GeoGuesser Bench: Benchmarking Vision–Language Agents on Interactive Geolocation

Anonymous authors
Paper under double-blind review

Abstract

Vision–language model (VLM) agents that act in the open world need two abilities: world knowledge to interpret what they see, and exploration to find what they cannot yet see. We study how agents combine the two through interactive visual geolocation. GeoGuesser Bench contains 282 tasks in an offline Street View environment where agents rotate, zoom, and move between real panoramas. Its three task types shift the balance from knowledge to exploration: agents (1) report the coordinates of their starting location, (2) name the enclosing street, city, state, or country, or (3) find, read, and count nearby objects. The coordinate tasks in (1) also include 42 real GeoGuessr World Championship rounds, whose expert guesses provide a human reference. We evaluate 23 VLMs: the best accuracy is 62.5% on place naming and 35.0% on object finding, and the best agent averages 4352.4 GeoGuessr points on the championship rounds, against 4770.3 for the better expert in each round. Beyond these headline numbers, world knowledge and exploration are only weakly correlated across models: agents that localize well do not necessarily explore well. To understand where agents fail, we categorize their errors into vision, navigation, and tool-use failures. Further ablations show that the strongest agents, such as Claude Opus 4.8, still perform well with limited views, while weaker agents degrade; customizing the environment thus adjusts task difficulty. The environment's verifiable rewards and restorable state also support training, and our Planner–Worker–Compressor harness cuts planner tokens by 57–61% by compacting the interaction history. We hope GeoGuesser Bench and its environment serve as a high-quality foundation for simulating real-world exploration at scale.

1 Introduction

AI agents are moving from the screen into the physical world. Success in the physical world requires more than recognizing what is already visible. Agents need two abilities that are easy to conflate: *knowing*, the world knowledge needed to interpret an observation, and *looking*, the exploration needed to gather missing evidence by deciding where to look, how to move, and when enough has been seen. A model may recognize an architectural style, a road marking, or a script from a single image, yet fail when the decisive clue lies around a corner. Building such agents requires interactive testbeds that are grounded in the real world yet run digitally, at scale, and reproducibly, since physical deployment is too slow, costly, and uncontrolled to serve as a benchmark.

Current evaluations rarely exercise both abilities together. Vision–language models (VLMs) increasingly act through tools (Yang et al., 2023; Liu et al., 2023; Surís et al., 2023), but static visual benchmarks pose questions about a fixed image, and agent benchmarks mostly target function calling and digital workflows (Patil et al., 2025; Trivedi et al., 2024) or configurable puzzle, navigation, and manipulation environments (Wang et al., 2026). It thus remains unclear whether models with strong visual world knowledge can also explore a real environment for evidence that is not initially visible.

Visual geolocation from Street View is a natural testbed for this question. Street View meets these requirements: its imagery carries the diversity and ambiguity of the real world, yet an agent can rotate, zoom, and move between panoramas without deploying hardware. Success can hinge on knowledge, such as recognizing architecture, signage, or vegetation, or on exploration, such as walking to a street sign or storefront. The game GeoGuessr has turned the task into a competitive sport (Haas

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.

et al., 2024; Plonk It, n.d.) whose top players offer a strong human reference. Most geolocation benchmarks, however, score a single prediction from a fixed image (Haas et al., 2024), and interactive geolocation work (Zheng et al., 2026) remains centered on localization: it does not ask agents to seek out other information in the scene or compare them with expert players.

We introduce GeoGuesser Bench, 282 tasks in an interactive Street View environment in which agents explore through tools and return answers that are verified automatically. Three task types shift the balance from knowing to looking. Precise Localization asks for the coordinates of the starting location and rewards world knowledge. Place Naming asks for the enclosing street, city, state, or country, whose clues may require turning or moving to bring into view. Navigation Challenge asks agents to find, read, or count objects that are not visible from the start and must be actively discovered. The coordinate tasks include 42 World Championship rounds, with the finalists' guesses as a human reference.

We evaluate 23 VLMs, and four findings stand out. First, knowledge-heavy and exploration-heavy tasks rank models differently: GPT-5.4 ranks third on one coordinate split but ties for 18th on object finding. Second, even the best agent still falls short of human experts on the championship rounds, averaging 4352.4 points against 4770.3 for the experts. Third, stronger agents hold up when the environment gets harder, while weaker agents break down: when each view is cut to its top or bottom half, Claude Opus 4.8 retains 98% of its score, whereas Gemini 3.1 Pro retains only 33%. Finally, a failure case study of four models shows that navigational reasoning, rather than visual recognition or tool use, is the most common cause of failure, appearing in 60–80% of failed episodes on every task.

The environment is also designed for training. It runs fully offline as a self-contained package, which makes it easy to set up and use, keeps observations from drifting as Street View imagery is updated, and closes shortcuts to reward hacking, since agents cannot look answers up. Its rewards are verifiable, and its full state (panorama and camera pose) can be restored at any step, which makes branching rollouts from intermediate states (Yi et al., 2026) inexpensive. Its difficulty is also configurable: restricting the agent's views or tools (e.g., removing movement) changes what evidence it can gather, and restricted views cost weaker agents far more than the strongest, which can support a training curriculum.

Long episodes are also costly: every step appends an image and text that the model rereads at every later step, so the input grows quickly with episode length. We therefore introduce a simple context-management harness, Planner–Worker–Compressor, that records the exploration state in a short summary before each planning call while keeping selected key frames. It prunes planner tokens by 57–61% while largely preserving accuracy.

Contributions.

- **Benchmark.** GeoGuesser Bench: 282 real-world Street View tasks that range from knowledge-heavy localization to exploration-heavy information seeking, with an expert reference from GeoGuessr World Championship rounds (Sections 3 and 4).
- **Environment.** An offline, configurable environment with verifiable rewards and restorable state for evaluating and training agents, and a Planner–Worker–Compressor harness that reduces planner tokens by 57–61% (Section 6).
- **Findings.** An evaluation of 23 VLMs, with ablations and a failure analysis, showing that knowledge and exploration rank models differently, that agents still trail human experts, that stronger agents are more robust to restricted views, and that navigational reasoning is the main failure mode (Section 5).

2 Related Work

Visual geolocation. Most geolocation methods predict a location from a single image, whether by classifying over geographic cells (Weyand et al., 2016; Theiner et al., 2021), fusing image and segmentation features (Pramanick et al., 2022), aligning images with GPS coordinates contrastively (Cepeda et al., 2023), combining geographic priors with supervised training (Haas et al., 2024), augmenting multimodal LLMs with retrieval (Zhou et al., 2024; Jia et al., 2024), fusing semantics with hierarchical geographic embeddings (Daruna et al., 2026), or re-ranking candidates by distance (Jia et al., 2025). Agent-based methods add retrieval and OCR tools but still reason over one fixed view (Li et al.,



Figure 1: **Example GeoGuesser Bench episodes.** Read left to right; calls below each new view specify the transition. Boxes mark natural (green) and human-made (red) clues. From top: Precise Localization, Place Naming, and Navigation Challenge. Prompts are abbreviated or illustrative.

2026). WanderBench (Zheng et al., 2026) makes geolocation interactive by organizing panoramas into navigable graphs. GeoGuesser Bench also asks agents to act, but additionally tests information seeking beyond localization (Navigation Challenge) and provides an expert-player reference.

Interactive and embodied benchmarks. StreetLearn (Mirowski et al., 2018) and Touchdown (Chen et al., 2019) evaluate goal-driven and instruction-following navigation in Street View, where success depends on reaching a place rather than on recognizing where one is. VisGym (Wang et al., 2026) evaluates VLMs across configurable puzzle, navigation, and manipulation environments, and Agent-X (Ashraf et al., 2025) evaluates multi-step visual reasoning with tools. Tool-use benchmarks (Patil et al., 2025; Trivedi et al., 2024; Ning et al., 2024) test function calling in text or digital environments, and vision–language–action models target physical control (Ma et al., 2024). GeoGuesser Bench complements these with global real-world imagery and answers that require both world knowledge and exploration.

Agent frameworks and context management. MM-REACT (Yang et al., 2023), LLaVA-Plus (Liu et al., 2023), and ViperGPT (Surís et al., 2023) pair language or vision–language controllers with vision tools, and ReAct (Yao et al., 2023) interleaves reasoning with actions in a closed loop. Toolformer (Schick et al., 2023) and Gorilla (Patil et al., 2023) train models to call APIs; see Qin et al. (2023) for a survey of tool learning. Long interaction histories remain a bottleneck: VisGym finds that VLMs perform worse with unbounded histories than with truncated ones (Wang et al., 2026). Our Planner–Worker–Compressor harness (Section 6) addresses this by summarizing the textual history while keeping selected visual observations.

3 Benchmark Construction

GeoGuesser Bench contains 282 entries, each pairing a Street View starting location with a question and a reference answer (Figure 1). All entries are solved in the same interactive environment (Section 4), but they differ in how much they draw on world knowledge versus exploration: by design, Precise Localization depends mainly on knowledge, Navigation Challenge mainly on exploration, and Place Naming on both. Section 5 tests whether agents actually solve them this way.

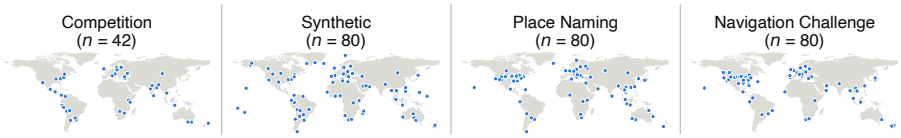


Figure 2: **Starting locations of each split.** Competition and Synthetic are the two Precise Localization splits; Place Naming and Navigation Challenge favor urban and suburban areas (Section 3.2). Empty regions largely reflect scarce or absent Street View imagery, as in mainland China and much of Russia and Africa.

3.1 Task Types

Precise Localization (80 + 42 entries). The agent reports the latitude and longitude of its starting location, as in GeoGuessr. Precise answers usually require combining many individually ambiguous cues, both natural (vegetation, soil, terrain) and human-made (signage, road markings, architecture). The task has two splits: the *Synthetic* split, 80 real locations that we selected ourselves, and the *Competition* split, 42 rounds of professional GeoGuessr play with recorded expert guesses (Section 3.2).

Place Naming (80 entries). The agent names the region containing the starting location at the granularity the question specifies: street (11 entries), city or town (31), state or province (11), or country (27). Each location has regional indicators nearby, such as street signs, the language of written text, or architectural style, but the agent may need to turn or move to bring them into view. The task thus combines cultural knowledge with moderate navigation.

Navigation Challenge (80 entries). The agent answers a question about a specific object or feature near the start. Question types include Identification, which asks for a property of a target, such as a business name, the text on a sign, or a color, and Enumeration, which asks how many instances of an object lie in an area or along a route. Many prompts give relative directions (e.g., "head towards the stoplight"). No target is visible in the initial view, so every question requires changing the viewpoint, and many require moving to other panoramas. The task therefore demands the most precise navigation and, on average, the least world knowledge.

3.2 Curation

Locations. We selected the Synthetic locations manually to span diverse countries, climates, and settings, from city centers to rural roads. Place Naming and Navigation Challenge locations favor urban and suburban areas, where cultural signals are strong and objects are distinct (Figure 2). For the Synthetic split and Place Naming, whose answers are properties of the location itself, we excluded locations whose initial view shows a famous landmark that a model could recognize without exploring. Together with the offline environment (Section 4), this filtering is designed so that answers must come from the scene rather than from recalling a landmark or searching externally.

Questions and accepted answers. For each Place Naming location, we chose a target region and identified visual clues near the start that point to it; where necessary, we changed the granularity (e.g., from city to street) or moved the start within the region so that the clues fit the target. For each Navigation Challenge location, we chose a target that exploration can reliably reach and wrote a question whose answer is a verifiable fact in the scene; we confirmed that the target is not visible in the initial view and that the question can be answered within the 40-step budget. Each Place Naming and Navigation Challenge question specifies an answer format (e.g., "Give only the number as an integer") and lists accepted answers that cover equivalent forms, such as "Ave" and "Avenue" or "Czechia" and "Czech Republic." For Precise Localization, the answer is the starting coordinate.

Competition split. The 42 Competition locations are rounds from the Grand Finals of the GeoGuessr World Cup in 2023 and 2024 (renamed the GeoGuessr World Championship in 2025). For each round, we record the location together with the distance and score of the better of the two players' guesses, which form the human reference (Section 4). Because these rounds were broadcast publicly, models may have seen them during pretraining; the Synthetic split, which does not come from public competition, serves as a check against such memorization (Section 5.1).

4 EVALUATION SETUP

Environment. The environment exposes Google Street View panoramas to the agent through tools (Figure 1). An episode starts at the entry's starting location with the camera facing north, level, at the default zoom (a 90° horizontal field of view). The prompt contains no image: the agent sees the scene only through tools that return a 1920×1080 rendering of the current view. Renderings omit Street View's interface overlays, including the address box, road labels, and navigation arrows, so the interface never reveals the location. The agent has the following tools: `scroll_*` tools rotate (left, right) or tilt (up, down) the camera, and `zoom_in` and `zoom_out` change magnification, each by an amount the agent chooses; and `move_*` tools (`move_north`, `move_northeast`, ...) step to the adjacent panorama in a compass direction, if one exists. Camera tools return the current view.

Offline package. The environment is a self-contained package that runs fully offline. The panoramas and panorama links reachable from each entry are bundled with the benchmark, so episodes make no network requests and observations do not drift as Street View imagery is updated. Running offline also closes shortcuts to reward hacking: agents cannot query map services, reverse-geocode their position, or search the web, and tools return only rendered views or movement status, never coordinates, addresses, or reference answers. The task set, tool set, step budget, and observation settings are configurable, and only the web-search ablation (Section 5.3) enables internet access. We will release the package with the benchmark.

Episodes and models. A step is one model response, which may contain several tool calls. An episode ends when the agent replies without calling a tool, and that reply is its answer; the agent may take at most 40 tool-calling steps. The system prompt describes how answers are scored, asks for a final reply with an `answer` field and a brief `context` justification, and permits "I do not know" as an answer. An agent that uses all 40 steps receives one final turn in which it is asked for its answer; we say that an episode *ends at the step limit* only if the agent still gives no answer in this turn. We evaluate the 23 VLMs in Table 1 through their function-calling interfaces with identical prompts, tools, and budget. Every entry counts toward its task's result: an episode that ends at the step limit, ends in an error, or yields no parseable answer scores 0 on Precise Localization and counts as incorrect on Place Naming and Navigation Challenge.

Scoring. For Precise Localization, we parse the first valid latitude–longitude pair in the final reply and compute its geodesic distance d (km) to the ground truth on the WGS-84 ellipsoid. GeoGuessr does not publish its scoring rule, but prior work and player analyses report an exponential decay (Haas et al., 2024; Dede, n.d.), $S(d) = 5000 \exp(-10d/D)$, where D is the haversine length of the diagonal of the smallest latitude–longitude rectangle containing the map ($D = 14{,}916.862$ km for the world map; Appendix B). The score ranges from 0 to 5000. For Place Naming and Navigation Challenge, we report accuracy over all 80 entries of each task; a reply is correct only if its `answer` field strictly matches one of the accepted answers, which include equivalent aliases (Appendix C).

Human reference. On the Competition split, the better player's guess in each round averages 4770.3 points, a score-equivalent distance (Appendix B) of 70.2 km. This is a reference point, not a matched comparison: players had one minute per round and placed their guesses on an interactive map, whereas agents have up to 40 steps but must state coordinates without a map. Taking the better of two guesses per round also makes the reference optimistic relative to a single expert.

5 RESULTS

5.1 AGENT PERFORMANCE

Precise Localization. No model reaches the human baseline on the Competition split (Table 1). The best agent, Claude Opus 4.8, averages 4352.4 points (207 km) against the expert reference of 4770.3 (70 km), and the runner-up trails it by only 46 points, well within their standard errors of about 200. The two Precise Localization splits rank the 23 models almost identically (Spearman $\rho = 0.97$), and 19 models score higher on Synthetic than on Competition, so we see no sign that exposure to broadcast competition rounds inflates Competition scores.

Much of the spread between models reflects whether an agent answers at all rather than how accurately it localizes. Pooled over models, 45.7% of Competition and 38.6% of Synthetic episodes score zero

Table 1: **Results on GeoGuesser Bench.** For the Competition and Synthetic splits of Precise Localization (Section 3.1), cells show the mean GeoGuessr score over all entries, with invalid responses scored 0, and the score-equivalent distance in km (Eq. 1). The human row averages the better of the two players' guesses in each Competition round (Section 4). Models are sorted by average rank, the mean of a model's ranks in the four result columns (1 is best; ties share the mean rank). **Bold** marks the best model in each column.

	Precise Localization (Score | Dist)		**Place Naming**(%)	**Navigation Challenge**(%)
Model	**Competition**	**Synthetic**		
Human experts	4770.3 | 70.2	–	–	–
Claude Opus 4.8	**4352.4** ±128 | **206.9**	**4549.5** ±83 | **140.8**	60.00 ±3.6	26.25 ±3.3
Claude Opus 4.7	4306.4 ±146 | 222.8	4501.4 ±96 | 156.7	56.25 ±3.6	31.25 ±3.4
GPT-5.5	3815.1 ±182 | 403.5	3467.8 ±161 | 545.8	**62.50** ±3.5	32.50 ±3.4
Claude Fable 5	2821.1 ±248 | 853.7	3164.2 ±175 | 682.5	**62.50** ±3.5	**35.00** ±3.5
Qwen3.7-Plus	3964.4 ±133 | 346.2	3996.0 ±104 | 334.4	42.50 ±3.6	23.75 ±3.1
GPT-5.4 Pro	2503.8 ±244 | 1031.7	2667.4 ±176 | 937.3	60.00 ±3.6	**35.00** ±3.5
GPT-5.4	3923.1 ±122 | 361.8	4221.1 ±85 | 252.6	50.00 ±3.6	12.50 ±2.4
GLM-5V-Turbo	3138.2 ±200 | 694.8	3859.8 ±131 | 386.1	46.25 ±3.6	15.00 ±2.6
Claude Opus 4.6	3023.9 ±231 | 750.2	3363.5 ±165 | 591.4	45.00 ±3.6	18.75 ±2.9
MiniMax-M3	3061.5 ±205 | 731.7	2905.8 ±159 | 809.6	33.75 ±3.4	17.50 ±2.8
Qwen3.6-Plus	3045.5 ±173 | 739.5	2956.4 ±141 | 783.8	36.25 ±3.5	15.00 ±2.6
GPT-5.5 Pro	1163.4 ±211 | 2175.0	1556.1 ±168 | 1741.2	47.50 ±3.6	31.25 ±3.4
Gemini 3 Flash	2851.0 ±240 | 838.0	2493.3 ±183 | 1038.0	38.75 ±3.6	18.75 ±2.9
Claude Sonnet 4.6	2649.1 ±245 | 947.5	2883.8 ±174 | 820.9	47.50 ±3.6	12.50 ±2.4
Gemini 3.5 Flash	1698.3 ±233 | 1610.7	2184.0 ±180 | 1235.5	46.25 ±3.6	17.50 ±2.8
Gemini 3.1 Pro	593.6 ±164 | 3178.8	1124.8 ±153 | 2225.3	47.50 ±3.6	30.00 ±3.4
Kimi K2.5	2016.2 ±228 | 1354.8	2717.1 ±174 | 909.7	42.50 ±3.6	15.00 ±2.6
Kimi K2.6	1667.9 ±229 | 1637.7	2100.3 ±177 | 1293.8	50.00 ±3.6	13.75 ±2.5
Qwen3.5-397B-A17B	2221.5 ±209 | 1210.1	2467.0 ±160 | 1053.8	28.75 ±3.3	8.75 ±2.1
GPT-5.2	523.9 ±146 | 3365.1	1243.2 ±151 | 2076.1	33.75 ±3.4	17.50 ±2.8
Grok 4.3	805.8 ±173 | 2722.9	1546.8 ±162 | 1750.1	32.50 ±3.4	5.00 ±1.6
Qwen3.5-122B-A10B	145.7 ±78 | 5274.1	807.2 ±126 | 2720.3	12.50 ±2.4	6.25 ±1.8
Qwen3.5-35B-A3B	252.4 ±83 | 4454.4	364.8 ±89 | 3905.0	10.00 ±2.2	3.75 ±1.4

(Table 8), mostly because the agent either ends at the step limit, using all 40 steps without answering even when prompted in the final turn (20.7% and 19.1% of episodes; Section 4), or abstains before the limit (19.0% and 14.6%). Both are failures to commit to an answer rather than inaccurate answers: the system prompt permits "I do not know," but it also describes the scoring rule, under which any guess scores above zero, so declining to guess is never optimal. When agents do answer, they are often accurate (Table 9): on the Synthetic episodes they answer, GPT-5.2 and Grok 4.3, whose zeros are all abstentions, average 4324 and 4419 points, more than GPT-5.4 (4221), which answers all 80. This conditional score is only a diagnostic, since each model answers a different subset of episodes.

The strongest Precise Localization agents also explore little: the four best by mean Precise Localization score average at most 3.1 `move_*` calls per episode on this task, yet three of them move 8.2–9.3 times per Navigation Challenge episode (Figure 3a). Section 5.3 tests whether movement is needed at all.

Place Naming. The best models name the correct region in 62.5% of questions. Accuracy falls with granularity (Figure 3b): the five strongest models identify the country in 81.5–88.9% of cases but the street in only 27.3–45.5%. Running out of steps also costs accuracy: 16.1% of all Place Naming episodes end at the step limit.

Navigation Challenge. This is the hardest task: no model exceeds 35.0%. Agents also work longest here with 19 of the 23 models taking more steps on Navigation Challenge than on either the Synthetic split or Place Naming (Table 5), and 28.5% of Navigation Challenge episodes end at the step limit, accounting for 35.3% of errors.

Rankings across tasks. Rankings on one task only partly predict the others. Across the 23 models, Precise Localization scores correlate weakly with Navigation Challenge accuracy (Spearman $\rho = 0.39$ for both splits) and moderately with Place Naming accuracy ($\rho = 0.48$ and 0.55), whereas Place

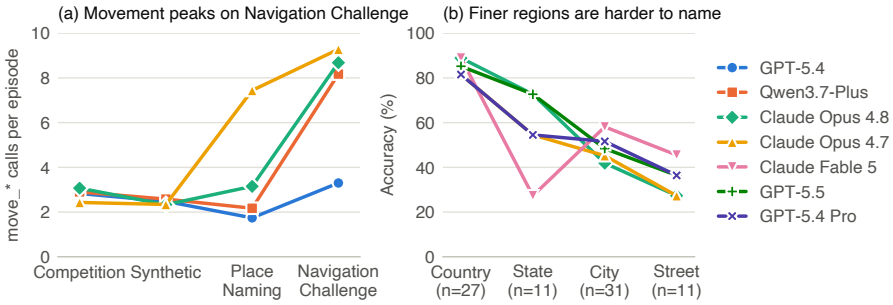


Figure 3: **Exploration and region granularity.** (a) Mean `move_*` calls per completed episode for the four strongest models by mean Precise Localization score: all four move little on Precise Localization, and three move far more on Navigation Challenge. (b) Accuracy of the five strongest Place Naming models by the granularity the question asks for. Claude Fable 5's dip at the state level comes from the step budget: all eight state questions it missed ended at the step limit. Tables 6 and 7 list the values.

Table 2: **Failure modes by task.** Share of failed episodes with each mode, pooled over the four annotated models; an episode can have several modes. **Bold**: most common.

Task	FM1 vision	FM2 navigation	FM3 tool use
Comp.	0.136	**0.780**	0.152
Synth.	0.154	**0.802**	0.123
Place Naming	0.326	**0.665**	0.035
Navigation Challenge	0.351	**0.601**	0.145

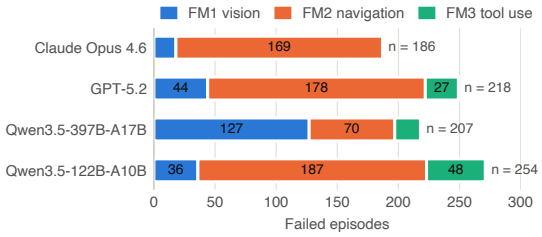


Figure 4: **Failure modes by model.** Failed episodes assigned each mode, pooled over the four task splits; n is the number of failed episodes, and an episode can have several modes.

Naming and Navigation Challenge correlate more strongly ($\rho = 0.70$). With 23 models these estimates are imprecise (a Fisher-z 95% interval for $\rho = 0.39$ spans -0.02 to 0.69), but individual rank reversals are large. GPT-5.4 is third on the Synthetic split but ties for 18th on Navigation Challenge, where it takes only 10.7 steps per episode, consistent with too little exploration; conversely, Claude Fable 5 ties for first on Place Naming and Navigation Challenge but ranks 11th on Competition, partly because it exhausts the budget in 15 of 42 episodes.

5.2 Failure Mode Analysis

To characterize errors, we assigned each failed episode of four models (Claude Opus 4.6, GPT-5.2, and Qwen3.5 at 122B and 397B parameters) one or more of three failure modes (Table 10). A Precise Localization episode counts as failed if it scores worse than 4000 points (333 km), or gave no valid answer; a Place Naming or Navigation Challenge episode fails if its answer does not match the ground truth for that question. We labeled 160 of the 865 failed episodes manually (10 per model and task split) and used them as examples for LLM-based classifier agents, which labeled the remaining 705. **FM1, Vision Failure**, covers errors in interpreting visual input or applying pretrained knowledge, such as hallucinated landmark recognition, misclassified cues, or fixation on distractors (Figures 13 and 14). **FM2, Navigational Reasoning Failure**, covers errors in spatial reasoning and planning, such as inefficient routes, failing to follow the prompt's directions or to replan, and losing track of objects or explored areas (Figures 11 and 12). **FM3, Tool-use Failure**, covers malformed, invalid, or unsupported tool requests, such as misreading a documented tool or calling one that does not exist, which cause API errors, refusals to engage with the task, or corrupted context (Figure 15). Figure 10 in Appendix F shows three of these examples side by side.

Navigational reasoning failure is the most common mode on every task split, appearing in 60–80% of failed episodes (Table 2), and it accounts for most failures of three of the four models (Figure 4).

Table 3: **Ablation results across all conditions.** Mean GeoGuessr score (0–5000) per model on the two Precise Localization splits, Competition (C, $n = 42$) and Synthetic (S, $n = 80$), with invalid responses assigned zero. *80 Steps*: step limit raised from 40 to 80, navigation enabled. *No Movement*: `move_*` tools removed, look-and-zoom only, 40 steps. *Top/Bottom Half*: only the named half of each 1920×1080 capture retained, 40 steps, navigation enabled. *Web Search*: `search_engine_query` and `fetch_url_content` added, full images, 40 steps. Models are sorted by average rank over the ten columns; best value per column in bold. GPT-5.5's Web Search cell on Synthetic averages 79 entries because one question was not attempted (1809 if it is counted as zero). Claude Opus 4.6's Top Half, Bottom Half, and Web Search runs used OpenRouter; its other runs used the provider's API.

Model	80 Steps		No Movement		Top Half		Bottom Half		Web Search	
	C	S	C	S	C	S	C	S	C	S
Claude Opus 4.8	**4594**	**4805**	**4646**	**4556**	**4368**	**4386**	**4369**	**4313**	**4551**	**4380**
GLM-5V-Turbo	3520	4089	4249	4356	3061	3960	3093	3480	3544	4022
Qwen3.7-Plus	3884	3930	3686	3998	2925	3171	3271	3097	3080	3339
Claude Opus 4.6	3708	3741	3749	3018	2527	2583	1928	2879	2688	3069
Claude Sonnet 4.6	2791	3193	3626	4203	2337	2922	2446	2707	3206	2767
GPT-5.5	4407	4391	4100	4178	2212	2305	1403	1553	2388	1832
Kimi K2.5	3038	3167	2668	3119	2258	2413	718	1819	3999	3979
Gemini 3.5 Flash	2744	3228	3939	4019	948	2123	938	1179	0	125
Qwen3.5-397B	2175	2149	1661	1861	1871	2235	1556	1147	2444	2662
Kimi K2.6	1218	1804	781	2161	471	1599	856	1185	238	507
Gemini 3.1 Pro	1729	2473	1939	2532	0	1062	0	437	119	312
GPT-5.2	767	1295	572	1544	604	902	112	667	1696	2209
Grok 4.3	650	1400	578	623	407	1133	213	820	835	1022

Within the Qwen3.5 family, the larger model has fewer tool-use failures (21 versus 48) but more vision failures (127 versus 36), possibly due to longer interactions with the environment (taking more steps on every task; Table 5), which bloats context with redundant visual evidence.

5.3 How Agents Gather Evidence

We probe how agents gather evidence for Precise Localization with five single-factor interventions on 13 models (Table 3): an 80-step limit, no `move_*` tools, only the top or bottom half of each view, and web search. We compare each with the 40-step scores in Table 1.

Step budget and movement. A larger budget and the removal of movement both raise scores for most models, largely because agents run out of steps less often. With 80 steps, 9 of the 13 models improve on each split, and the share of episodes that end at the step limit falls from 19.1% to 7.4% (pooled over models and splits). Three of the four models that do not improve almost never end at the 40-step limit in the main setting.

Removing movement at the original 40-step budget raises scores over the main setting for 9 of 13 models on Competition and 10 on Synthetic, and over the 80-step condition for 6 and 7. Episodes become shorter (12.3 versus 20.0 steps on average) and rarely reach the limit (4.5%, versus 19.1% at 40 steps and 7.4% at 80); Claude Opus 4.8 scores 4646 on Competition without moving at all, more than with movement at either budget. For the strongest agents, a single panorama thus suffices for strong Precise Localization performance, and for most models the main cost of movement is the budget it consumes. The exceptions are two models that tend to abstain, which score lower without movement than with it at either budget and abstain more often.

Image crops. Keeping only the top or bottom half of each view lowers scores for 10 to 13 of the 13 models, depending on the half and split. Stronger models are more robust: averaged over both crops and splits, the fraction of its main-setting score that a model retains (33–98%) rises with its mean main-setting Precise Localization score (Spearman $\rho = 0.73$; Fisher-z 95% interval 0.29 to 0.91), with GPT-5.5 the clearest exception (third in the main setting, it retains 52%). Neither half is consistently more informative. Although the top half beats the bottom half for 12 of 13 models on Synthetic and for only 7 on Competition, every model tilts toward the missing half more often than away from it, which costs steps. The share of episodes ending at the step limit rises from 19.1% to 28.1% (top) and 30.2% (bottom).



Figure 5: **Visual coverage within one panorama.** Bright regions fall within the reconstructed camera views of an example no-movement episode; dark regions were never viewed. The mask shows which directions were viewed, not what the agent attended to.

Web search. Web search improves 7 of 13 models on Competition but only 4 on Synthetic. It helps agents that would otherwise abstain: Kimi K2.5 nearly doubles its Competition score (2016 to 3999), abstaining in 4 rather than 47 episodes across both splits. It hurts agents that spend their budget searching: Gemini 3.5 Flash scores zero on all 42 Competition episodes. External information therefore does not reliably help; agents must still find useful clues and budget their steps.

Visual coverage. Agents rarely inspect the full panorama. We measure coverage, the fraction of the viewing sphere covered by the union of an agent's camera views in a no-movement episode (Figure 5; Appendix D). Across ten models, about two-thirds of episodes with any tool call never change pitch, and most end below the 0.49 of the sphere covered by one full horizontal sweep at the default field of view (Figure 9). High scores do not require broad coverage: Claude Opus 4.8's no-movement run above has a median coverage of 0.31 and never tilts the camera. One explanation is that many geographic cues, such as signage, road markings, vegetation, and facades, lie near the horizon, echoing the equator bias of people viewing 360° panoramas (Sitzmann et al., 2018).

6 Mitigating Context Growth in Long Episodes

Each step appends an image and text that the agent rereads at every later step, so its context grows linearly with episode length and its total input grows quadratically. For Gemini 3 Flash on Navigation Challenge, the median input rises from about 1.1K tokens at the first step to 32.6K at the 40th, and a median episode consumes 649K input tokens. Long histories also cause failures: in the main evaluation, 93 episodes (49 of them on Navigation Challenge) ended because the request exceeded the provider's size limit, 69 for Claude models and 24 for GLM-5V-Turbo. We therefore test whether compressing the history can reduce token usage.

Framework. We split the agent into three models (Figure 6 in Appendix A). A *Planner* (Gemini 3 Flash) decides at each round whether to answer or to keep exploring, in which case it issues a natural-language instruction. A *Worker*, a fine-tuned Qwen2.5-VL-7B, turns the instruction into one atomic action given the current image (move forward, turn by a small or large angle, look up or down, zoom, or stop). A *Compressor* (Gemini 2.5 Flash-Lite) reads the task, the most recent actions, and notes on the key frames stored in the trajectory memory, and writes a short `brief` of the exploration state and a `focus_hint` that points to evidence relevant to the next decision. The Planner receives this summary in place of the raw textual history, together with the task and the key frames, so its textual context stays short while visual evidence remains available.

Token usage. Compared with the previous harness without Compressor, our Planner–Worker–Compressor harness reduces Planner token usage from 11.06M to 4.36M on Precise Localization, from 6.20M to 2.44M on Place Naming, and from 11.18M to 4.80M on Navigation Challenge—reductions of 60.6%, 60.6%, and 57.1%, respectively (Table 4 in Appendix A). These counts cover only the Planner and exclude Worker and Compressor tokens.

7 Conclusion

GeoGuesser Bench tests whether VLM agents can combine world knowledge with active exploration in real Street View scenes. Across 23 agents, knowledge-heavy and exploration-heavy performance diverge, the best agent trails an expert reference, and many failures come from abstaining or exhausting the step budget without committing to an answer, rather than from misreading the scene; finding and counting objects remains hard for every agent. The environment's verifiable rewards and restorable state also make it a training ground for such agents, which we leave to future work.

AI USE STATEMENT

In this work, we used generative AI tools to support data analysis, interpret results, and propose hypotheses for observed trends, including recomputing statistics and checking reported numbers against per-episode result files. We did not use generative AI tools to select locations, write questions or reference answers, label failure modes, or design experiments; generating synthetic datasets, theoretical modeling, mathematical claims, proofs, and translation do not apply to this work. Additionally, we used generative AI tools to draft and edit text, format tables and references, and assemble figures. We have reviewed all AI-assisted work, and the authors checked reported numbers against the underlying result files. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Tajamul Ashraf, Amal Saqib, Hanan Ghani, Muhra AlMahri, Yuhao Li, Noor Ahsan, Umair Nawaz, Jean Lahoud, Hisham Cholakkal, Mubarak Shah, Philip Torr, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Salman Khan. Agent-X: Evaluating deep multimodal reasoning in vision-centric agentic tasks. *arXiv preprint arXiv:2505.24876*, 2025.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization, 2023. URL https://arxiv.org/abs/2309.16020.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. TOUCHDOWN: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Angel Daruna, Nicholas Meegan, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Geosurge: Geo-localization using semantic fusion with hierarchy of geographic embeddings, 2026. URL https://arxiv.org/abs/2510.01448.

Nicolas Dede. The maths of GeoGuessr, n.d. URL https://latb.io/geoguessr/articles/the-maths. Accessed 2026-04-23.

Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12893–12902, 2024.

Pengyue Jia, Yiding Liu, Xiaopeng Li, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, Dawei Yin, and Xiangyu Zhao. G3: An effective and adaptive framework for worldwide geolocalization using large multi-modality models, 2024. URL https://arxiv.org/abs/2405.14702.

Pengyue Jia, Seongheon Park, Song Gao, Xiangyu Zhao, and Sharon Li. Georanker: Distance-aware ranking for worldwide image geolocalization, 2025. URL https://arxiv.org/abs/2505.13731.

Qiujun Li, Zijin Xiao, Xulin Wang, Zhidan Ma, Cheng Yang, and Haifeng Li. Locationagent: A hierarchical agent for image geolocation via decoupling strategy and evidence from parametric knowledge, 2026. URL https://arxiv.org/abs/2601.19155.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. LLaVA-Plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied AI. *arXiv preprint arXiv:2405.14093*, 2024.

Piotr Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. Learning to navigate in cities without a map. In *Advances in Neural Information Processing Systems*, 2018.

Kangyun Ning, Yisong Su, Xueqiang Lv, Yuanzhe Zhang, Jian Liu, Kang Liu, and Jinan Xu. WTU-EVAL: A whether-or-not tool usage evaluation benchmark for large language models. *arXiv preprint arXiv:2407.12823*, 2024.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. *arXiv preprint arXiv:2305.15334*, 2023.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley Function Calling Leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392, 2025.

Plonk It. Beginner's guide to GeoGuessr, n.d. URL https://www.plonkit.net/beginners-guide. Accessed 2026-03-20.

Shraman Pramanick, Ewa M. Nowara, Joshua Gleason, Carlos D. Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild, 2022. URL https://arxiv.org/abs/2204.13861.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2302.04761.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.11366.

Significant Gravitas. Auto-GPT: An autonomous GPT-4 experiment. https://github.com/Significant-Gravitas/AutoGPT, 2023.

Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018. doi: 10.1109/TVCG.2018.2793599.

John P. Snyder. *Map Projections: A Working Manual*. Number 1395 in U.S. Geological Survey Professional Paper. U.S. Government Printing Office, Washington, D.C., 1987. doi: 10.3133/pp1395.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable semantic photo geolocalization. *CoRR*, abs/2104.14995, 2021. URL https://arxiv.org/abs/2104.14995.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2407.18901.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Zirui Wang, Junyi Zhang, Jiaxin Ge, Long Lian, Letian Fu, Lisa Dunlap, Ken Goldberg, XuDong Wang, Ion Stoica, David M. Chan, Sewon Min, and Joseph E. Gonzalez. VisGym: Diverse, customizable, scalable environments for multimodal agents. *arXiv preprint arXiv:2601.16973*, 2026.

Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet - photo geolocation with convolutional neural networks. *CoRR*, abs/1602.05314, 2016. URL http://arxiv.org/abs/1602.05314.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.03629.

Junkeun Yi, Damon Mosk-Aoyama, Baihe Huang, Ritu Gala, Charles Wang, Sugam Dipak Devare, Khushi Bhardwaj, Abhibha Gupta, Oleksii Kuchaiev, Jiantao Jiao, Jian Zhang, and Venkat Srinivasan. PivotRL: High accuracy agentic post-training at low compute cost. *arXiv preprint arXiv:2603.21383*, 2026.

Yushuo Zheng, Huiyu Duan, Zicheng Zhang, Xiaohong Liu, and Xiongkuo Min. Learning to wander: Improving the global image geolocation ability of LMMs via actionable reasoning. *arXiv preprint arXiv:2603.10463*, 2026.

Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, pp. 2749–2754. ACM, July 2024. doi: 10.1145/3626772.3657673. URL http://dx.doi.org/10.1145/3626772.3657673.

A Planner–Worker–Compressor Details

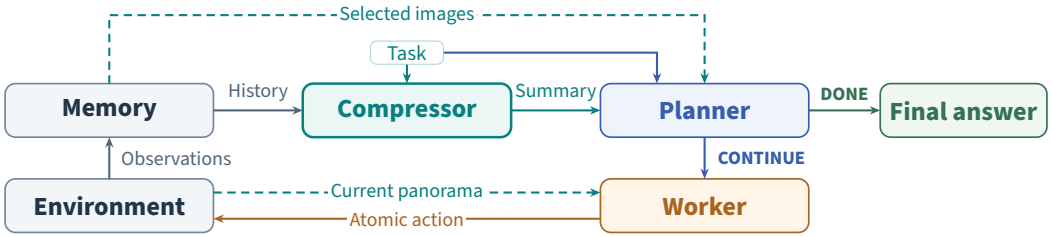


Figure 6: **Planner–Worker–Compressor framework.** The Compressor condenses recent actions and key-frame notes into a `brief` and `focus_hint`; selected images remain available to the Planner. The Planner either returns an answer or instructs the Worker to act, producing new observations for Memory. Dashed arrows denote image inputs to the models.

Table 4: **Planner tokens and task results with and without Compressor.** M denotes millions of tokens; reductions are computed from the displayed counts. Precise Localization reports the GeoGuessr score (the mean distance error falls from 166.89 km to 89.56 km with Compressor); Place Naming and Navigation Challenge report accuracy.

	Precise Localization	Place Naming	Navigation Challenge
Planner tokens, no Compressor	11.06M	6.20M	11.18M
Planner tokens, Compressor	4.36M	2.44M	4.80M
Token reduction	60.6%	60.6%	57.1%
Result, no Compressor	4686	30.00%	24.39%
Result, Compressor	4737	26.25%	24.39%

Configuration. The Planner is `gemini-3-flash-preview` and the Compressor `gemini-2.5-flash-lite`. The Planner's output has four fields, `thought`, `status`, `instruction`, and `answer`: status `CONTINUE` passes the instruction to the Worker, and `DONE` returns the answer. Table 4 reports planner tokens and task results with and without Compressor.

Relation to prior agent frameworks. ReAct (Yao et al., 2023) interleaves chain-of-thought reasoning with tool actions in a closed reason–act–observe loop. Reflexion (Shinn et al., 2023) adds verbal self-criticism to this loop; Auto-GPT (Significant Gravitas, 2023) lets a single LLM decompose a goal and dispatch sub-tasks to itself; and Voyager (Wang et al., 2023) accumulates a library of executable skills through interaction in an embodied setting. These frameworks share the premise that long-horizon problems are best solved by alternating reasoning with observation. The framework in Section 6 follows this premise but departs from prior work in two ways: it splits high-level reasoning and low-level visual action across two models of very different capacity, and it places an explicit Compressor between trajectory memory and the planner, which separates context summarization from high-level reasoning.

B Precise Localization Scoring

The score $S(d)$ (Section 4) halves with every $D \ln 2/10 \approx 1{,}034$ km of error (Figure 7). We average scores rather than distances because the score is defined for invalid responses, which receive 0, and is not dominated by a few intercontinental misses. Tables also report the score-equivalent distance

$$\bar{d} = -\frac{D}{10} \ln \frac{\bar{S}}{5000}, \tag{1}$$

the error whose score equals the mean score $\bar{S}$; it is not the mean error.

Distance computation. GeoGuessr measures guess error with the haversine distance (Dede, n.d.). Given two latitude-longitude coordinates (ϕ_1, λ_1) and (ϕ_2, λ_2) on a spherical surface of radius r, the equation for their haversine distance is given by (Snyder, 1987):

$$\text{Hav}(\mathbf{x_1}, \mathbf{x_2}) = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad (2)$$

This quantity represents the shortest distance between the two coordinates over the surface of the sphere; GeoGuessr uses $r = 6371$ km (Dede, n.d.). Our evaluator instead computes the geodesic distance on the WGS-84 ellipsoid with the `geopy` library. For any pair of points, the two distances differ by less than 0.6%, which changes the score $S(d)$ (Section 4) by at most about 10 points.

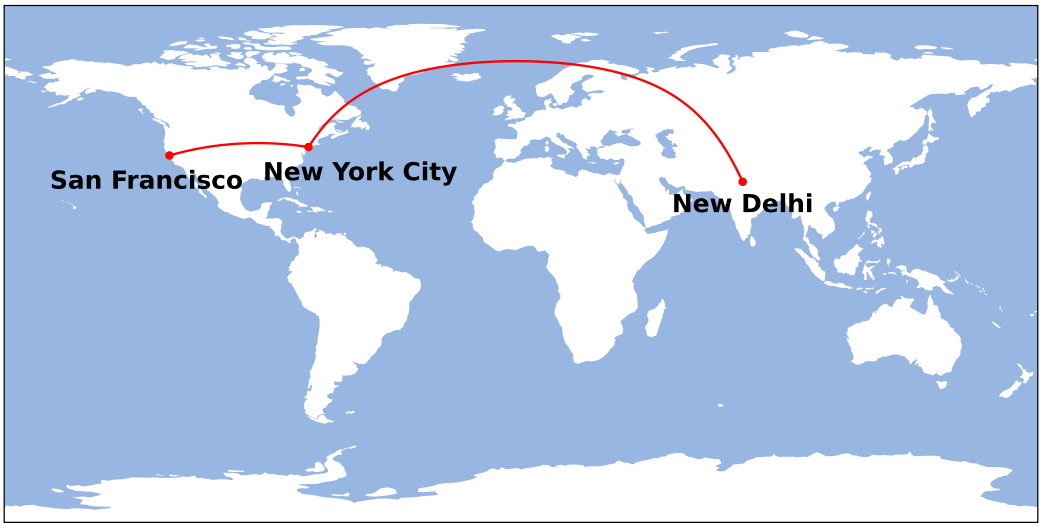


Figure 7: **Haversine Distance Visualization.** Haversine distance from New York City to San Francisco (˜4129 km; score=313) and New Delhi (˜11755 km; score=2).

C Place Naming and Navigation Challenge Scoring

We extract the `answer` field of the final reply, lowercase it and each accepted answer, and strip punctuation; the reply is correct only if the result equals one of the entry's accepted answers. Strict matching gives no credit for hedging: an answer that lists several candidates ("Kenya or Uganda") or negates one ("not Kenya") is incorrect, whereas a looser rule such as whole-word containment would accept both when "Kenya" is accepted. Its risk is the opposite error, a correct answer written in a form that the accepted answers do not list; to limit it, the accepted answers cover equivalent forms of each reference answer, such as "Czechia" and "Czech Republic" (Section 3.2).

Label audit. We audited the automatic labels for both kinds of error. For false positives, we checked replies marked correct for answers that are hedged, negated, or otherwise wrong; for false negatives, we checked replies marked incorrect for answers equivalent to the reference answer. The audit found no mislabeled reply.

D Visual Context Measurement

The no-movement coverage analysis uses ten of the 13 ablation models (Figure 9); Claude Opus 4.6, GPT-5.2, and Qwen3.5-397B are excluded because their camera traces were not available for this analysis. Each model contributes 42 Competition and 80 Synthetic episodes, including zero-score responses.

Coverage definition. For an episode confined to one panorama, angular coverage is the fraction of the viewing sphere covered by the union of reconstructed camera views:

$$C = \frac{1}{4\pi}\int_{\mathbb{S}^2} f(\mathbf{d})\, d\Omega, \qquad f(\mathbf{d}) = \begin{cases} 1, & \text{if } \mathbf{d} \in \bigcup_t F_t, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

where $\mathbf{d}$ is a unit viewing direction and F_t is the spherical footprint of the perspective view after tool call t. Overlapping views count once. Figure 5 illustrates this union in an equirectangular projection, whose areas require latitude weighting to measure spherical coverage. Coverage measures reconstructed view availability, not semantic information or the agent's attention: zooming into a sign may improve legibility without adding angular area.

Camera replay and spherical integration. We reconstruct camera states by replaying each trace from heading, pitch, and horizontal field of view $(0°, 0°, 90°)$. The replay updates heading and pitch with each `scroll_*` call, and divides or multiplies the field of view by the zoom factor for `zoom_in` or `zoom_out`, respectively. Pitch is clamped to $\pm 90°$ and field of view to $[10°, 120°]$. Each recorded tool call emits one reconstructed view, including multiple calls within an agentic step. No extra initial view is inserted. An empty trace is assigned zero coverage, meaning no reconstructed tool-call views, rather than proof that the agent received no initial visual input. This replay assumes that recorded camera commands took effect; it does not independently validate tool execution or image delivery.

For each view, a unit direction belongs to its footprint if it lies in front of the camera and its horizontal and vertical perspective coordinates fall within the field-of-view bounds. We form the Boolean union of these footprints on a 2048×1024 longitude–latitude grid. A cell centered at latitude ϕ contributes $\cos\phi\,\Delta\phi\,\Delta\lambda$ steradians; division by 4π gives Eq. 3. The raster grid controls numerical integration resolution and is distinct from the 1920×1080 rendered viewport. As a geometric check, a single $90°$ horizontal view at this aspect ratio covers approximately 0.1127 of the sphere.

Horizontal-sweep coverage. Figure 8 shows how a single perspective view maps onto the sphere. For the 1920×1080 viewport and default horizontal field of view $h = 90°$, the vertical field of view is $v = 2\arctan[(1080/1920)\tan(h/2)] \approx 58.72°$. A dense horizontal sweep at zero pitch therefore covers a belt with fraction $\sin(v/2) \approx 0.4903$ of the sphere. This corresponds to the vertical cutoff in coverage values pooled across models and datasets in Figure 9. The 0.4903 bound applies to horizon-centered views with horizontal field of view at most $90°$; either changing pitch or zooming out beyond $90°$ can exceed it.

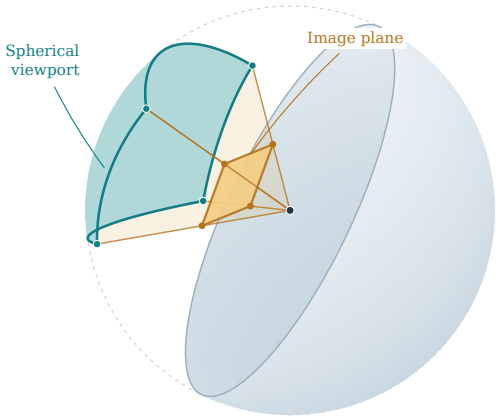


Figure 8: **Camera field of view on the unit sphere.** A perspective camera at the origin projects rays through a rectangular image plane (gold) onto the unit sphere. Normalizing each ray yields the spherical viewport (teal), the footprint F_t of a single view in Eq. 3, whose curved boundary is set by the camera field of view rather than by a latitude–longitude rectangle. The default $h = 90°$ horizontal field of view and 16:9 aspect ratio give $v \approx 58.72°$ vertically. The cutaway removes the forward hemisphere to expose the observer and viewing pyramid; the dashed silhouette marks the full sphere.

Coverage and performance. Figure 9 plots angular coverage against GeoGuessr score, pooling observations across ten models and the Competition and Synthetic subsets. Among traces with at least one tool call, 254/380 (66.8%) on Competition and 496/738 (67.2%) on Synthetic keep pitch

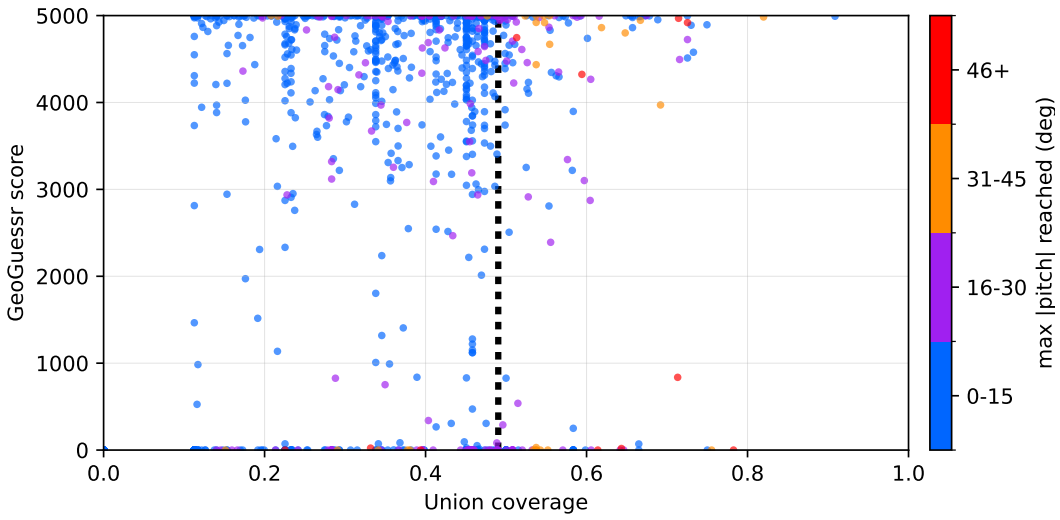


Figure 9: **Panorama coverage versus performance–No Movement, Precise Localization.** Observations from ten models on Competition and Synthetic are pooled in one plot of panorama coverage (Eq. 3) against GeoGuessr score; color denotes maximum absolute pitch. Invalid responses remain at zero score, and zero coverage denotes an empty reconstructed trace. The dotted line at approximately 0.49 indicates the coverage limit of a full horizon sweep at the default field of view; pitch changes or wider fields of view can exceed it.

at zero throughout. The remaining 40 and 62 entries, respectively, have no recorded tool calls; they remain in the plot at $C = 0$ but are excluded from these behavioral percentages. This concentration at the horizon limits angular coverage even when agents turn through all headings. Color denotes the maximum absolute pitch reached, with $0°$–$15°$ grouped in the same bin; exact zero-pitch counts are computed from the traces, not inferred from color. Discrete rotation choices produce repeated coverage values, and the plot is descriptive rather than a controlled dose–response experiment.

High scores do not require broad angular coverage. For example, Opus 4.8 never changes pitch in its 42 Competition no-movement traces, has a median coverage of 0.3075, and achieves a mean score of 4646. GLM-5V-Turbo has a greater median coverage, 0.4732, but a lower mean score of 4249 on the same split. This comparison illustrates why angular extent alone cannot explain performance: models also differ in which regions they inspect and how they interpret them. It does not establish that narrower coverage improves accuracy. Figure 5 illustrates the distinction between inspecting a panorama and exposing its entire surface.

E Failure-Mode Annotation

Five annotators labeled the 865 failed episodes of the four models in Section 5.2. The episodes were divided among the annotators, and each annotator inspected the traces assigned to them and gave each episode one or more of the three failure modes in Table 10. To measure agreement, 300 episodes were labeled by more than one annotator, and the annotators' labels agreed on 95% of them.

F Additional Figures and Tables

Table 5: **Average steps per episode (mean ± standard deviation).** Number of agent steps (tool-call rounds plus the final answer) the model takes before committing a response, per task split. Statistics are computed over *completed* episodes only; episodes that ended in an API error, such as requests exceeding the provider's size limit (HTTP 413) on long, image-heavy episodes of Claude Fable 5 and Claude Opus 4.6, are excluded. Claude Opus rows use the direct Anthropic endpoint. The step budget is 40.

	Precise Localization		Place Naming	Navigation Challenge
Model	**Competition Steps**	**Synthetic Steps**	**Steps**	**Steps**
Claude Fable 5	29.1 ± 12.5	23.9 ± 15.3	20.0 ± 16.0	31.1 ± 11.1
Claude Opus 4.8	12.8 ± 8.9	9.7 ± 7.9	11.6 ± 11.3	27.0 ± 13.2
Claude Opus 4.7	11.8 ± 8.0	9.8 ± 9.0	14.7 ± 13.4	25.1 ± 12.3
Claude Opus 4.6	24.4 ± 12.5	24.5 ± 14.5	24.8 ± 15.3	33.5 ± 10.7
Claude Sonnet 4.6	29.4 ± 13.1	23.2 ± 13.9	24.1 ± 15.3	34.5 ± 10.3
GPT-5.5 Pro	38.3 ± 6.0	36.0 ± 8.3	24.6 ± 15.6	30.0 ± 11.4
GPT-5.5	30.0 ± 8.8	27.2 ± 11.1	17.4 ± 12.2	24.9 ± 10.2
GPT-5.4 Pro	35.7 ± 7.2	33.2 ± 9.3	23.4 ± 12.2	25.8 ± 10.3
GPT-5.4	10.8 ± 2.7	10.1 ± 3.0	7.2 ± 3.5	10.7 ± 8.1
GPT-5.2	9.9 ± 2.9	9.3 ± 3.1	9.1 ± 5.5	12.6 ± 9.4
Gemini 3.1 Pro	39.5 ± 5.4	37.0 ± 8.8	29.1 ± 14.0	32.8 ± 12.0
Gemini 3.5 Flash	35.2 ± 9.2	33.7 ± 10.8	30.1 ± 13.7	35.7 ± 10.1
Gemini 3 Flash	28.3 ± 13.1	28.7 ± 14.0	31.4 ± 13.2	36.4 ± 8.9
Kimi K2.6	26.3 ± 12.1	22.3 ± 13.2	24.3 ± 13.2	31.9 ± 12.1
Kimi K2.5	20.9 ± 10.5	21.8 ± 10.8	20.6 ± 12.5	32.9 ± 11.1
Grok 4.3	3.3 ± 1.4	3.5 ± 1.3	3.5 ± 1.6	4.1 ± 3.0
GLM-5V-Turbo	22.0 ± 11.0	18.5 ± 9.0	16.4 ± 12.0	22.3 ± 13.3
MiniMax-M3	13.9 ± 11.3	11.6 ± 8.6	7.8 ± 9.7	25.2 ± 14.9
Qwen3.7-Plus	12.6 ± 4.3	11.2 ± 5.2	9.8 ± 8.8	20.3 ± 13.8
Qwen3.6-Plus	4.0 ± 3.6	4.1 ± 3.8	3.6 ± 3.1	16.6 ± 14.4
Qwen3.5-397B-A17B	12.0 ± 7.6	15.4 ± 8.0	8.1 ± 8.2	16.7 ± 12.9
Qwen3.5-122B-A10B	3.5 ± 2.0	3.6 ± 2.0	2.5 ± 1.1	9.6 ± 10.5
Qwen3.5-35B-A3B	4.2 ± 2.1	4.5 ± 2.4	6.2 ± 5.6	8.8 ± 7.6

Table 6: **Average movement calls per episode.** Mean number of `move_*` tool calls per episode for the four strongest models by mean Precise Localization score (Competition and Synthetic), by task type; plotted in Figure 3a. Counts are over completed episodes only.

Model	**Comp.**	**Synth.**	**Place Naming**	**Navigation Challenge**
Claude Opus 4.8	3.07	2.33	3.15	8.68
Claude Opus 4.7	2.43	2.34	7.44	9.28
GPT-5.4	2.83	2.48	1.74	3.30
Qwen3.7-Plus	2.90	2.58	2.17	8.17

Table 7: **Place Naming accuracy by question granularity.** Accuracy (%) of the five strongest models on Place Naming, by the geographic granularity the question asks for; plotted in Figure 3b. **Bold**: best per granularity.

Model	Country (n=27)	State (n=11)	City (n=31)	Street (n=11)
Claude Fable 5	**88.9**	27.3	**58.1**	**45.5**
Claude Opus 4.8	**88.9**	**72.7**	41.9	27.3
Claude Opus 4.7	81.5	54.5	45.2	27.3
GPT-5.5	85.2	**72.7**	48.4	36.4
GPT-5.4 Pro	81.5	54.5	51.6	36.4

Table 8: **Precise Localization score distribution and failure causes.** For each Precise Localization split: the median GeoGuessr score, the number of zero-score episodes (Competition: of 42; Synthetic: of 80), and how many of them ended at the 40-step limit, with no answer in the final turn (Step; Section 4), or with an abstention before the limit (Abst.); the remaining zeros are API errors or episodes without a final reply. The last two columns count the Place Naming and Navigation Challenge episodes (of 80) that ended at the step limit in the same sense. Rows follow Table 1.

	Competition				Synthetic				Step limit	
Model	Median	Zero	Step	Abst.	Median	Zero	Step	Abst.	Place Naming	Navigation Challenge
Claude Opus 4.8	4923	2	2	0	4984	2	2	0	4	27
Claude Opus 4.7	4949	4	1	3	4988	4	2	2	9	17
GPT-5.5	4763	7	7	0	4968	18	18	0	7	10
Claude Fable 5	4872	18	15	0	4994	29	28	0	22	28
Qwen3.7-Plus	4411	2	0	1	4608	1	0	1	3	17
GPT-5.4 Pro	3662	20	20	0	3988	35	35	0	11	10
GPT-5.4	4209	1	0	1	4871	0	0	0	0	1
GLM-5V-Turbo	4091	11	6	0	4892	8	3	0	7	16
Claude Opus 4.6	4416	14	9	0	4978	22	21	0	24	37
MiniMax-M3	4138	12	4	6	4119	23	1	11	3	24
Qwen3.6-Plus	3691	6	0	6	3481	6	0	3	0	13
GPT-5.5 Pro	0	32	32	0	0	53	53	0	27	30
Gemini 3 Flash	4400	17	17	0	2420	39	39	0	47	56
Claude Sonnet 4.6	4460	19	15	0	4841	29	23	0	27	39
Gemini 3.5 Flash	0	27	27	0	0	44	44	0	35	53
Gemini 3.1 Pro	0	37	35	2	0	62	62	0	38	48
Kimi K2.5	0	23	1	22	4123	30	5	25	13	39
Kimi K2.6	0	27	8	19	0	43	14	29	18	41
Qwen3.5-397B-A17B	1954	14	1	13	3000	25	1	23	1	10
GPT-5.2	0	37	0	37	0	57	0	57	0	2
Grok 4.3	0	34	0	34	0	52	0	52	0	0
Qwen3.5-122B-A10B	0	40	0	39	0	62	0	61	0	5
Qwen3.5-35B-A3B	0	37	0	1	0	66	0	5	0	1



Figure 10: **Failure-mode examples.** (a) FM2, Place Naming ("What street are we on?"), intermediate view: the agent overlooks blue signage and crosswalks that suggest a nearby street sign, and answers "I do not know." (b) FM1, the same question at another location: from its final view, the agent answers "Las Vegas Boulevard South" based on the skyline and roadside businesses instead of exploring toward the intended view. (c) FM1, Navigation Challenge ("How many distinct, third floor patios does this building have?"): the model fixates on a house with a gabled roof rather than the intended building. Figures 12–14 give the full prompts and responses.

Table 9: **Score on answered episodes.** Mean GeoGuessr score | distance in km (Eq. 1) over episodes with a non-zero score on the two Precise Localization splits, Competition (n=42) and Synthetic (n=80), and the percentage of episodes that score zero, pooled over both (n=122). Zero scores come from abstentions, episodes that end at the step limit, and errors (Table 8). **Bold** marks the highest score and the lowest zero-score rate.

Model	Competition Score | Dist	Synthetic Score | Dist	Zero-score rate %
Claude Opus 4.8	4570.0 | 134.1	4666.2 | 103.1	3.28
Claude Opus 4.7	4759.7 | 73.5	4738.3 | 80.2	6.56
GPT-5.5	4578.1 | 131.5	4474.6 | 165.6	20.49
Claude Fable 5	4937.0 | 18.9	4963.4 | 11.0	38.52
Qwen3.7-Plus	4162.6 | 273.4	4046.5 | 315.6	2.46
GPT-5.4 Pro	4780.0 | 67.1	4742.0 | 79.0	45.08
GPT-5.4	4018.8 | 325.9	4221.1 | 252.6	**0.82**
GLM-5V-Turbo	4251.8 | 241.8	4288.7 | 228.9	15.57
Claude Opus 4.6	4535.9 | 145.3	4639.3 | 111.7	29.51
MiniMax-M3	4286.2 | 229.8	4078.3 | 303.9	28.69
Qwen3.6-Plus	3553.1 | 509.6	3196.1 | 667.5	9.84
GPT-5.5 Pro	4886.3 | 34.3	4610.7 | 120.9	69.67
Gemini 3 Flash	4789.6 | 64.1	4864.9 | 40.9	45.90
Claude Sonnet 4.6	4837.5 | 49.3	4523.6 | 149.4	39.34
Gemini 3.5 Flash	4755.2 | 74.9	4853.3 | 44.4	58.20
Gemini 3.1 Pro	**4986.5** | **4.0**	**4999.2** | **0.2**	81.15
Kimi K2.5	4456.9 | 171.5	4347.4 | 208.6	43.44
Kimi K2.6	4670.2 | 101.8	4541.3 | 143.5	57.38
Qwen3.5-397B-A17B	3332.2 | 605.3	3588.3 | 494.9	31.97
GPT-5.2	4400.8 | 190.4	4324.0 | 216.7	77.05
Grok 4.3	4230.7 | 249.2	4419.3 | 184.2	70.49
Qwen3.5-122B-A10B	3059.8 | 732.5	3587.7 | 495.1	83.61
Qwen3.5-35B-A3B	2120.6 | 1279.5	2084.7 | 1304.9	84.43

Table 10: Taxonomy of observed failure modes. Each mode represents a distinct category of error exhibited by visual agents during geospatial reasoning tasks.

ID	Failure Mode	Description	Affected Stage
FM 1	Vision Failure	Errors in interpreting visual input or applying pretrained knowledge, including hallucinated landmark recognition, target object misclassification, and faulty perception of scene content.	Vision
FM 2	Navigational Reasoning Failure	Errors in spatial and navigational reasoning, including poor route planning, imprecise movement execution, and lost spatial awareness across observations.	Planning / Execution
FM 3	Tool-Use Failure	Errors in tool interface use, including malformed or invalid calls, hallucinated tool capabilities, and refusal to engage due to misunderstood tool semantics.	Tool Use



Navigation Challenge Prompt: How many red lanterns are hanging along the next block south on Grant Avenue?

```
'Navigation': move_south x6
'Response': 6
'Reasoning': 'I moved south along Grant Avenue and captured the view of
    the next block. I counted 6 red lanterns hanging directly above
   the street on a wire/string, which are the street lanterns and not
   part of any store decorations.'
```

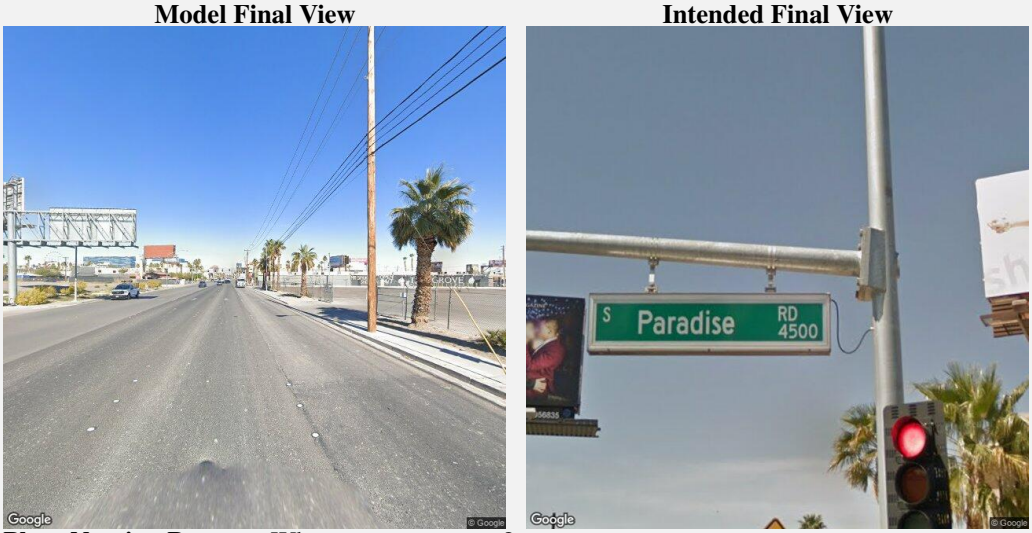
Figure 11: **FM2 (navigational reasoning failure) example: lack of object permanence.** Final view of Qwen3.5-122B-A10B for a Navigation Challenge prompt. The model miscounts the lanterns hanging just above and ignores the lanterns it passed on the way to this viewpoint.



Place Naming Prompt: What street are we on?

```
'Response': I do not know
'Reasoning': 'I have navigated through the area...but I cannot see any
    street signs or clear identifiers that would reveal the street name
    . The images show a road with double yellow lines, snow-covered
    ground, pine trees on one side, and buildings on the other, with
    mountains in the distance, but no visible street name signage.'
```

Figure 12: **FM2 (navigational reasoning failure) example: poor route planning.** Intermediate view of Qwen3.5-397B-A17B for a Place Naming prompt. The agent overlooks the blue signage and crosswalks, which suggest that a street sign is nearby.



Place Naming Prompt: What street are we on?

```
'Response': Las Vegas Boulevard South
'Reasoning': 'The skyline and roadside businesses indicate the Las
    Vegas Strip area, and the corridor shown matches Las Vegas
    Boulevard South.'
```

Figure 13: **FM1 (vision failure) example: hallucinated landmark recognition.** Final view of GPT-5.2 and the intended final view for a Place Naming prompt. Instead of exploring further and grounding its answer in concrete features of the scene, the agent infers its location from sparse evidence and pretrained knowledge.



Navigation Challenge Prompt: You are in the back parking lot of an apartment building. How many distinct, third floor patios does this building have?

Figure 14: **FM1 (vision failure) example: fixation on a distractor.** Intended target and the target chosen by Claude Opus 4.6 for a Navigation Challenge prompt. The model fixates on the house with a gabled roof.

Precise Localization Prompt: Identify the most precise coordinates of this starting location.

```
'Response': I do not know
'Reasoning': '...The available tools only allow me to navigate and
    capture street view images, but I cannot analyze visual content to
    identify geographic coordinates.'
```

Figure 15: **FM3 (tool-use failure) example.** Response of Qwen3.5-122B-A10B to a Precise Localization prompt before taking any action. The agent misunderstands what `capture_view` does (Section 4).