GeoGuesser Bench: Benchmarking Vision–Language Agents on Interactive Geolocation

Anonymous ACL submission

Abstract

1 Introduction

VLMs increasingly serve as the perceptual backbone of autonomous agents, with function-calling frameworks extending them to closed-loop decision making (Yang et al., 2023; Liu et al., 2023; Surís et al., 2023). Existing evaluations fall into two regimes: 2-D digital workflows (Trivedi et al., 2024) or single-shot visual QA. Neither exercises what physical-world deployment demands—actively controlling a viewpoint in 3-D, accumulating evidence across many rounds, and producing a precise quantitative answer.

Visual geolocation from street imagery is a demanding instance, requiring world knowledge (architecture, signage, climate), fine-grained perception, and deliberate exploration since disambiguating evidence rarely appears from a single heading. GeoGuessr (Haas et al., 2024; Plonk It, n.d.; Nicolas Dede, n.d.) has made this a competitive sport whose top players routinely localize unseen panoramas to within kilometers, providing a natural human baseline. Existing benchmarks evaluate single-shot localization from a fixed image (Haas et al., 2024) or test navigation in synthetic environments lacking real-world diversity.

We introduce GeoGuesser Bench: 282 real-world locations paired with an interactive Google Street View environment a VLM agent must navigate, look around, and zoom in. Entries stratify across world knowledge and navigational proficiency. Exact Geolocation predicts the starting panorama's coordinates; a 42-location subset from the GeoGuessr World Championship Grand Finals provides a calibrated human-expert reference. Coarse Geolocation identifies an encompassing region (street, city, country) from cultural cues. Precise Environmental Challenge counts environmental features near the spawn, stressing route planning and spatial continuity.

Evaluating frontier VLMs reveals a substantial gap to humans: the best systems achieve roughly half the GeoGuessr score of competition players. Failures cluster into hallucinated landmark recognition, breakdowns in spatial continuity, and malformed tool calls. Larger models are not uniformly better—stronger world knowledge does not eliminate navigational deficits, and scaling step budget helps unevenly across model families.

A monolithic VLM re-encoding its full panorama history incurs $\sim 10^7$ tokens per task. We propose a *Planner–Worker–Prefilter* framework that decomposes the agent into a large Planner (high-level reasoning), small Worker (atomic action grounding), and lightweight Prefilter (trajectory compression), more than halving planner cost while setting a new state of the art on Type 1 and Type 3.

Contributions. (i) GeoGuesser Bench, with 282 stratified entries and a calibrated human-expert competition subset; (ii) an interactive tool-driven Street View environment; (iii) a frontier-VLM evaluation identifying three failure modes, with ablations on FOV, step budget, and mobility; (iv) a context-compressed Planner–Worker–Prefilter framework halving planner token cost while improving accuracy.

2 Related Work

Visual geolocation. PIGEON (Haas et al., 2024) performs strong single-image localization via geographic priors and supervised training, extending a line of static-image methods that includes classification over learned geocells (Weyand et al., 2016), semantic geocell partitioning for interpretability (Theiner et al., 2021), dual-branch transformer architectures fusing image and segmentation features (Pramanick et al., 2022), CLIP-style con-

trastive alignment between images and GPS coordinates (Cepeda et al., 2023), retrieval-augmented generation with multimodal LLMs (Zhou et al., 2024), (Jia et al., 2024), hierarchical geographic embeddings combining retrieval and classification (Daruna et al., 2026), and distance-aware reranking over candidate locations (Jia et al., 2025). These efforts treat geolocation as static-image prediction; even recent agent-based methods that invoke external retrieval and OCR tools (Li et al., 2026) still reason over a single fixed view. GeoGuesser Bench reframes it as interactive, testing world knowledge alongside active perception, planning, and tool use.

Tool-use, embodied, and visual–spatial benchmarks. Leaderboard (Patil et al., 2025), AppWorld (Trivedi et al., 2024), and WTU-Eval (Ning et al., 2024) evaluate function-calling and stateful tool use, but observations are screenshots, documents, or structured text in 2-D digital workflows. Vision–language–action surveys (Ma et al., 2024) note that embodied benchmarks largely prioritize low-level manipulation over high-precision spatial estimation in visually rich scenes, and Agent-X (Ashraf et al., 2025) shows agents struggle with logical coherence across multi-step visual chains, though its setting remains image-based. GeoGuesser Bench complements these by requiring the same tool-use decision quality in a continuous 3-D scene at global scale, with quantitative outputs and a calibrated human-expert reference from the GeoGuessr World Championship.

Agent and tool-learning frameworks. MM-REACT (Yang et al., 2023), LLaVA-Plus (Liu et al., 2023), and ViperGPT (Surís et al., 2023) pair LLM controllers with vision skills; Bee (Zhang et al., 2025) shows dual-level chain-of-thought improves multimodal reasoning. Toolformer (Schick et al., 2023) teaches LLMs to call APIs via self-supervision; Gorilla (Patil et al., 2023) mitigates argument hallucination via retrieval-aware fine-tuning; broader surveys (Qin et al., 2023) cast tool use as a closed-loop intention–planning–action–observation cycle. Our Planner–Worker decomposition adopts this division of labor but specializes it for embodied geolocation, with the Prefilter compressing high-cost panoramic observations—a problem text-oriented frameworks do not face.

3 Benchmark Data Curation

The proposed dataset is designed to evaluate agent visual understanding, 3D spatial reasoning, and 3D navigation through the context of dynamic geolocation and visual question-answering.

3.1 Dataset Stratification

The dataset consists of 282 distinct coordinate pair + question entries, categorized into three primary categories to facilitate a nuanced analysis of model world knowledge and navigational proficiency. See Fig. 2.

Type 1 Exact Geolocation (n=80 +42): Comprised of 80 synthetically curated entries and 42 entries from GeoGuessr competition data, questions of this type simulate the act of geolocation or geoguessing, as commonly seen in the game GeoGuessr. Agents are prompted to output a coordinate estimation of their starting location. Optimal solutions require a high degree of world knowledge but a low degree of navigational proficiency with successful responses taking into account various, ambiguous environmental features such as vegetation, soil patterns, and landscape elevation as well as societal features such as signage and architectural styles. Although additional context may be achieved through further exploration of the environment, solutions are not based solely on individual environmental features (See Section 6.3 for a discussion on the necessary capabilities of model geolocation). See Fig.

Type 2 Coarse Geolocation (n=80): Questions of this type require agents to geolocate the general region containing the starting location (street names, countries, etc.). Agents are prompted to output the string name of the encompassing region. Optimal solutions rely less on natural world features but necessitate nuanced cultural knowledge and a relatively moderate degree of navigational proficiency and understanding. They involve bringing regional indicators that map directly to correct outputs into view such as the language of signage, road patterns, and architectural styles. Due to the solution pattern of the task, Type 2 questions are not a mapping of coordinate estimates to geographic regions (See Section 6.1). See Fig. **??**.

Type 3 Precise Environmental Challenge (n=80): Questions of this type require agents to answer visual question-answering prompts about the environment. Agents are prompted to output the string solution to the prompt. Optimal solutions

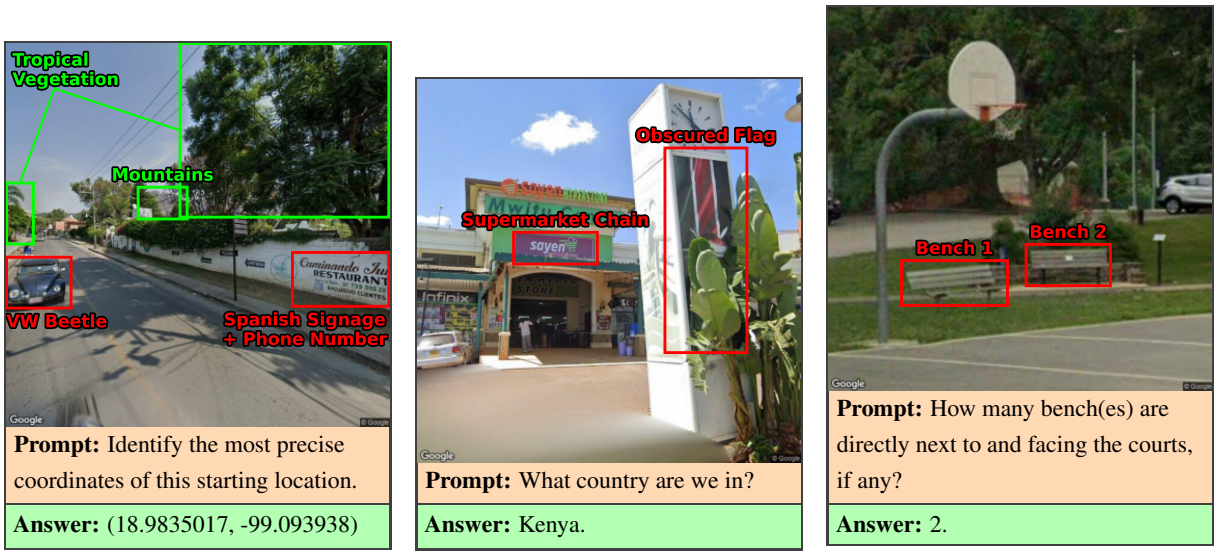


Figure 1: Examples of the three task types (Type 1, 2, 3 from left to right) with annotated relevant clues.

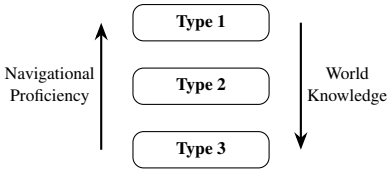


Figure 2: Task types ordered by navigational precision and world knowledge required.

require on average the least world knowledge but the highest degree of navigational precision and efficiency. These questions emphasize the reasoning and navigational path utilized to bring the verifiable target object(s) into view rather than inference of a solution based on environmental indicators. Type 3 questions are categorized into two sub-tasks:

1. **Identification (Finding) (n=52):** Agents locate one or more specific environmental landmark out of direct vision from the starting location, requiring efficient route planning as well as active comprehension and application of novel visual stimuli.

2. **Enumeration (Counting) (n=28):** Agents identify and count the number of a specific environmental feature, evaluating a model's ability to maintain spatial continuity during active exploration. See Fig. **??**.

3.2 Curation Methodology

Data collection utilized random sampling with filtering across all three question types to generate an initial location and question prompt:

Random Sampling: We randomly sampled valid street view coordinates to produce an initial global distribution then manually filtered locations to enforce geographical diversity. This ensures the evaluation of agent geolocational ability across various cultures, biomes, and geological landscapes. (See Fig. 3).

Filtering: We then manually inspected the starting viewpoint of each coordinate to ensure models are unable to immediately recognize famous locations from their pretraining. After the initial coordinate generation, we employed additional filtering for Type 2 and Type 3 prompt generation:

- **Type 2:** For Type 2 questions, we first determined a valid geographic region associated with each coordinate. We then identified visual clues that strongly map to the intended region. This process may involve adjusting the granularity of the target region (e.g., city to street) or adjusting the initial coordinates within the same region to better fit the task constraints. We prioritized suburban and urban environments that exhibit strong cultural and societal signals while adhering to the prevention of premature landmark recognition.

- **Type 3:** Type 3 questions were constructed by inspecting the local environment to identify the target object(s) that may be consistently and reasonably located through environmental exploration. These questions were designed to require active navigation and viewpoint adjustment through confirmation that the target

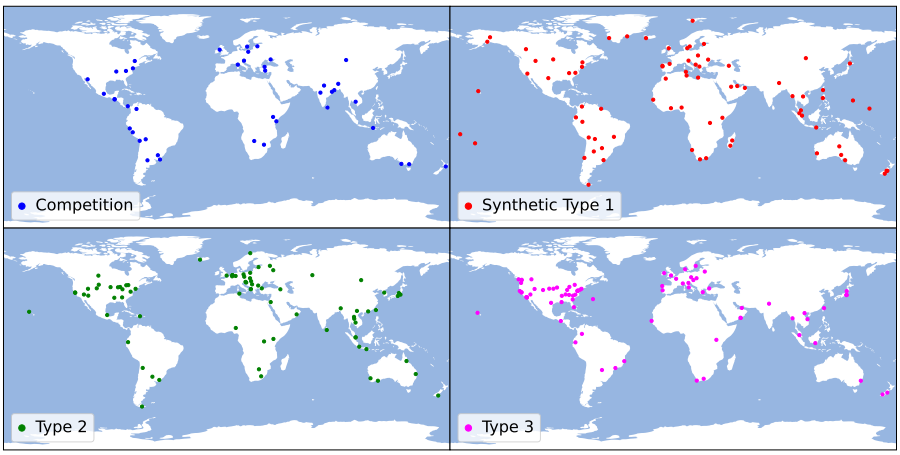


Figure 3: **Global Distribution of Question Types.** Note how the relative density of Type 2 and Type 3 questions increases around urban centers.

object(s) were not immediately visible from the initial viewpoint. Additionally, all questions were manually verified to be solvable within 40 agentic steps. We prioritized suburban and urban environments for their high visual contrast and findable/countable environmental features.

Type 1 Competition Subset: In addition to the 80 synthetically curated Type 1 entries, we provide a distinct Type 1 subset for evaluation consisting of 42 locations and the corresponding best player's performances. Taken from the Grand Finals of the Geoguessr World Championship in 2023, 2024, and 2025, each sample consists of the data from one game: the latitude-longitude coordinate pair of the initial location, the distance offset of the best expert estimation, and the GeoGuessr score of that guess. In each game, the competition provided a random location from GeoGuessr's global-scale map. Apart from the 1 minute time limit and player access to a map for placing their coordinate estimation, the objective and allowed navigational tools of the competitors are identical to those of the models. This set represents the highest known level of human expertise, functioning as an effective reference in comparing the capabilities of humans and models when both prompted Exact Geolocation queries. By including 42 human-benchmarked coordinates in addition to 80 synthetic coordinates, we provide expert comparison while also accounting for the case of model pretraining on the competition subset.

4 Environment Design

The GeoGuessr Research environment exposes a vision-language model to a street-view panorama as an interactive scene rather than a fixed image: at each step the model selects from a small action space and receives a fresh observation, deciding its next move based on its own strategy. Each run begins with the panorama at a starting coordinate and a high-level task, and proceeds as an open-ended sequence of tool calls up to a 40-step limit. The action space comprises `capture_view`, four `scroll_*` actions for camera rotation, `zoom_in`/`zoom_out` for FOV, `check_direction` for listing navigable directions, and eight compass `move_*` actions for stepping to a neighboring panorama. See Fig. 4.

5 Evaluation

Our approach to agent response evaluation differs between question types. Synthetic and Competition Type 1 responses require coordinate accuracy while Type 2 and Type 3 responses necessitate semantic correctness. Across all datasets, models are constrained to 40 agentic steps, with responses exceeding this limit being considered either 0 score or incorrect.

5.1 Type 1 (Exact Geolocation) Evaluation

Agent responses to Synthetic and Competition questions are evaluated through the haversine (or great circle) distance between the agent's coordinate prediction and the ground-truth coordinate, assuming a spherical Earth (See Appendix B).

In addition, Exact Geolocation response evaluations also include a metric inspired by the pop-

Figure 4: **Demonstration of Agent Tool-Use Trace.** Shown above is a 4 step tool-use trace of an agent and the corresponding effects of the tool calls on the agent's viewpoint. Agents may invoke multiple tools in the same step.

ular geolocation game Geoguessr. Although the exact scoring equation used by Geoguessr is not explicitly mentioned in official documentation, it has been shown (Haas et al., 2024; Plonk It, n.d.; Nicolas Dede, n.d.) to follow an exponentially decaying relationship as the distance between the estimated location and true location increases:

$$5000 \cdot \exp\left(-10 \cdot \frac{\text{Hav}}{\text{Rect}}\right) \quad (1)$$

Where 5000 is the maximum score of a guess and Hav is the haversine distance between the coordinates of a guess and the ground truth in km. **Rect=14916.862 km** and is the haversine diagonal of the smallest rectangle that would contain every location in the playable map, in this case a rectangle containing the entire Geoguessr world map (See Fig. 5). One advantage of the Geoguessr score is that it is a compact and variable system that increasingly differentiates estimates as they approach the ground truth. In addition, in the case where a model is unable to respond with a coordinate estimate, raw estimation errors in km are undefined so instead a score of 0 is assigned to the response. Another benefit of the GeoScore system is its recognition and accessibility through the GeoGuessr game, allowing for direct comparisons to human performance.

5.2 Type 2 (Coarse Geolocation) and Type 3 (Precise Environmental Challenge) Evaluation

Grounded in unambiguous environmental features, Coarse Geolocation and Precise Environmental Challenge questions have reproducible ground-truths, limiting the solution space of each question to predictable short phrases, individual words, or numbers. Correct responses are determined by performing sub-string matching on the model responses to a list of semantically equivalent ground-truth responses. We calculate the accuracy of each model over each data subset based on correct responses to total responses.

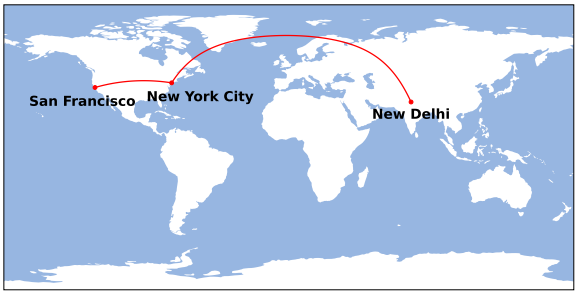


Figure 5: **Haversine Distance Visualization.** Haversine distance from New York City to San Francisco (~4129 km; score=313) and New Delhi (~11755 km; score=2).

6 Results

6.1 Agent Performance

The performances of various SOTA function calling models on the dataset are summarized in Table 1.

In the Synthetic and Competition Type 1 tasks, we observe a distinct deficiency of most agents in exact geolocation. A small number of agents approach average estimation errors at the region level (4373 score | 200 km) (Haas et al., 2024), as compared to human experts (4770.3 score | 70.2 km), on the Competition dataset. In addition, as seen in Table 3, 4 of the top-5 performing models required 3 or fewer movement tool calls, demonstrating that the exact geolocation task biases world knowledge over spatial understanding and exploration. This also implies 2 valid strategies for geolocation (See Section 6.3).

In the Type 2 task, the top models correctly infer the region 62.5% of the time. In Table 5, we observe that the top-5 performing models achieve high accuracies for regions of interest on the magnitude of countries (3024.2 score | 750 km) while achieving non-trivial accuracies in regions orders

Table 1: **Model Task Performance. Human expert average = 4770.3 | 70.2 km**. For Competition and Exact Geolocation tasks, cells show **Score | Dist**, where Score is the average GeoGuessr score and Dist (km) is the corresponding distance according to Eq. 1. Accuracy (%) is reported for Coarse Geolocation and Precise Environmental Challenge tasks. **Bold** indicates best performance per task type.

Model	Competition Score | Dist	Exact Geolocation Score | Dist	Coarse Geolocation %	Env. Challenges %
Claude Fable 5	2821.1 | 853.7	3164.2 | 682.5	**62.50**	**35.00**
Claude Opus 4.8	**4352.4** | **206.9**	**4549.5** | **140.8**	60.00	26.25
Claude Opus 4.7	4306.4 | 222.8	4501.4 | 156.7	56.25	31.25
Claude Opus 4.6	3023.9 | 750.2	3363.5 | 591.4	45.00	18.75
Claude Sonnet 4.6	2649.1 | 947.5	2883.8 | 820.9	47.50	12.50
GPT-5.5 Pro	1163.4 | 2175.0	1556.1 | 1741.2	47.50	31.25
GPT-5.5	3815.1 | 403.5	3467.8 | 545.8	**62.50**	32.50
GPT-5.4 Pro	2503.8 | 1031.7	2667.4 | 937.3	60.00	**35.00**
GPT-5.4	3923.1 | 361.8	4221.1 | 252.6	50.00	12.50
GPT-5.2	523.9 | 3365.1	1243.2 | 2076.1	33.75	17.50
Gemini 3.1 Pro	593.6 | 3178.8	1124.8 | 2225.3	47.50	30.00
Gemini 3.5 Flash	1698.3 | 1610.7	2184.0 | 1235.5	46.25	17.50
Gemini 3 Flash	2851.0 | 838.0	2493.3 | 1038.0	38.75	18.75
Kimi K2.6	1667.9 | 1637.7	2100.3 | 1293.8	50.00	13.75
Kimi K2.5	2016.2 | 1354.8	2717.1 | 909.7	42.50	15.00
Grok 4.3	805.8 | 2722.9	1546.8 | 1750.1	32.50	5.00
GLM-5V-Turbo	3138.2 | 694.8	3859.8 | 386.1	46.25	15.00
MiniMax-M3	3061.5 | 731.7	2905.8 | 809.6	33.75	17.50
Qwen3.7-Plus	3964.4 | 346.2	3996.0 | 334.4	42.50	23.75
Qwen3.6-Plus	3045.5 | 739.5	2956.4 | 783.8	36.25	15.00
Qwen3.5-397B-A17B	2221.5 | 1210.1	2467.0 | 1053.8	28.75	8.75
Qwen3.5-122B-A10B	145.7 | 5274.1	807.2 | 2720.3	12.50	6.25
Qwen3.5-35B-A3B	252.4 | 4454.4	364.8 | 3905.0	10.00	3.75

Table 2: **Failure Mode Proportion by Question Type.** Proportion of each failure mode out of total failures per task type, aggregated across all models. **Bolded** are the most common failure modes of each task type.

Task Type	FM 1	FM 2	FM 3
Type 1, Comp.	0.132	**0.603**	0.322
Type 1, Synth.	0.149	**0.620**	0.290
Type 2	0.301	**0.540**	0.179
Type 3	0.297	**0.513**	0.272

Table 3: **Average Movement Calls per Episode.** Mean number of `move_*` (panorama-traversal) tool calls per episode for the five strongest models (ranked by mean Competition/Exact Geolocation score), by task type and pooled across all four datasets. Counts are over completed episodes only.

Model	Comp	Type1	Type2	Type3
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-5.5	16.64	11.56	7.47	8.01
GPT-5.4	2.83	2.48	1.74	3.30
Qwen3.7-Plus	2.90	2.58	2.17	8.17

of magnitudes (4996.7 score | 1 km) below their respective average estimation errors in Table 1. This indicates that the Type 2 task is not simply an abstraction of the Type 1 task (i.e. mapping coordinate estimations to regions), while also demonstrating that models are indeed grounding their responses in concrete environmental clues. We attribute the comparable step counts of agents in the Type 2 task as compared to the Type 1 task to the fact that Type 2 questions are guaranteed to contain cultural indicators within the immediate, urban vicinity of the agents' starting location; meanwhile, Type 1 questions possess higher ambiguity in a more rural environment with no such grounding.

In the Type 3 task, no model achieves an accuracy above 35%. Furthermore, 16/21 models tested required a higher average number of agentic steps before responding than in both the Type 1 and Type

Table 4: **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 type. Statistics are computed over *completed* episodes only; entries that errored before producing a trace (e.g. Anthropic 413 `request_too_large` on long image-heavy episodes for Claude Fable 5 and Claude Opus 4.6) are excluded. Claude Opus rows use the direct Anthropic endpoint. The maximum step budget is 40.

Model	Competition Steps	Exact Geolocation Steps	Coarse Geolocation Steps	Env. Challenges 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 5: **Coarse Geolocation Accuracy by Question Granularity.** Accuracy (%) of the five strongest models on Coarse Geolocation, broken down by the geographic granularity asked for in the question. Sample sizes: Street (n=11), City (n=31), State (n=11), Country (n=27). **Bold** indicates best performance per granularity.

Model	<1 km	<25 km	<200 km	<750 km
Claude Fable 5	**45.5**	**58.1**	27.3	**88.9**
Claude Opus 4.8	27.3	41.9	**72.7**	**88.9**
Claude Opus 4.7	27.3	45.2	54.5	81.5
GPT-5.5	36.4	48.4	**72.7**	85.2
GPT-5.4 Pro	36.4	51.6	54.5	81.5

2 tasks, indicating the emphasis on thorough environmental exploration. In Fig. 6, we observe that models achieve a higher performance on the Identification (tool-use emphasis) subtask as compared to the Enumeration (spatial understanding emphasis) subtask. These results demonstrate the emphasis of Precise Environmental Challenge tasks on tool precision and spatial understanding/planning, as well as the deficiency of frontier models in both capabilities, but especially the latter.

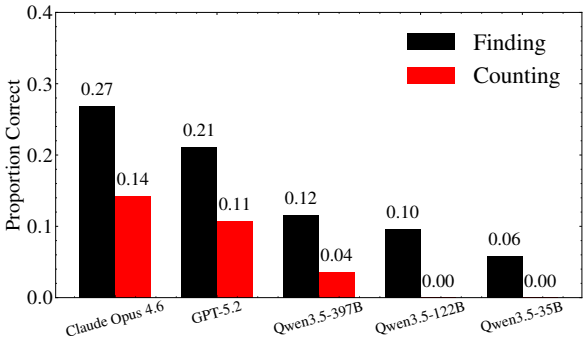


Figure 6: **Precise Environmental Challenge Sub-task Accuracy.** Shown is the accuracy of all tested models on both Type 3 sub-tasks.

Table 6: Valid response counts (and %) for each model across all task types. For Competition and Exact Geolocation tasks, a valid response is one where Score > 0. For Coarse Geolocation and Precise Environmental Challenge tasks, a valid response is one where the model completed within 40 steps. **Bold** indicates best performance (highest valid rate) per column.

Model	Competition (n=42)	Exact Geolocation (n=80)	Coarse Geolocation (n=80)	Env. Challenges (n=80)
Claude Opus 4.6	19/42 (45%)	38/80 (48%)	33/80 (41%)	26/80 (32%)
GPT-5.2	5/42 (12%)	23/80 (29%)	**80/80 (100%)**	**78/80 (98%)**
Qwen3.5-397B-A17B	**28/42 (67%)**	**55/80 (69%)**	79/80 (99%)	70/80 (88%)
Qwen3.5-122B-A10B	2/42 (5%)	18/80 (22%)	**80/80 (100%)**	75/80 (94%)
Qwen3.5-35B-A3B	5/42 (12%)	14/80 (18%)	**80/80 (100%)**	79/80 (99%)

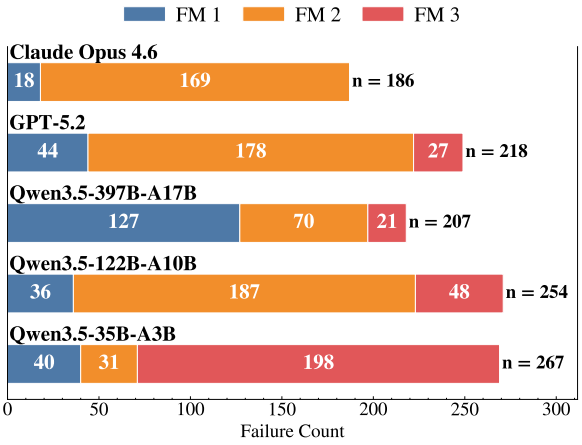
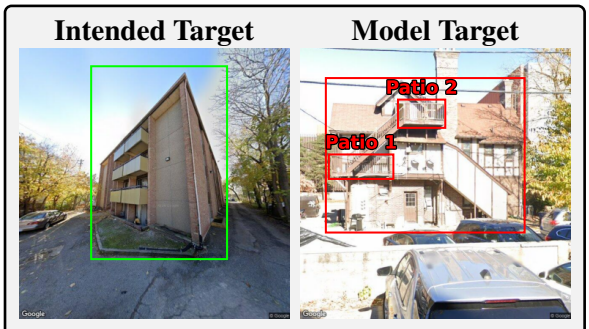


Figure 7: **Failure Mode Distribution by Model.** Absolute counts of each failure mode per model, aggregated across all four task types. Bar length corresponds to total failures per model; segment widths give the count of each failure category. Models may exhibit simultaneous failures so segment sums may exceed the labeled total n.



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

Figure 8: **FM 2 Subset Confusion Example.** Shown is the target subset of Claude Opus 4.6 and the intended subset for the given prompt. The model incorrectly fixates on the house with a gable roof in its response.

6.2 Failure Mode Analysis

Several recurring failure modes (FM) account for the majority of incorrect or invalid model responses, highlighting the deficiencies of these models in their world knowledge and navigational/spatial understanding (See Fig. 7 and Table 2). We categorize these observed modes as follows:

FM 1 Vision Failure: This failure mode describes errors arising from the model's interpretation of visual inputs and evocation of pretrained world knowledge, including hallucinated landmark recognition and failed image segmentation. This may manifest as misclassification of environmental clues or model fixation on red herrings. (See Fig. 8).

FM 2 Navigational Reasoning Error: This failure mode describes failures in spatial and navigational reasoning, including inefficient route planning and breakdowns in spatial continuity across successive observations. This may manifest as model inability to replan when given indicative context or enact directional instructions in the query. (See Fig. 16).

FM 3 Tool-Use Failure: This failure mode is defined as failures in the model's interaction with its tool interface that produce malformed, invalid, or unsupported requests. An agent may miscomprehend the use case of an available tool despite having access to function documentation, or it may hallucinate access to functions that are not currently exposed to it. These failures result in API errors, agent dismissal of the given prompt, or muddied agent context due to a bad request, ultimately preventing productive engagement with the environment.

In Fig. 7, we observe that large models exhibit a large proportion of FM 2 failures, especially Opus-

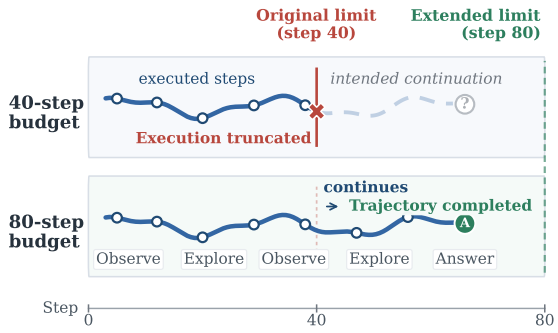


Figure 9: Caption

4.6, GPT-5.2, and Qwen-3.5 122B. This indicates a lack of spatial understanding and planning capabilities in 3D spaces. We also observe that that Qwen family models experience increased visual degeneracy and tool/navigational proficiency as model size increases. See Fig. 12, 13, and 14. We conclude frontier models with superior visual (i.e., world knowledge) and tool-calling capabilities continue to exhibit deficiencies in navigational reasoning and spatial awareness.

6.3 Analyzing the Roles of Visual Information and Navigation

To better understand the impact of visual information and navigation on geolocation performance, we conducted a series of ablation studies that systematically varied the degree of visual information and navigational freedom available to the agents. We **increased step limit** from step=40 to step=80, **immobilized** the agents by withholding their access to navigational tools, **cropped the bottom/top half** of agent viewpoint images, and exposed **web-search** tools. Ablations were conducted across proprietary and open source models of varying sizes.

Increased Step Limit:

We raised the per-entry step budget of the agents from the base limit from step=40 to step=80 in the Competition and Type 1 datasets. This allowed for the agents to properly exhibit their long-horizon geolocational capabilities and emergent strategies. See Table **??**.

We observe that Opus-4.6 improves drastically (~2100 → ~3700) while GPT-5.2 and Qwen-3.5's performance remains approximately constant. This demonstrates that Claude benefits greatly from additional exploration by exceeding the previous step limit and reasoning on a longer timescale.

Allowing the agents a higher step budget allows them to gather more diverse context across multiple panoramas or the complete context of a single panorama, depending on agent policy. As expected, Opus benefited the most as the model with a high proportion of responses in the basic configuration reaching the 40 step limit. Qwen-3.5 and GPT-5.2 had a low number of instances with the limits reached so the increased budget had a negligible impact on their performances. See Table 6.

Immobilization: We reverted the step budget to 40 and disabled the agents' access to movement `move_*` tools, leaving only the `capture_view`, `scroll_*`, and the `zoom_*` tools. The agent may therefore look and zoom in any direction from within the initial panorama, but may not navigate to any neighboring panorama. As such, we tested the ability of agents to gather sufficient visual context from a single panorama. We evaluated the same three models on the same two datasets. See Table **??**.

The effect of immobilization varies greatly across models. Claude exhibited a similar performance increase to when it was allowed a step budget of 80, Qwen's performance worsened by ~600 points, and GPT's performance remained constant.

When restricting the models to a single panorama with a 40 step limit, they were forced to gather a majority of the visual context of a single panorama (See Fig. 10). We predicted that the reduced variety of visual context would increase the uncertainty in model responses and negatively impact their geolocational performance. Although this was the case for Qwen, GPT's performance remained constant and Claude's performance improved compared to the basic configuration and nearly matched the performance of the 80 step ablation. Also, due to the 40 step limit, we observe that limiting exploration to a single panorama caused Claude to converge onto estimations within fewer steps.

Top/Bottom-Half FOV:

We sought to understand the effect of clues based on their vertical positioning within the environment. We did this by post-processed model viewpoint images to only include either the top or bottom half. This biased model behaviors to fixate on clues either on the top or bottom half, such as the type of foliage on trees or street markings on the ground. When exposed only the top half of their viewpoint as compared to the bottom half, 10/11 models on

Model	80 Steps		No Navigation		Top-Half FOV		Bottom-Half FOV		Web Search	
	C	T1	C	T1	C	T1	C	T1	C	T1
Claude Opus 4.8	**4594**	**4805**	**4646**	**4556**	**4368**	**4386**	**4369**	**4313**	**4551**	**4380**
Claude Opus 4.6	3708	3741	3749	3018	2527	2583	1928	2879	2688	3069
GPT-5.5	4407	4391	4100	4178	2212	2305	1403	1553	2388	1832
GPT-5.2	767	1295	572	1544	604	902	112	667	1696	2209
Gemini 3.1 Pro	1729	2473	1939	2532	0	1062	0	437	119	312
Gemini 3.5 Flash	2744	3228	3939	4019	948	2123	938	1179	0	125
Claude Sonnet 4.6	2791	3193	3626	4203	2337	2922	2446	2707	3206	2767
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
Qwen3.5-397B	2175	2149	1661	1861	1871	2235	1556	1147	2444	2662
Kimi K2.5	3038	3167	2668	3119	2258	2413	718	1819	3999	3979
Kimi K2.6	1218	1804	781	2161	471	1599	856	1185	238	507
Grok 4.3	650	1400	578	623	407	1133	213	820	835	1022

Table 7: **Ablation results across all conditions.** Mean GeoGuessr score (0–5000) per model on the Competition (C) and Type1 (T1) datasets under each ablation: *80 Steps* (step limit raised 40→80, navigation enabled); *Immobilized* (`move_*` tools removed, look-and-zoom only, 40 steps); *Top-/Bottom-Half FOV* (vertical half of each 1920×1080 capture removed, 40 steps, navigation enabled); *Web Search* (`search_engine_query` and `fetch_url_content` added, full images, 40 steps). Best value per column in bold.

Figure 10: **Immobilization Demonstration.** Example visual context gathered by Kimi K2.6 when restricted to the initial panorama.

the Synthetic dataset and 6/11 on the Competition dataset experienced performance increases.

Web Search: Models were exposed web search tools in order to trivialize the world knowledge component of geolocation. However, only 4/11 models experienced improvements in both datasets, while 5/11 experienced degeneracies in both datasets.

Ablation Analysis: These results imply that the visual context gathered by models within a single panorama resulted in performance comparable to the visual context gathered across multiple panoramas. Claude's performance increase from the basic configuration to the 80 step ablation demonstrates that its policy depended on exploration, otherwise its basic configuration performance would have matched that of the immobilization ablation. Assuming models respond once they receive sufficient visual context, we claim that the agents' exploration policies are flawed in such a way that extra navigation across multiple panoramas is required to match the visual evidence obtainable from a single, arbitrary panorama. While the visual capabilities of agents are impressive, this deficiency in their exploration policy demonstrates that frontier agents still lack in their 3D spatial and navigational understanding.

7 Context Compression in the Agent Framework

Single-shot VLMs underperform on GeoGuesser Bench because an isolated panorama rarely discriminates among locations with similar architecture, vegetation, or climate; reliable geolocation requires *active exploration*—iteratively selecting viewpoints, accumulating complementary observations, and committing only when evidence is sufficient. We realize this with a closed-loop agent that observes the current panorama, selects a viewpoint, executes atomic actions, and returns an answer once exploration terminates. Two design choices distinguish our framework from prior agent designs (ReAct (Yao et al., 2023), Reflexion (Shinn et al., 2023), Auto-GPT (Significant Gravitas, 2023), Voyager (Wang et al., 2023); see Appendix A): (i) a Planner–Worker split that routes global reasoning to a large model and fine-grained visual action grounding to a small one, and (ii) an explicit prefilter that compresses trajectory memory before the Planner sees it. The framework is applied uniformly across all three GeoGuesser Bench tasks and, as shown in Section 7.2, substantially im-

proves accuracy over the strongest single-shot baselines.

The split is motivated by both capability and cost: the Worker is invoked several times per planning round, so routing atomic actions through a small fine-tuned VLM reduces token cost and latency without compromising high-level decisions, while the structured planner-to-worker interface (one instruction in, one action out) lets either component be swapped independently. The prefilter is motivated by a token-cost analysis: re-encoding the growing trajectory at each round drives a single task to $\sim 10^7$ planner tokens, most of which redundantly re-encode past history. A lightweight VLM that summarizes the trajectory into a brief and focus hint cuts planner token consumption by more than half and, as shown in Table 9, can also improve accuracy by directing attention to the most informative evidence.

The loop comprises four components:

- **Planner.** A large vision-language reasoning model (e.g., Gemini 3 Flash). Conditioned on the task, the current exploration context, and a set of selected visual observations, it decides whether to continue or terminate.
- **Worker.** A small vision-language action model (e.g., Qwen2.5-7B-VL). Given the current image and the Planner's natural-language instruction, it predicts a single atomic action.
- **Environment.** A Street View navigator. The action space comprises *Move Forward*, *Turn Left/Right* at small or large angular increments, *Look Up/Down*, *Zoom In/Out*, and *Stop*.
- **Memory.** A trajectory store of recent steps and selected key frames, conditioning the Planner on past evidence rather than the most recent frame alone.

7.1 Framework Instantiation: Prefilter Planner

We instantiate the framework with a concrete *prefilter planner*. Conditioned on the task, the most recent actions, and the notes associated with selected key frames, the prefilter, a model substantially smaller than the Planner, produces a compact two-field summary consisting of a short `brief` that describes the current state and a `focus_hint` that identifies the evidence most pertinent to the next decision. The Planner subsequently reasons over this compressed summary in place of the raw history, which is the mechanism by which the planner token cost is reduced. As illustrated in Figure 11, the prefilter module is positioned between memory construction and high-level reasoning, performing trajectory compression prior to invocation of the main Planner.

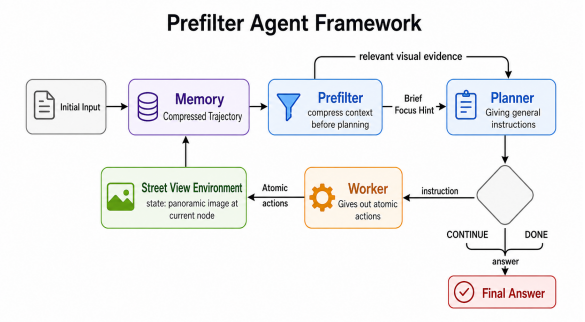


Figure 11: Prefilter agent framework. A lightweight prefilter model compresses the trajectory into a brief and focus hint, enabling the Planner to reason over a smaller yet more targeted context while preserving the Planner–Worker execution interface.

Importantly, the prefilter framework preserves the downstream interface unchanged: the Planner continues to emit `thought`, `status`, `instruction`, and `answer`. The introduction of prefiltering therefore modifies only the context-construction strategy, leaving the decision schema and the execution loop intact, which renders the with- and without-prefilter conditions directly comparable in the experiments that follow.

7.2 Comparison with Prior Results

Prior to presenting the token-cost analysis, we first verify that the compressed-context framework does not incur an accuracy penalty. Table 8 compares the prefilter agent against the strongest results reported earlier in the paper. The largest improvement is observed on Task 1 (Exact Geolocation), where our framework attains an estimated GeoGuessr score substantially exceeding the previous best model result. The proposed method also achieves the best performance on Task 3 (Environmental Challenges). Although performance on Task 2 (Coarse Geolocation) remains below the strongest previously reported score, the prefilter agent establishes new state-of-the-art performance on two of the three tasks, with a particularly large margin on the most demanding exact-geolocation setting, while, as shown in Table 9, achieving these results at a fraction of the planner token cost.

Task	Prior Best	Ours
Task 1 (Exact Geoloc.)	2467.0	**4737**
Task 2 (Coarse Geoloc.)	**37.50%**	30.00%
Task 3 (Env. Challenges)	22.50%	**24.39%**

Table 8: Comparison between the proposed agent framework and the strongest previously reported results in the paper. The proposed method outperforms prior results on Task 1 and Task 3, with a particularly large margin on exact geolocation.

	Task 1	Task 2	Task 3
Tokens (no PF)	11.06M	6.20M	11.18M
Tokens (PF)	4.36M	2.44M	4.80M
Reduction	60.6%	60.7%	57.1%
Result (no PF)	4686	30.00%	24.39%
Result (PF)	4737	26.25%	24.39%

Table 9: Ablation on the prefilter (PF) model. Disabling prefilter sharply increases token usage but does not improve accuracy. Task 1 reports estimated GeoGuessr score (MDE drops from 166.89 km to 89.56 km with PF); Task 2 and Task 3 report accuracy.

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. 2025. Agent-X: Evaluating deep multimodal reasoning in vision-centric agentic tasks. *arXiv preprint arXiv:2505.24876*.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Preprint*, arXiv:2309.16020.

Angel Daruna, Nicholas Meegan, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. 2026. Geosurge: Geo-localization using semantic fusion with hierarchy of geographic embeddings. *Preprint*, arXiv:2510.01448.

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

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

Pengyue Jia, Seongheon Park, Song Gao, Xiangyu Zhao, and Sharon Li. 2025. Georanker: Distance-aware ranking for worldwide image geolocalization. *Preprint*, arXiv:2505.13731.

P. Langley. 2000. Crafting papers on machine learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1207–1216, Stanford, CA. Morgan Kaufmann.

Qiujun Li, Zijin Xiao, Xulin Wang, Zhidan Ma, Cheng Yang, and Haifeng Li. 2026. Locationagent: A hierarchical agent for image geolocation via decoupling strategy and evidence from parametric knowledge. *Preprint*, arXiv: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. 2023. LLaVA-Plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*.

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

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

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

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. 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*, pages 48371–48392.

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

Plonk It. n.d. Beginner's guide to GeoGuessr. Accessed 2026-03-20.

Shraman Pramanick, Ewa M. Nowara, Joshua Gleason, Carlos D. Castillo, and Rama Chellappa. 2022. Where in the world is this image? transformer-based geo-localization in the wild. *Preprint*, arXiv: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, and 22 others. 2023. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*.

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

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

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

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

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

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

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. 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)*. ArXiv:2407.18901.

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

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

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

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

Yi Zhang, Bolin Ni, Xin-Sheng Chen, Heng-Rui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Han Hu, Meng-Hao Guo, and Shi-Min Hu. 2025. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open MLLMs. *arXiv preprint arXiv:2510.13795*.

Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. 2024. 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, page 2749–2754. ACM.

A Relation to Prior Agent Frameworks

Our design builds on a substantial body of work on LLM-based agents. ReAct (Yao et al., 2023) establishes the canonical paradigm in which a language model interleaves chain-of-thought reasoning with tool actions inside a closed reason–act–observe loop. Reflexion (Shinn et al., 2023) augments this loop with verbal self-criticism, while Auto-GPT (Significant Gravitas, 2023) popularizes fully autonomous orchestration in which a single LLM iteratively decomposes a user goal and dispatches sub-tasks to itself; Voyager (Wang et al., 2023) extends the paradigm to embodied settings by incrementally accumulating a library of executable skills through environment interaction. These frameworks share the common premise that long-horizon problems are best solved by alternating reasoning steps with environment observations. The framework in Section 7 instantiates this premise in the visual geolocation setting while departing from prior work in (i) decomposing high-level reasoning and low-level visual action across two models of very different capacity, and (ii) interposing an explicit prefilter stage between trajectory memory and the planner, which decouples context summarization from high-level reasoning.

B Haversine Distance Formula

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):

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

This quantity represents the shortest distance between the two coordinates over the surface of the sphere.

C Additional Figures and Tables

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

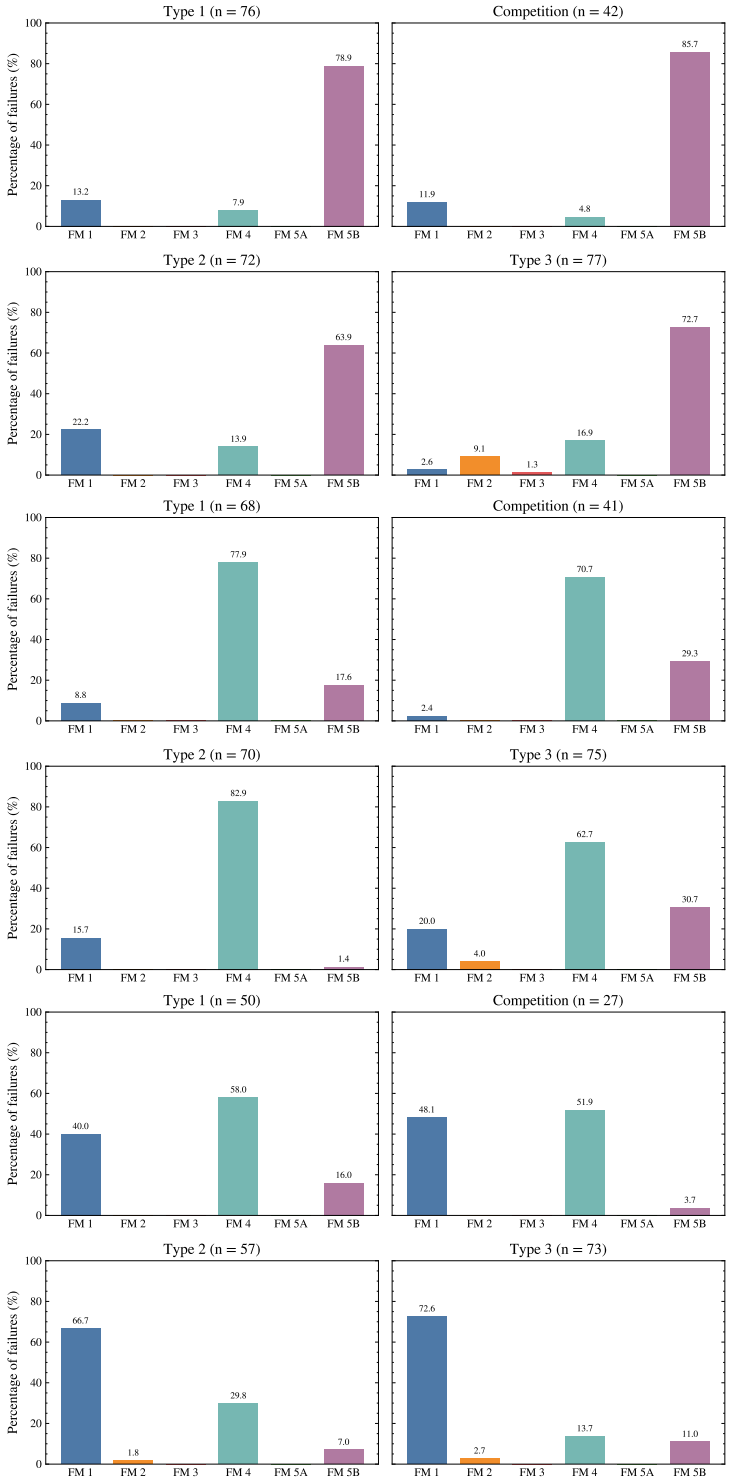


Figure 12: **Qwen3.5 Model Family Failure Distribution**.

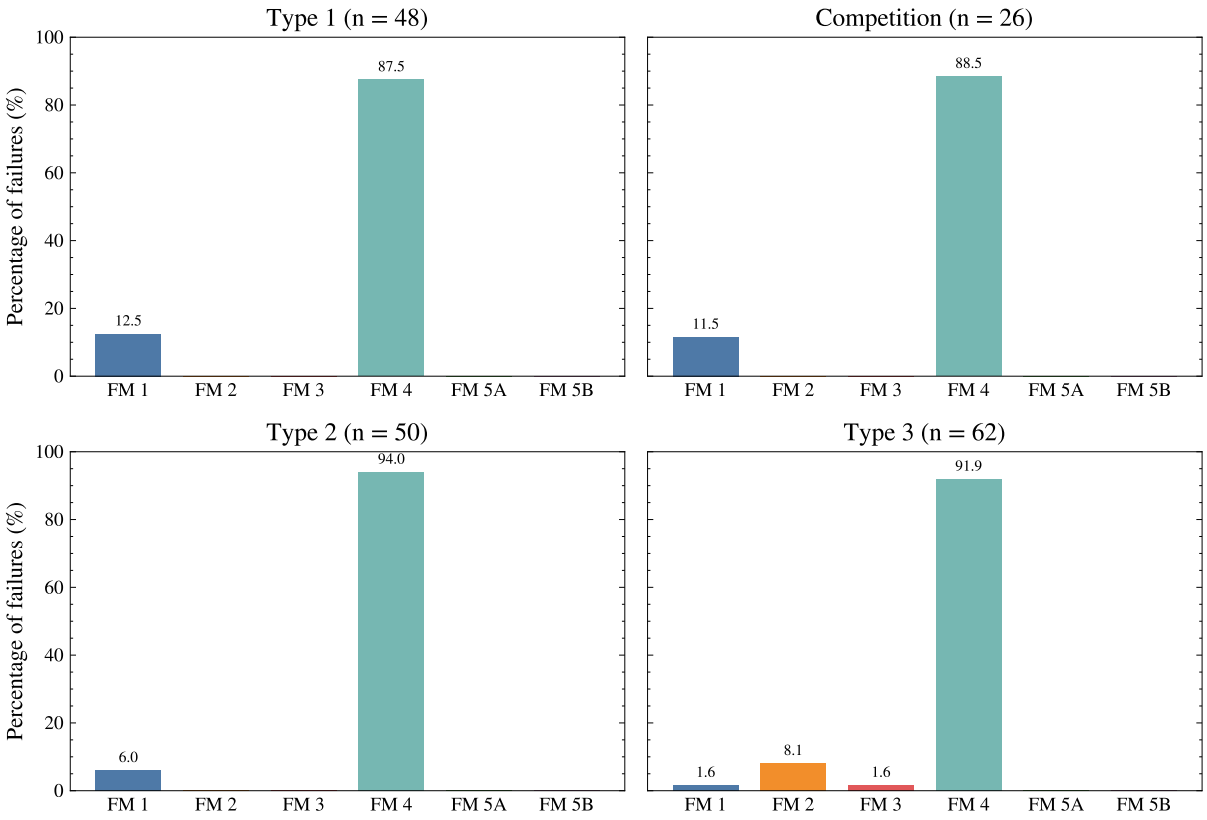


Figure 13: **Claude Opus 4.6 Failure Distribution**.

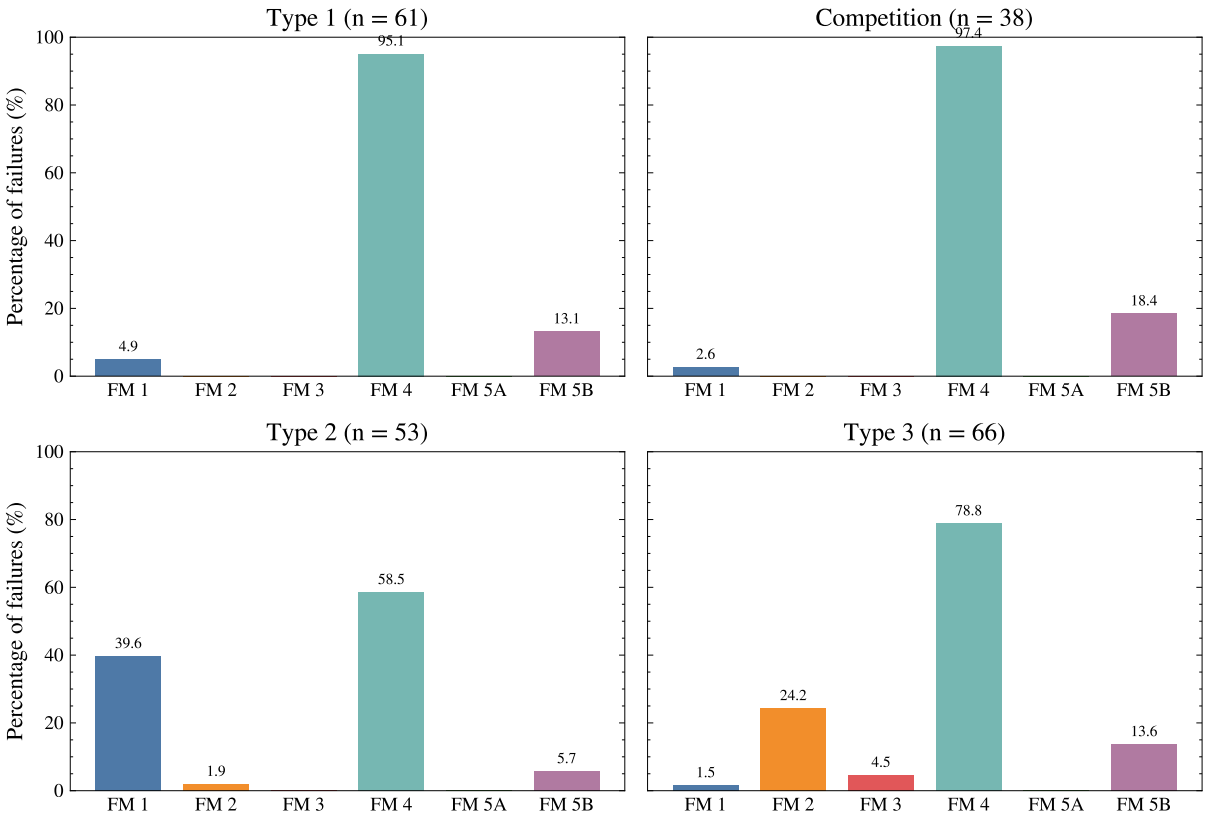
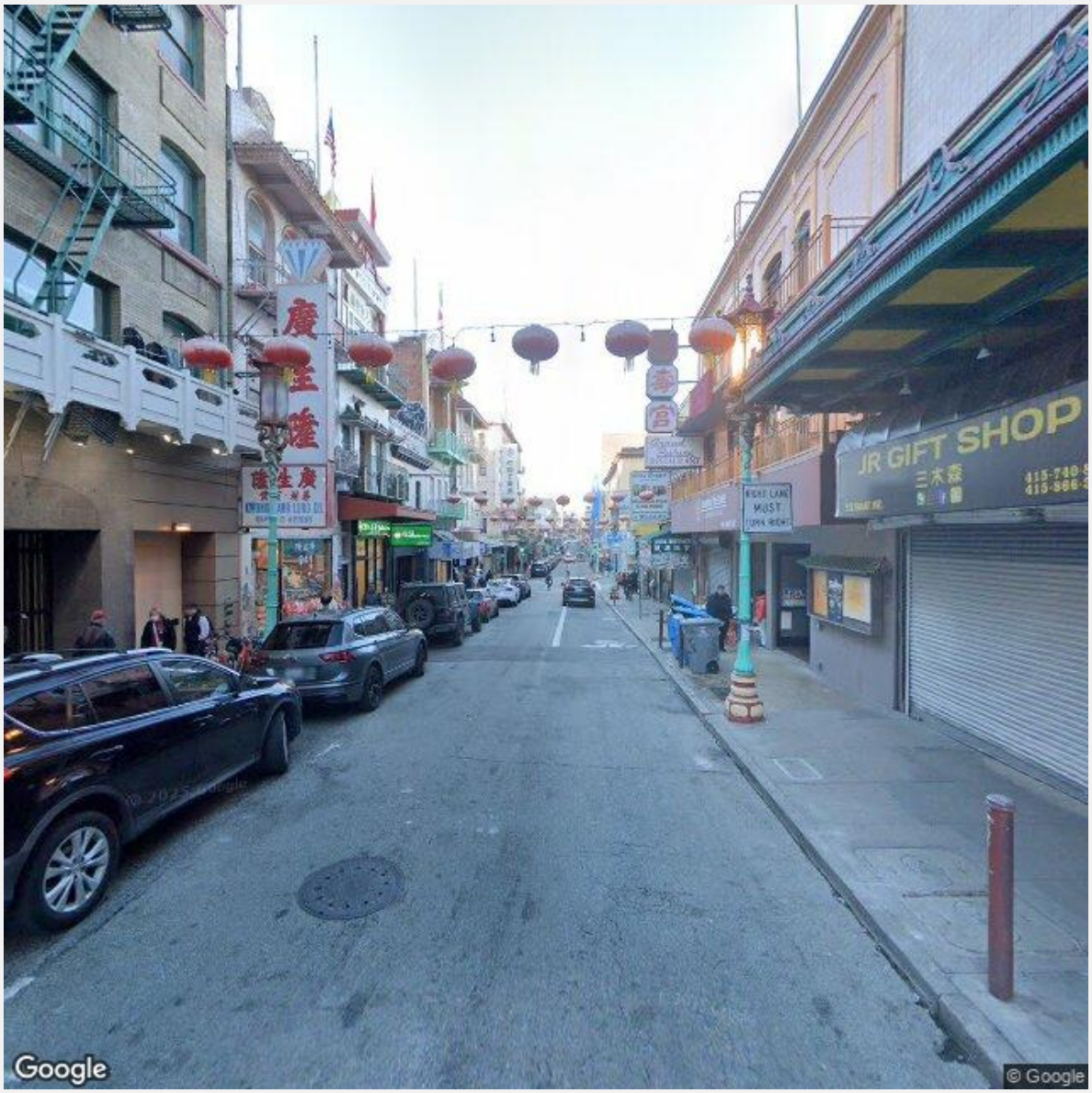


Figure 14: **GPT 5.2 Failure Distribution**.



Precise Environmental 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 15: **Lack of Object Permanence Example.** Shown is the final view of Qwen 3.5-122B-A10B for the given Type 3 prompt. Not only does the model miscount the number of lanterns hanging just above, it disregards the count of the other lanterns it passed by on the way to this viewpoint.

Coarse Geolocation 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 16: **Navigational Reasoning Failure Example.** Shown is an intermediate view of Qwen 3.5-397B-A17B for the given Type 2 prompt. The agent disregards the blue signage and crosswalks as likely indicators of street signs.