	Rank ρ
Radius 0.25 m	100.0	2.4	0.2	1.00
Radius 0.35 m	67.6	8.5	0.5	1.00
Grid 0.04 m	91.6	5.3	0.9	1.00
Grid 0.06 m	89.5	6.0	1.1	1.00
Goal tolerance 0.05 m	99.8	0.0	0.1	1.00
Goal tolerance 0.15 m	100.0	0.0	0.1	1.00

Table S8: Geometry sensitivity on 819 route questions and nine-model outputs. Ref. is the fraction retaining a graph solution; edge and success flips are measured against the nominal contract.

B.3 Visual-Input Dependence

The input-dependence control keeps the task and candidate count fixed on paired questions and compares the full input with text-only input and a mismatched image–marker overlay. Table S9 reports route success rates for each condition.

Model	Full	Text only	Mismatched overlay
GPT-5.5	30.0/0.0/4.0	0.0/0.0/0.0	10.0/0.0/0.0
Claude Opus 4.8	28.0/0.0/0.0	0.0/0.0/0.0	4.0/0.0/0.0
Qwen 3.6	16.0/0.0/0.0	0.0/0.0/0.0	2.0/0.0/0.0

Table S9: Input-dependence route SR (%) on fixed paired questions. Cells report Point/Embodied/Intent Path.

For Point Path, full-input success rates of 30.0%, 28.0%, and 16.0% fall to zero with text-only input and to 10.0%, 4.0%, and 2.0% with mismatched overlays for GPT-5.5, Claude Opus 4.8, and Qwen 3.6, respectively. GPT-5.5's Intent Path success rate falls from 4.0% to zero under both controls; the remaining Embodied and Intent entries are zero across conditions. These comparisons show that the successful route predictions depend on the paired visual input.

C Complete Dataset and Route Statistics

Release scale and route structure. Table S10 separates release scale from benchmark route structure. The scale block distinguishes scenes, scene–camera views, route units shared by paired questions, target instances, and questions. The structure block shows that route questions are not dominated by direct, unambiguous cases: they contain a median of four same-family objects, Embodied and Intent references require a median of two edges, and embodied routes spend a median 40% of their length in clearance-sensitive passages.

(a) Release scale

Split	Scenes	Views	Routes	Targets	Qs.
Train	2,942	6,483	7,843	6,144	31,852
Val	32	200	368	179	1,345
Benchmark	255	365	309	281	1,111

(b) Benchmark route structure

Signal	N	Median	P90	Max
Same-type target ambiguity	819	4	8	16
Reference segments (Point Path)	309	1	2	3
Reference segments (Body Path)	309	2	3	9
Reference segments (Intent Path)	201	2	3	9
Reference path length	819	3.2 m	4.8 m	8.5 m
Embodied narrow-passage fraction	510	40%	81%	100%

Table S10: Release scale and benchmark route structure. Views are scene–camera pairs; routes are scene–view–route tuples shared by paired route questions.

Failure after correct endpoints. Table S11 first selects predictions whose first edge is legal and whose endpoint is acceptable, then divides them into routes with an illegal later edge and routes whose complete edge sequence is legal. Of these predictions, 42.2% of Point, 75.8% of Embodied, and 69.4% of Intent routes still fail on a later edge. The rates remain nearly unchanged on the half of questions with fewer displayed waypoints, showing that dense waypoint overlays do not explain the failure.

(a) All benchmark questions

Task	First edge legal + endpoint correct	Illegal later edge	Complete route legal
Point	805	340 (42.2%)	465 (57.8%)
Embodied	120	91 (75.8%)	29 (24.2%)
Intent	85	59 (69.4%)	26 (30.6%)

(b) Fewer-waypoint subset (50% of questions)

Task	First edge legal + endpoint correct	Illegal later edge	Complete route legal
Point	495	212 (42.8%)	283 (57.2%)
Embodied	77	59 (76.6%)	18 (23.4%)
Intent	52	35 (67.3%)	17 (32.7%)

Table S11: Route outcomes after the first step and endpoint are correct, pooled over nine models. Each row divides these predictions into routes with an illegal later edge and routes whose complete edge sequence is legal. Panel (b) repeats the diagnostic on the half of questions with fewer displayed waypoints.

Effect of reference length. Table S12 stratifies route outcomes by the number of edges in the verified reference. Illegal-edge incidence rises with reference length, but it is already 90.6% for Embodied and 88.4% for Intent on questions whose reference contains only one edge, increasing to 94.0% and 92.4% for references with at least three edges. Route length therefore exacerbates, but does not by itself explain, the embodied-route failure.

Task	Edges	N	End hit	Success	Illegal edge
Point	1	250	29.5	17.8	47.2
Point	$\geq$2	59	26.7	12.1	59.5
Embodied	1	135	4.4	1.6	90.6
Embodied	2	90	5.2	0.6	92.2
Embodied	$\geq$3	84	4.9	0.5	94.0
Intent	1	91	5.4	2.3	88.4
Intent	2	56	7.5	1.4	90.1
Intent	$\geq$3	54	2.5	0.0	92.4

Table S12: Route results by reference length, pooled over nine VLMs (%). N counts questions before model expansion.

D HUMAN–MODEL COMPARISONS AND QUALITATIVE EXAMPLES

D.1 SAME-QUESTION HUMAN–MODEL COMPARISON

We compare human participants with nine models on a randomly sampled subset comprising 10 questions from each of the five tasks, for 50 questions in total. The human reference exceeds the strongest VLM on the same questions by 25.6 points in EgoPath Score. Across the three route tasks, the human answers outperform the strongest VLM results by 43.3 percentage points in mean valid-path rate and 33.4 percentage points in mean success rate. This exploratory comparison provides a same-interface reference rather than an estimate of a population-level human ceiling.

Evaluator	Score	Valid path
Human calibration	**54.2**	**70.0**
GPT-5.5	28.6	26.7
Claude Opus 4.8	26.6	23.3
Gemini 3.1 Pro	23.2	26.7
Llama 4	18.9	20.0
MiniMax M3	17.6	16.7
Qwen 3.6	15.6	20.0
Mistral L3	11.3	16.7
Kimi K2.6	9.7	13.3
Grok 4.3	0.8	10.0

Table S13: Same-question calibration on a fixed 50-question subset (10 per task; %). Score combines chance-adjusted traversability skill and route success; valid path is averaged over route tasks.

D.2 QUALITATIVE EXAMPLES

Figures S1–S10 present two real questions from each EgoPathBench task. Every panel within a figure repeats the identical first-person image and visual candidate interface, then overlays the recorded answer from the human calibration or one of the nine evaluated VLMs. The overlays are generated from the returned waypoint IDs and the released evaluator records; they were not visible to respondents.

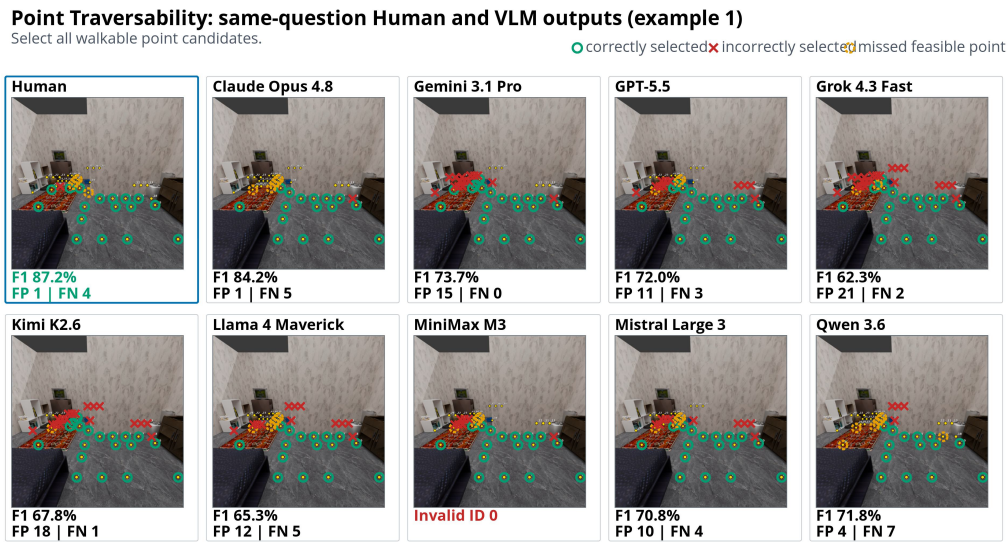

Figure S1: Same-question outputs for Point Traversability. Each panel repeats the identical first-person input and overlays the actual Human or model response. Green rings denote correctly selected feasible points, red crosses denote incorrectly selected points, and dashed orange rings denote feasible points omitted from the returned set.

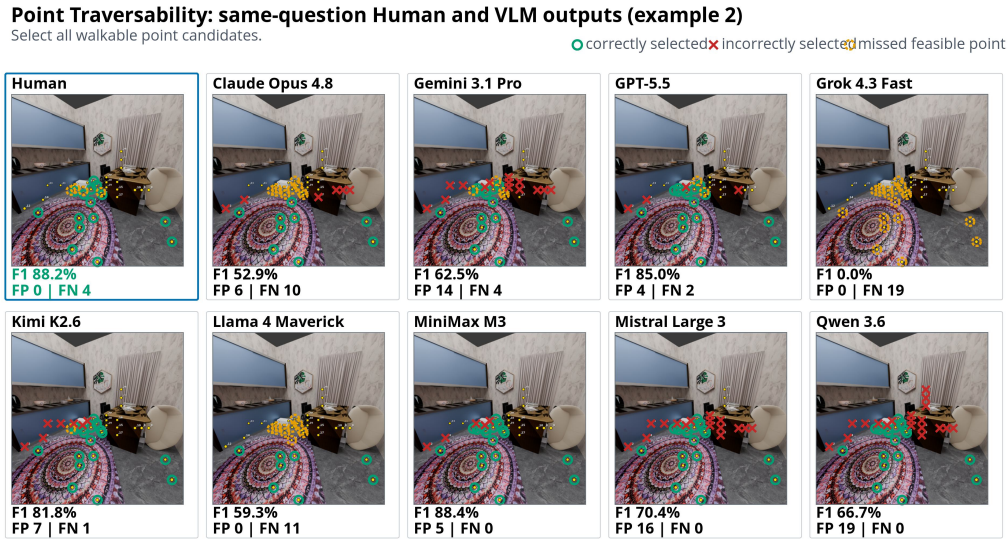

Figure S2: Same-question outputs for Point Traversability. Each panel repeats the identical first-person input and overlays the actual Human or model response. Green rings denote correctly selected feasible points, red crosses denote incorrectly selected points, and dashed orange rings denote feasible points omitted from the returned set.

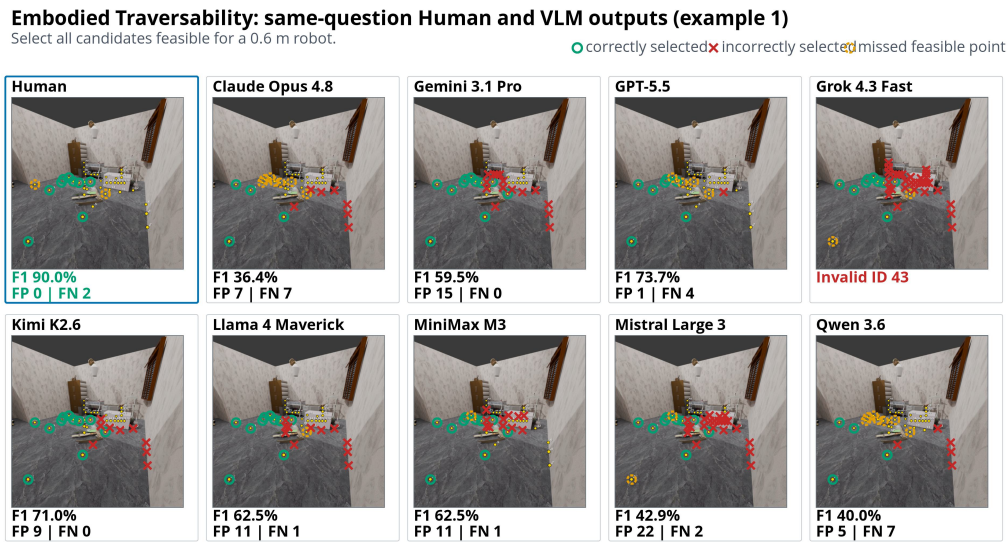

Figure S3: Same-question outputs for Embodied Traversability. Each panel repeats the identical first-person input and overlays the actual Human or model response. Green rings denote correctly selected feasible points, red crosses denote incorrectly selected points, and dashed orange rings denote feasible points omitted from the returned set.

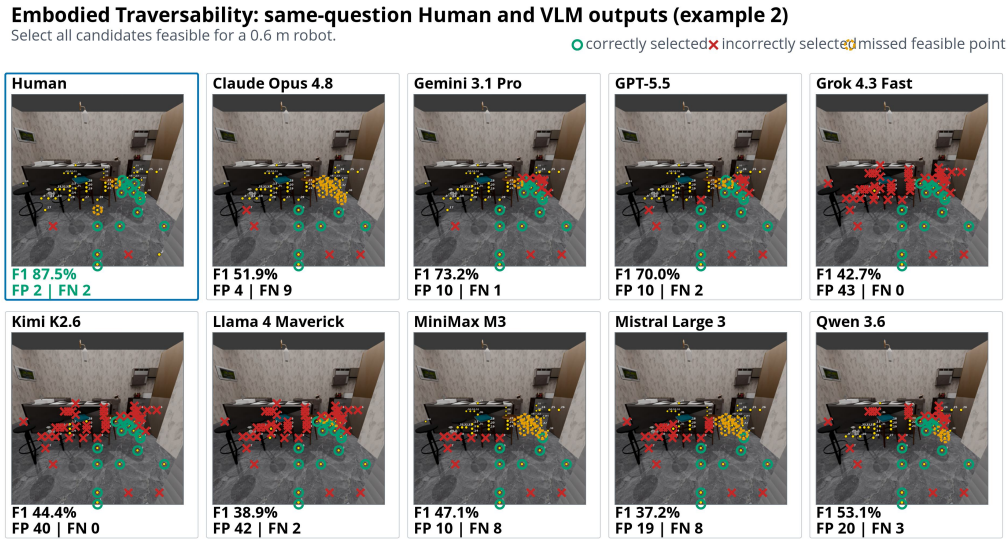

Figure S4: Same-question outputs for Embodied Traversability. Each panel repeats the identical first-person input and overlays the actual Human or model response. Green rings denote correctly selected feasible points, red crosses denote incorrectly selected points, and dashed orange rings denote feasible points omitted from the returned set.

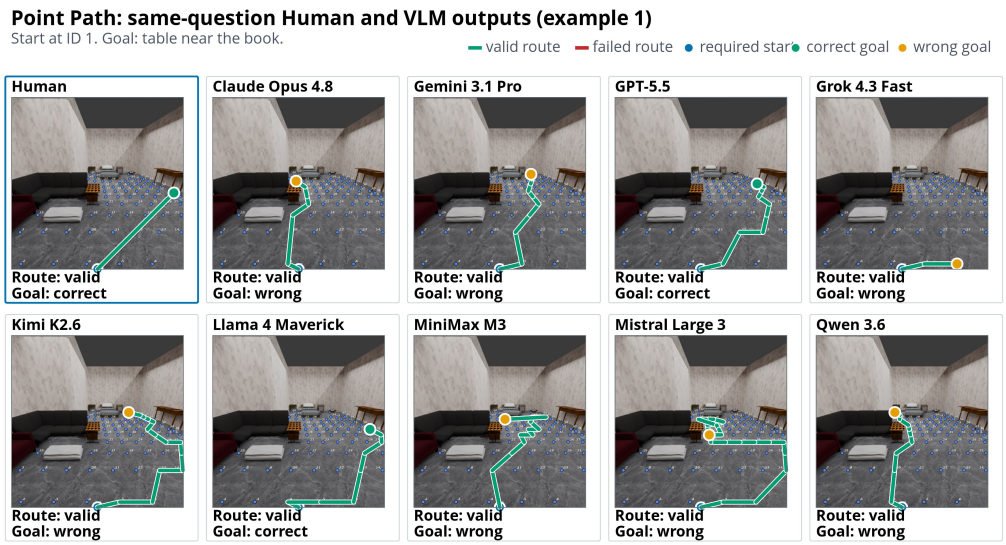

Figure S5: Same-question outputs for Point Path. Each panel repeats the identical first-person input and overlays the actual Human or model response. The complete returned route is green when all consecutive edges are legal and red otherwise. The required start is blue; the returned endpoint is green for an acceptable goal and orange for a wrong goal.

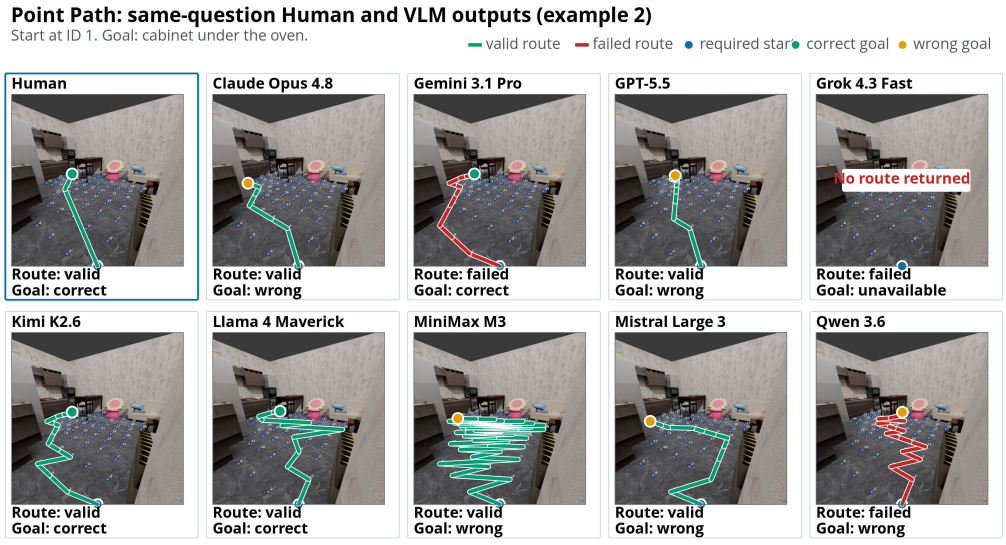

Figure S6: Same-question outputs for Point Path. Each panel repeats the identical first-person input and overlays the actual Human or model response. The complete returned route is green when all consecutive edges are legal and red otherwise. The required start is blue; the returned endpoint is green for an acceptable goal and orange for a wrong goal.

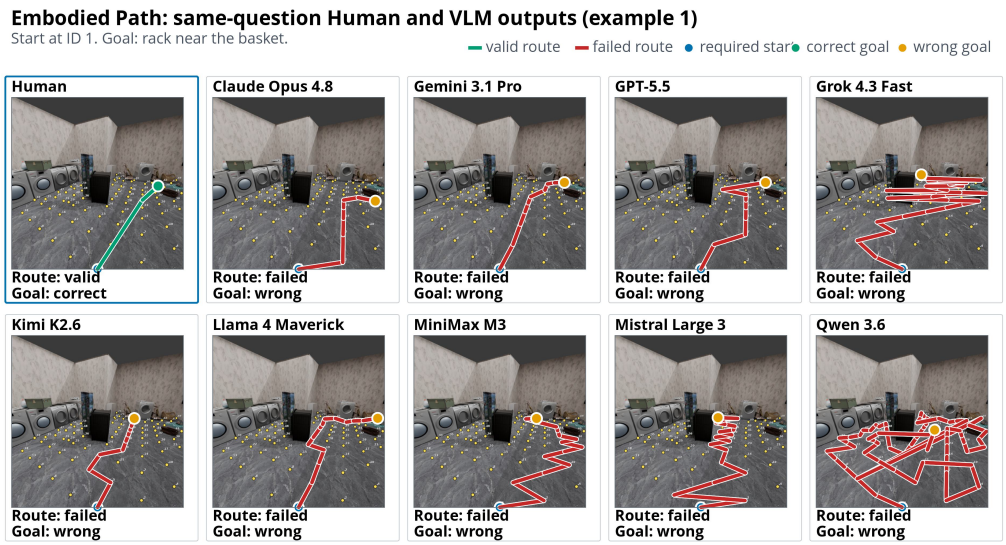

Figure S7: Same-question outputs for Embodied Path. Each panel repeats the identical first-person input and overlays the actual Human or model response. The complete returned route is green when all consecutive edges are legal and red otherwise. The required start is blue; the returned endpoint is green for an acceptable goal and orange for a wrong goal.

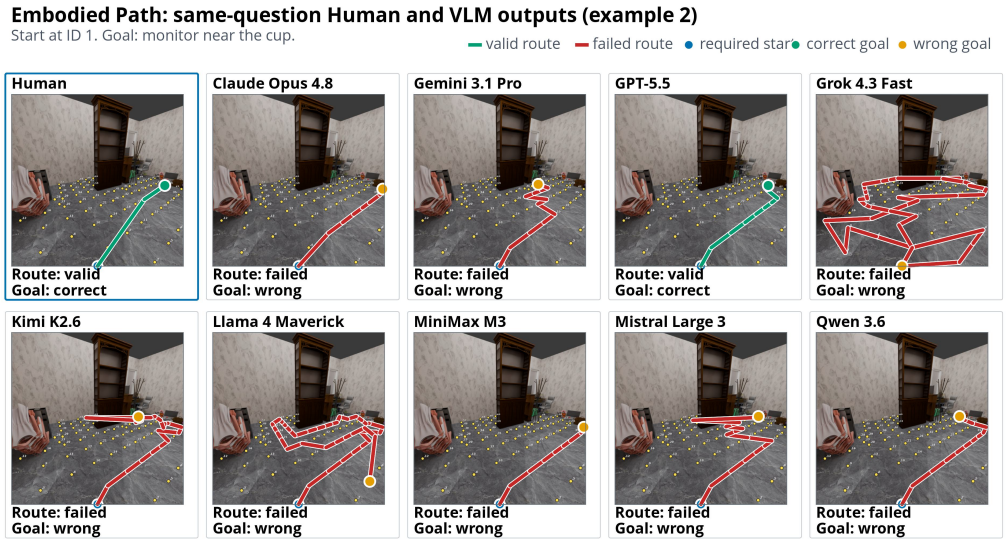

Figure S8: Same-question outputs for Embodied Path. Each panel repeats the identical first-person input and overlays the actual Human or model response. The complete returned route is green when all consecutive edges are legal and red otherwise. The required start is blue; the returned endpoint is green for an acceptable goal and orange for a wrong goal.

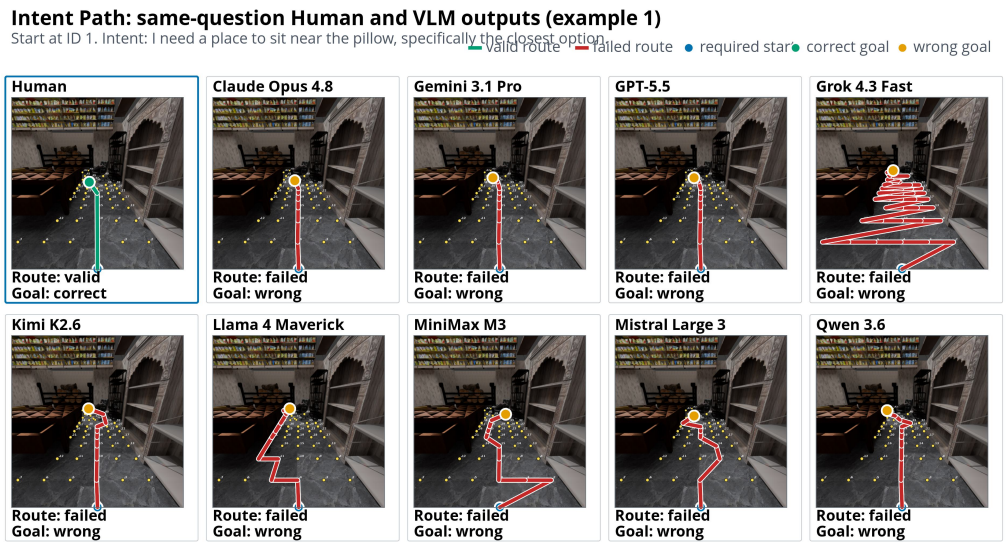

Figure S9: Same-question outputs for Intent Path. Each panel repeats the identical first-person input and overlays the actual Human or model response. The complete returned route is green when all consecutive edges are legal and red otherwise. The required start is blue; the returned endpoint is green for an acceptable goal and orange for a wrong goal.

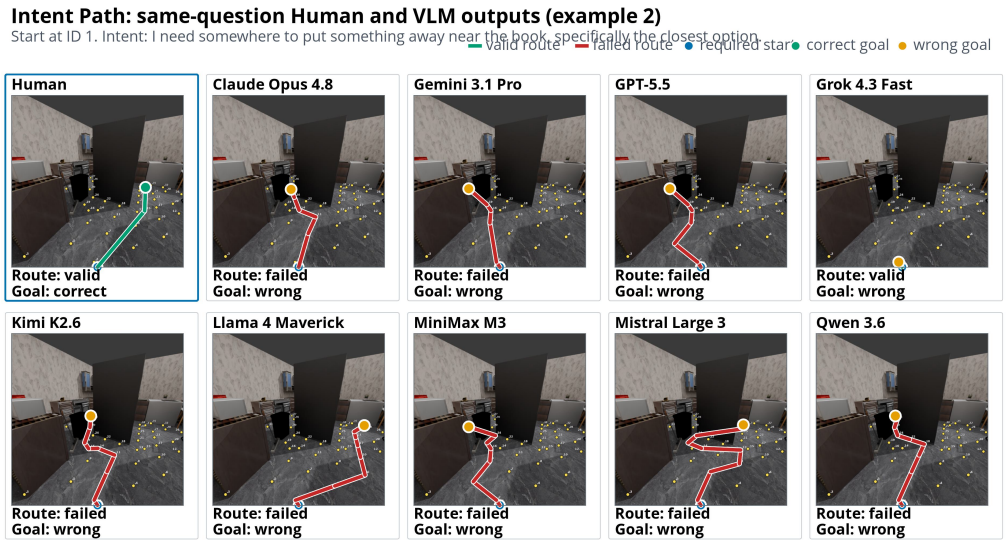

Figure S10: Same-question outputs for Intent Path. Each panel repeats the identical first-person input and overlays the actual Human or model response. The complete returned route is green when all consecutive edges are legal and red otherwise. The required start is blue; the returned endpoint is green for an acceptable goal and orange for a wrong goal.