GeoAgent: Evaluating VLM Geolocalization Through Embodied Navigation

Anonymous ACL submission

Abstract

Modern Vision-Language Models (VLMs) perform well above the human baseline in geolocalization tasks. However, most efforts to study AI behavior on the task remain limited to static image-based retrieval, classification, and predictions. We argue that faithful recreation of the task should involve embodied navigation, where a multimodal agent autonomously explores its surroundings to gather observations before submitting a prediction. To this end, we introduce **GeoAgent**, an agentic environment-based benchmark that requires agents to navigate Street View environments to sequentially refine their geolocalization through reasoning. Our analysis shows modern VLMs struggle with discerning regional patterns while succeeding at country and continent-level guesses. When compared to static image-based baselines, agentic navigation significantly improves accuracy across established metrics. We also note severe bias in a developed/developing region context across frontier model architectures and poor self-improvement capabilities given incorrect priors. Overall, our work establishes the challenges of embodied navigation and geospatial reasoning. We publicly release our code and the GeoAgent environment: anonymous.4open.science/r/geoagent-F40E/

1 Introduction

Geolocalization from visual inputs has emerged as a challenging spatial reasoning task for modern AI systems. Among various tests, the popular game GeoGuessr[1] attracts the most attention (Haas et al., 2024; Cheng et al., 2025; Talreja et al., 2026; Wazzan et al., 2024; Jay et al., 2025; Suresh et al., 2018), which requires accurate geospatial predictions within a limited Street View-like navigable environment. Early AI approaches to this game framed GeoGuessr as an image classification task (Haas et al., 2024). More recently, Vision-Language Models (VLMs) have demonstrated remarkable zero-shot performance on geolocalization tasks (Cheng et al., 2025; Yi and Shan, 2025; Zhang et al., 2025; Li et al., 2025a), owing to their web-scale generalizability in interpreting visual cues such as signage, architectural styles, and vegetation patterns.

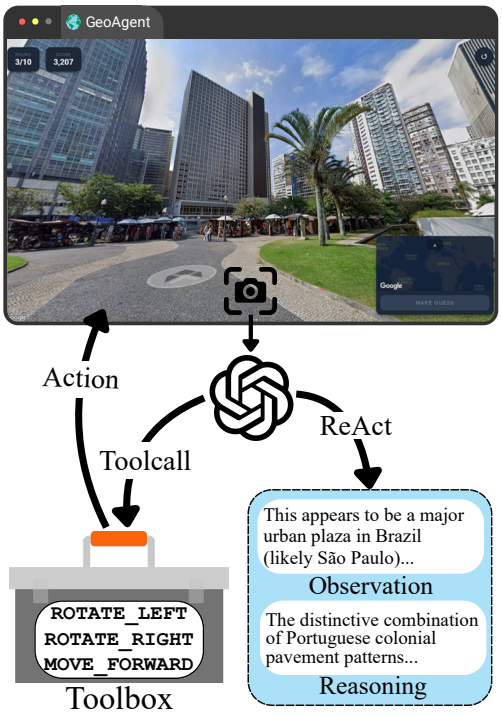

Figure 1: An overview of the GeoAgent framework, with an example VLM (GPT-5-Mini) interacting. Each agent is provided with a toolbox containing nine actions to explore the environment and submit a guess through iteratively refining their geolocalization prediction.

Most current evaluations assess models on single, static images, which fail to capture how humans actually perform geolocalization in a GeoGuessr-like setting, which involves active exploration of their environment. (Vyas et al., 2022) attempts to provide richer contextual clues utilizing

[1] https://www.geoguessr.com

video clips, but remains constrained by a passive observer. In real world environments, critical identifying features cannot be represented entire with a constrained set of static images. Passive inference traps the model in a fixed trajectory, leaving no room for resolving visual ambiguities. Embodied navigation aims to resolve this bottleneck, by allowing the agent to actively reason and execute navigational actions, such as movement and rotation, with the explicit goal of maximizing information gain. This approach of active, iterative exploration mirrors how human players would approach geolocalization and represents an unexplored frontier for evaluating spatial reasoning in VLMs. Beyond coverage, the fundamental limitation lies in the fact that passive inference denies the agent the ability to seek disambiguating evidence. Human geolocalization is an iterative process of active evidence-gathering and hypothesis formation. A fixed image set cannot replicate this loop, since the most informative view for a given location may simply not be among those captured. Evaluating VLMs only under passive inference therefore tests recognition, not reasoning under uncertainty. This is the distinction we achieve with GeoAgent.

Modern advances in tool calling (Patil et al., 2025; Yao et al., 2023; Erdogan et al., 2024; Ehtesham et al., 2025; Singh et al., 2025) and agentic web navigation (Wu et al., 2025; Deng et al., 2023; Zhou et al., 2023) have made simulating web-based workflows possible. Leveraging these capabilities, we introduce **GeoAgent**, both an open-source, modular, and reusable agentic environment and a benchmark that evaluates VLMs through embodied navigation in fully explorable Street View environments (Figure 1). The task involves two distinct steps: exploration actions with tool calls and geospatial reasoning for the final location guess. This enables us to study the following research questions that static evaluation does not capture:

- **RQ1:** Do VLMs improve geolocalization accuracy through active exploration, and does improvement depend on the correctness of initial hypotheses?
- **RQ2:** Do frontier VLMs exhibit systematic geographic bias between developed and developing regions?
- **RQ3:** How do exploration strategies differ across VLM architectures, and does action efficiency correlate with task performance?

Our benchmark comprises 1,200 fully navigable Street View locations across 100 cities spanning developed and developing regions worldwide, with difficulty stratified by recognizability tiers, as shown in Figure 2. We conduct comprehensive agentic behavioral evaluations of four frontier VLMs and compare them to static multiview baselines, revealing that models demonstrate limited self-correction capabilities when initial hypotheses are incorrect. Our analyses also reveal substantial bias among developed and developing geolocations across frontier model capabilities. To the best of our knowledge, our work represents the first comprehensive study of agentic behavior under geospatial embodied navigation constraints.

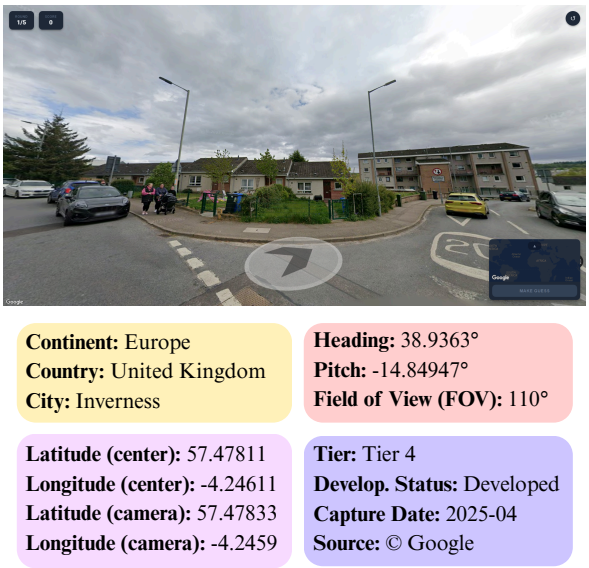

Figure 2: A screenshot of the navigable GeoAgent interface with the collected data.

2 Related Work

2.1 Vision-based Geolocalization

Early foundational works treated planetary-scale geolocalization as a discrete prediction task. *PIGEON* (Haas et al., 2024) established state-of-the-art performance using semantic geocell clustering and contrastive pretraining on *GeoGuessr* data. Yi and Shan (2025) democratized this with *GeoLocSFT*, achieving competitive results via efficient supervised fine-tuning on small, high-quality datasets (∼2.7K pairs). Despite their effectiveness, these specialist models operate as black boxes on single static images, lacking interpretability.

To address this, recent literature leverages Vision-Language Models (VLMs) for explicit reasoning. *GeoReasoner* (Li et al., 2024) employs

Tier	# Cities	Samples	Share (%)	Example Cities	Developed	Developing
Tier 1	5	34	2.85	New York, Tokyo	5	0
Tier 2	10	86	7.14	Rome, Mumbai	5	5
Tier 3	25	257	21.42	Bern, Bangalore	10	15
Tier 4	25	343	28.56	Oulu, Mombasa	12	13
Tier 5	35	480	40.00	Inuvik, Siem Reap	14	21
Total	**100**	**1,200**	**100.00**	–	**46**	**54**

Table 1: Tier-wise composition of the final dataset (1,200 samples) with example cities and development status.

Action	Description
ROTATE:<degrees>	Rotate to a specific heading (0–360°)
ROTATE_LEFT	Rotate 90° left from current view
ROTATE_RIGHT	Rotate 90° right from current view
ROTATE_BEHIND	Rotate 180° to look behind
LOOK_UP	Tilt the camera up by 20°
LOOK_DOWN	Tilt the camera down by 20°
MOVE_FORWARD	Move forward along the road or path
RETURN_START	Return to the starting position
GUESS	Submit the final location guess

Table 2: Available tool calls for agents to explore the GeoAgent environment.

reasoning-tuning to extract visual cues (e.g., signage, vegetation), while *GLOBE* (Li et al., 2025a) introduces a bi-objective optimization framework to assess image localizability before predicting. Concurrently, *Geo-R1* (Xu et al., 2025) utilizes reinforcement learning to instill a geospatial thinking paradigm, and *GeoGuess* (Cheng et al., 2025) establishes benchmarks for hierarchical reasoning (Country → City). However, evaluations like *IMAGEO-Bench* (Li et al., 2025b) and *FairLocator* (Huang et al., 2025) reveal that general-purpose VLMs (GPT-4V and Gemini) exhibit geographic biases in non-Western regions. *GAMa* (Vyas et al., 2022) substitutes static images with video clips for cross-view geo-localization, yet the observer cannot act to resolve ambiguity. Crucially, all these approaches remain confined to *passive* inference by analyzing fixed views, unlike GeoGuessr.

2.2 Agentic Web Navigation

Recently, AI evaluation has shifted toward autonomous interaction, primarily in 2D web environments. Benchmarks like *WebArena* (Zhou et al., 2023) and *Mind2Web* (Deng et al., 2023) evaluate agents on long-horizon tasks requiring DOM manipulation. Advanced methods have since optimized this agency: *Synapse* (Zheng et al., 2024) introduces memory-augmented prompting (trajectory-as-exemplar) for complex workflows, *RealWorldWebAgent* (Gur et al., 2024) employs efficient planner models (HTML-T5), and *WebWalker* (Wu et al., 2025) focuses on systematic subpage traversal. In the geospatial domain, agency remains nascent. *NAVIG* (Zhang et al., 2025) introduces limited tool use (map retrieval, knowledge bases), and *GeoComp* (Song et al., 2026) models human-like deduction chains. However, these geo-agents remain physically static, unable to navigate the environment, while web agents operate in semantic (not spatial) 2D spaces.

GeoAgent bridges this gap by treating geolocation as *embodied, interactive exploration*. Unlike web agents navigating DOM trees, we formalize navigation actions (movement, rotation) within 3D environments (Google Street View). By evaluating models' ability to actively reduce uncertainty through physical exploration, we establish a new paradigm for assessing spatial intelligence.

3 Method

3.1 Task Setup & Data Collection

GeoAgent involves a navigable web UI that can be traversed through tool calls to explore the surrounding street signs, vegetation, and architecture for the final guess. We design a minimal interface that simulates navigation environments akin to OpenGuessr[2] with Google Street View. VLMs must autonomously explore a provided environment, explain their observations, and reason through a guess in every step. We limit the number of exploration steps to 8 as a principled design choice reflecting realistic agentic deployment constraints. As shown in Appendix B, mean Haversine distance plateaus and occasionally degrades beyond action 5, indicating that agents either commit to a hypothesis early or fail to update it meaningfully with further exploration. An unconstrained action budget would therefore not improve benchmark signal. It would instead amplify confirmation bias and inflate API costs without yielding additional diagnostic value.

[2] https://openguessr.com

We initially collect a list of 100 cities across the globe based on two criteria: developed/developing status per the UN Council for Trade and Development[3] and recognizability tiers as outlined in Table 1. We classify the cities based on how many visitors they attract annually[4]. This ensures a difficulty classification that closely resembles web-scale training data makeup for most modern LLMs. For instance, urban agglomerates like New York City and Paris are ranked in Tier 1, while Butare, Rwanda, and Imphal, India are placed in Tier 5.

For every city, we randomly sample 100 latitude and longitude pairs within a 2-mile radius of their known coordinates as retrieved from Google's Geocoding API, resulting in a dataset of 10,000 records. Coordinates that result in no navigable Street View environments are dropped, and we finally retain 1,200 samples in our dataset with an unbalanced bifurcation as shown in Table 1. This ensures the validity and difficulty of our data, as it would resist solutions from pure pattern matching or training data regurgitation and instead force fine-grained geospatial reasoning. A sample of the collected data is shown in Figure 2. A total of nine tool calls are defined for the model to interact with the environment, as shown in Table 2.

3.2 Evaluation Metrics

We evaluate agent performance using three complementary metrics that capture geolocation accuracy, exploration effectiveness, and reasoning patterns.

Geospatial Distance. We measure geolocation quality using the great-circle (Haversine) distance between predicted and ground-truth coordinates. Given latitude/longitude pairs (ϕ_1, λ_1) (prediction) and (ϕ_2, λ_2) (ground truth) in radians:

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right), \tag{1}$$

$$c = 2\arctan 2(\sqrt{a}, \sqrt{1-a}), \tag{2}$$

$$d = R \cdot c, \tag{3}$$

where $\Delta\phi = \phi_2 - \phi_1$, $\Delta\lambda = \lambda_2 - \lambda_1$, and $R =$ 6371 km is Earth's radius. We report both mean and median distances, as the distribution is highly skewed by catastrophic failures.

[3] https://unctadstat.unctad.org/EN/Classifications.html

[4] https://www.statista.com/chart/12742/the-worlds-most-visited-cities

Improvement Rate. To quantify the exploration effect, we measure how much the agent's guess improves from its initial prediction to action k:

$$\text{improvement}_k = \left(1 - \frac{d_k}{d_1}\right) \times 100, \tag{4}$$

where d_1 is the distance at action 1 (initial guess) and d_k is the distance at action k. Positive values indicate the agent moved closer to the correct location. We compute this separately for rounds with correct versus incorrect final city guesses to analyze exploration patterns.

TF-IDF Vocabulary Analysis. We identify characteristic vocabulary used by VLMs about different geographic categories using term frequency-inverse document frequency (TF-IDF). For each round, we concatenate the `observations` and `reasoning` fields from all actions. We preprocess the resulting text by: (i) lowercasing, (ii) removing non-alphabetic characters, (iii) removing English stop words, (iv) extracting unigrams and bigrams (`ngram_range=(1,2)`), and (v) filtering terms appearing in $> 95\%$ of documents (`max_df=0.95`).

All extracted text segments from the same country are concatenated into a single document. For instance, the "Japan" document contains all observation and reasoning text from rounds where the ground-truth location was in Japan. Finally, we compute top-N TF-IDF terms per country.

3.3 Baselines

To compare against agentic exploration, we establish two multi-view baselines that stitch several shots of the environment into a single image for direct inference. The first, 4-view, captures a panorama at 0°, 90°, 180°, and 270°; the second, 8-view, augments this with a second panorama taken after a single MOVE_FORWARD action. These baselines enable a direct comparison between stepwise agentic exploration with ReAct-based reasoning and one-shot multi-image geolocalization (for examples, see Appendix D).

3.4 Models Used

We test five leading VLMs on our task. Among open options, we choose Gemma 3 27B and Llama 4 Scout 109B with Claude Haiku 4.5, Gemini 3.0 Flash, and GPT-5 Mini forming the closed subset. These choices are motivated by the tool-calling and reasoning capabilities of these models. All VLMs are tested through their proprietary APIs, other than

(a) 4-View Condition

Model	Mean Dist.	Median Dist.	Reasoning Len.	Continent Acc (%)	Country Acc (%)	City Acc (%)
Claude Haiku 4.5	2951.669 ± 48.885	1114.558 ± 27.436	329.889 ± 1.556	64.4 ± 0.2	43.5 ± 0.1	9.4 ± 0.2
GPT-5 Mini	**776.741** ± 10.927	232.525 ± 0.104	205.881 ± 0.061	**90.3** ± 0.7	**80.4** ± 0.6	30.3 ± 0.4
Gemini 3.0 Flash	1446.793 ± 74.819	**118.524** ± 11.726	232.314 ± 1.944	70.0 ± 0.5	65.6 ± 0.6	**39.5** ± 1.0
Gemma 3 27B	2669.221 ± 17.961	754.691 ± 0.542	**198.007** ± 0.245	49.5 ± 0.3	37.1 ± 0.2	12.7 ± 0.2
Llama 4 Scout	1586.080 ± 61.202	457.951 ± 3.666	200.437 ± 0.301	72.7 ± 0.3	59.9 ± 0.7	20.8 ± 0.3
Random Baseline	9958.248 ± 73.832	9951.232 ± 164.865	0.000 ± 0.000	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

(b) 8-View Condition

Model	Mean Dist.	Median Dist.	Reasoning Len.	Continent Acc (%)	Country Acc (%)	City Acc (%)
Claude Haiku 4.5	2603.224 ± 39.610	944.665 ± 36.292	338.618 ± 2.069	68.7 ± 0.4	47.9 ± 0.6	9.9 ± 0.6
GPT-5 Mini	**740.117** ± 44.042	205.691 ± 6.466	**200.493** ± 0.340	**91.5** ± 1.1	**82.9** ± 1.2	32.3 ± 0.6
Gemini 3.0 Flash	1247.272 ± 34.074	**84.078** ± 5.218	231.867 ± 1.321	71.9 ± 0.7	67.3 ± 0.8	**41.5** ± 0.5
Gemma 3 27B	3092.037 ± 12.714	888.743 ± 6.712	205.121 ± 0.312	44.3 ± 0.4	33.4 ± 0.3	12.0 ± 0.3
Llama 4 Scout	1592.635 ± 36.715	453.513 ± 8.914	203.163 ± 0.414	75.0 ± 0.4	61.5 ± 0.7	20.3 ± 0.8
Random Baseline	9992.070 ± 193.054	9951.671 ± 268.035	0.000 ± 0.000	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

(c) Agentic Navigation

Model	Mean Dist.	Median Dist.	Reasoning Len.	Continent Acc (%)	Country Acc (%)	City Acc (%)
Claude Haiku 4.5	2181.395 ± 86.864	688.333 ± 32.868	424.183 ± 1.983	68.4 ± 2.6	52.4 ± 2.7	14.3 ± 1.3
GPT-5 Mini	**899.989** ± 73.514	182.333 ± 11.590	**182.772** ± 0.718	**88.6** ± 0.2	**81.2** ± 0.7	34.5 ± 1.4
Gemini 3.0 Flash	1099.383 ± 99.775	**5.033** ± 0.153	283.474 ± 27.276	73.8 ± 1.2	69.8 ± 1.4	**50.1** ± 0.5
Gemma 3 27B	1689.328 ± 147.503	401.500 ± 48.195	268.524 ± 100.382	63.9 ± 1.9	54.0 ± 3.1	19.3 ± 1.7
Llama 4 Scout	1779.452 ± 35.500	579.667 ± 12.702	200.352 ± 0.200	70.1 ± 1.2	58.4 ± 1.3	14.7 ± 1.0
Random Baseline	10089.154 ± 94.161	10210.333 ± 184.114	0.000 ± 0.000	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

Table 3: Performance comparison across base models. **Bold** indicates best, underline indicates second-best (excluding Random Baseline). Accuracies reported in %. Lower distance and reasoning length are better; higher accuracy is better.

Model	Actions	Scrn.	Moves	Rotations
Claude Haiku 4.5	5.61±0.49	5.61±0.49	2.79±0.36	1.59±0.14
Gemini Flash 3.0	4.96±0.25	4.96±0.25	3.57±0.25	0.39±0.02
Gemma 3 27B	5.96±0.25	5.96±0.25	4.20±0.34	0.76±0.49
Llama 4 Scout	7.55±0.04	7.55±0.04	0.41±0.07	6.14±0.07
GPT-5 Mini	4.94±0.07	4.94±0.07	2.42±0.12	1.51±0.06
Random Baseline	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00

Table 4: Mean exploration behavior across all models, averaged over developed and developing regions (N=1200; 3 runs each). Llama 4 Scout saturates the action budget but almost exclusively rotates, while Gemini Flash 3.0 favors forward movement over rotation.

Llama 4 Scout, which we access through Groq. The temperature is set to 0.3 if available.

4 Results

Table 3 summarizes our main results, where Gemini 3.0 Flash takes the lead over the remaining models (Haversine distance of 1099.38 km, and 50.1% city prediction accuracies) with Claude Haiku 4.5 performing the worst (2181.40 km mean Haversine distance). Interestingly, GPT-5 Mini outperforms Gemini on continent accuracy (88.6%) and country accuracy (81.2%). Notably, all models show significantly worse performance on mean Haversine distance despite strong continent- and country-guessing capabilities. We also notice higher standard deviations on agentic navigation against the 4-view and 8-view baselines, but the numbers remain contextualized to the random baseline. The substantial gap between country-level accuracy (52.4–81.2%) and city-level accuracy (14.3–50.1%) across all models suggests that VLMs can identify broad geographic regions from macro-level cues (vegetation, climate indicators, driving side) but struggle with the "last-mile" localization based on municipal infrastructure, local business signage, or regional architectural variants. Gemma 3 27B demonstrates strong performance, beating both Claude Haiku 4.5 and Llama 4 Scout in city prediction accuracies and Haversine distance despite its smaller size. Another important observation is the inversion of reasoning length with performance on GeoAgent. GPT-5 Mini's outputs are the least verbose (182.77 characters on average) while Claude Haiku 4.5's are the longest (424.18 characters).

In comparison to the 4-view and 8-view baselines, we notice significantly improved geolocalization performance across both Haversine distance (1446.79 ± 74.82 km on 4-view to 1099.38 km on agentic for Gemini 3.0 Flash) and city-level accuracy, which sees the largest relative gains: Gemini 3.0 Flash improves from 39.5% (4-view) to 50.1% (agentic), a 26.8% relative increase, while GPT-5 Mini rises from 30.3% to 34.5%. This suggests

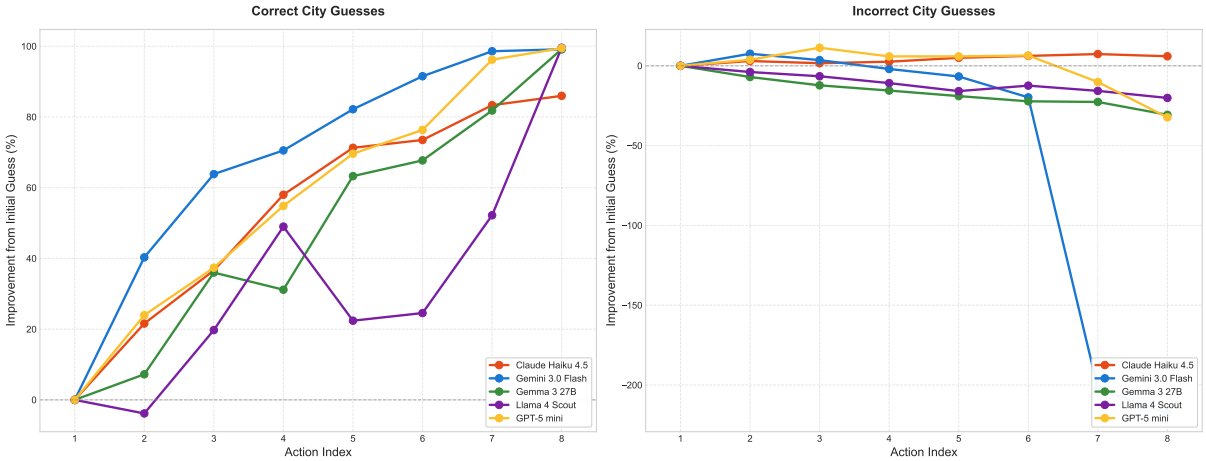

Figure 3: Aggregated performance improvement per action (out of 8) per model, stratified by final correct and incorrect guesses. We notice model performance generally improves with each extra action, but calibrates when the outcome is incorrect. Gemini 3.0 Flash is an outlier with a collapse on incorrect city guesses beyond 5 actions.

	Agentic				4-view				8-view			
	Cont. Acc (%)	Country Acc (%)	City Acc (%)	Mean Dist (km)	Cont. Acc (%)	Country Acc (%)	City Acc (%)	Mean Dist (km)	Cont. Acc (%)	Country Acc (%)	City Acc (%)	Mean Dist (km)
Claude Haiku 4.5	-1.7	-3.0	+3.5	-594	+15.7	+9.6	+1.0	-1936	+15.3	+7.8	+2.5	-1884
GPT-5 Mini	+7.1	+5.7	-2.6	-811	+7.7	+5.0	-2.3	-683	+7.8	+7.2	-0.8	-632
Gemini 3.0 Flash	-22.1	-25.9	-6.4	-290	-22.3	-22.4	-2.9	-452	-22.1	-23.0	-3.4	-352
Gemma 3 27B	-22.0	-22.0	-0.6	-394	-15.5	-19.7	+2.0	-1824	-15.6	-18.7	+0.4	-1868
Llama 4 Scout	-10.8	-14.7	-0.4	+501	-2.1	-2.9	-1.9	-836	+1.7	-3.9	-1.4	-1123

Figure 4: Developed–developing performance delta across all models and navigation modes. Each cell shows the difference (developed –developing); negative values indicate stronger performance on developed regions. All models exhibit a developed-region bias in mean Haversine distance (50–103% lower), and most secure higher continent, country, and city accuracies for developed regions. Gemma 3 27B is an exception, with slightly higher city accuracy in developing regions across all three conditions. Gemini 3.0 Flash shows the largest country-level deltas (>20%) in all modes, suggesting its strong overall performance is disproportionately driven by developed-region familiarity.

that step-wise exploration allows models to iteratively refine coarse regional guesses. Notably, not all models benefit equally from exploration: Llama 4 Scout's mean distance actually worsens from 1586.08 km (4-view) to 1779.45 km (agentic), and Gemma 3 27B sees degraded continent accuracy (49.5% $\rightarrow$ 63.9% under 4-view to agentic, but 44.3% under 8-view).

4.1 Can VLMs improve geoguessing performance through embodied navigation?

To study whether additional navigation improves geoguessing capabilities, we quantify agent behavior over each action in Figure 3. For guesses that are finally correct, we notice that most models show significant performance improvement with more navigation. For incorrect guesses, however, most models show limited sensitivity to additional exploration. Gemini 3.0 Flash, interestingly, shows an exponential drop in accuracy with added navigation steps beyond action 5 (Figure 3, right). Claude Haiku 4.5, in contrast, shows the most calibration as added navigation does not significantly improve or worsen its guesses. While exploration improves performance when initial hypotheses are approximately correct (RQ1), VLMs demonstrate poor self-correction capabilities. Additional actions rarely recover from incorrect priors, suggesting confirmation bias in sequential reasoning.

We further investigate what distinguishes recoverable from non-recoverable initial errors by analyzing the relationship between first-action distance and final accuracy. For rounds where models

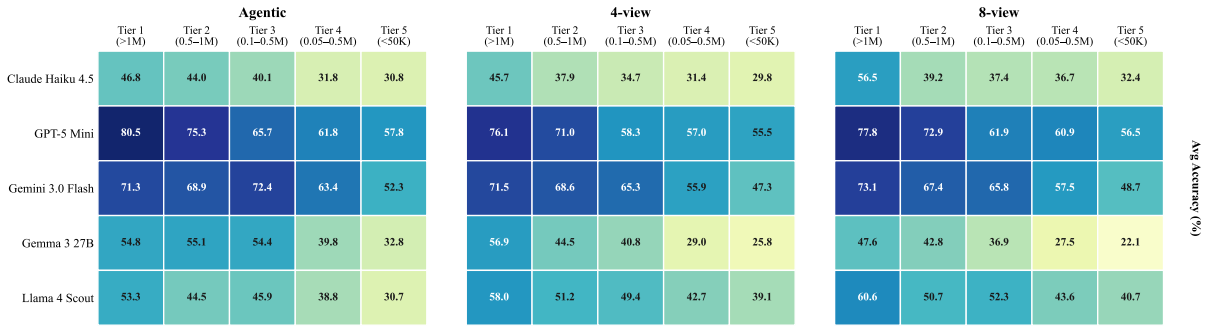

Figure 5: Agent behavior and localization accuracy across geographic tiers. Models struggle significantly with less identifiable locations, with city guessing accuracies dropping below 26% on Tier 5. Interestingly, Llama 4 Scout and GPT-5 Mini maintain their country guessing accuracies across Tiers while Gemma 3 27B and Claude Haiku 4.5 show significant collapse down the tier list. Mean Haversine distance generally monotonously increases by tier other than Llama 4 Scout, which reverses the trend.

ultimately succeed, the median initial distance is 847 km (IQR: 312–1,892 km), compared to 2,341 km (IQR: 1,156–4,872 km) for failed rounds. A logistic regression predicting final city accuracy from initial distance yields a negative coefficient ($\beta = -0.43, p < 0.001$), confirming that exploration refines already-reasonable hypotheses instead of re-discovery. Our statistical tests (Appendix C) show that these results are significant.

4.2 Agent Behavior Analysis

To better understand the agentic behavior of the tested VLMs, we dissect performance by development tiers and recognizability tiers. Table 4 shows how the number of exploration actions, defined as a sum of moves and rotations made by the model, differs. Addressing RQ3, we note that GPT-5 Mini is quite efficient, completing its task in 4.20-4.95 steps out of 8. Llama 4 Scout shows higher utilization at 7.48-7.57 steps, but most of this remains concentrated to rotations across the spawn point instead of moving forward or backward. Overall, VLMs are significantly better at guessing developed region locations than developing alternatives (Figure 4). Notably, we find that models explore developed regions more than developing (5.95 vs. 5.65 actions), and reasoning length is largely architecture-dependent. When models do explore developing-region locations fully, they exhibit higher rotation-to-move ratios (2.1:1 vs. 1.4:1 in developed regions). While this difference is not directly noticed in task accuracy (on average, developed regions lead by 2.07%), Haversine distance deviates by 50-103%, representing substantial bias across the tested models. The largest deltas appear at the country level under agentic navigation: Gemini 3.0 Flash shows a -25.9% country accuracy gap and GPT-5 Mini a -5.7% gap, while Claude Haiku 4.5 and Llama 4 Scout show -3.0% and -14.7% respectively. Gemma 3 27B is a notable exception, with slightly higher city accuracy in developing regions across all three conditions (Figure 4). Our statistical results in Appendix C show that taking more screenshots helps up to a point, then plateaus or slightly hurts city-level precision.

4.3 Reasoning Vocabulary Analysis

Finally, we conduct TF-IDF sampling to identify which visual cues and terminology are the most frequently encountered in geospatial reasoning. Table 5 presents these descriptors per country. Finding more qualitative evidence for RQ2, we note that models focus on visible identification cues across tested nations, such as "gingko" for Japan, "tram tracks" and "shutters" for Switzerland, and the "Chao Phraya" river for Thailand. Most models reason with nationality-based terminology across the regions tested; however, a few noticeable patterns emerge. In Claude Haiku 4.5, we notice that "Japanese" is the leading term for Taiwan, and Gemma 3 27B leads with "German" for Switzerland. False positives cluster around visually similar regions: Southeast Asian countries are frequently confused (Thai locations misclassified as Vietnamese 23% of the time, and vice versa 18%). Notably, models often invoke the correct visual descriptors, such as "tropical vegetation," "weathered infrastructure," and "motorcycle traffic" for Southeast Asian countries, but map these to incorrect conclusions, suggesting a geographic grounding bottleneck rather than identification. The absence of country-specific markers for Mozambique and

Country	Claude Haiku 4.5	Gemini 3.0 Flash	Gemma 3 27B	Llama 4 Scout	GPT-5 mini
Canada	canadian, plains, north american, flat, midwest	canadian, canada, alberta, prairies, north american	american, flat, canada, flat landscape, midwest	north, north american, suburban, houses, suggest north	prairie, north american, canadian, american, north
China	cultural, park, asian, cultural venue, china	shanghai, bund, china, screen, shanghai china	japanese, garden, lanterns, japanese garden, japanese writing	geographic features, features signs, geographic, logo, google logo	chinese, taiwan, taipei, traditional chinese, traditional
India	indian, india, auto, tropical, market	india, delhi, indian, rickshaws, auto	india, hindi, indian, hindi signage, auto rickshaws	india, tuk, english, urban, script	indian, india, tropical, south asian, south
Japan	japanese, japanese residential, ginkgo, japan, tokyo	japanese, japan, tokyo, utility, nakano	japanese, autumn, japanese writing, retaining, concrete retaining	japanese, plaque, japanese characters, presence japanese, japan	ginkgo, ginkgo leaves, japanese, leaves, tokyo
Norway	nordic, norway, norwegian, road, scandinavian	norway, norwegian, oslo, sweden, birch	norway, scandinavian, road, sweden, norwegian	picket, european, white picket, picket fence, white	norway, norwegian, scandinavian, picket, birch
Switzerland	swiss, tunnel, alpine, european, roofs ivy	switzerland, bern, swiss, zurich, shutters	german, tram tracks, tram tracks, park	shutters, chinese, european, architecture shutters, chinese characters	swiss, park, switzerland, european, chinese
Taiwan	japanese, chinese, urban, taiwan, taiwanese	taipei, taiwan, traditional chinese, chinese, xinyi	chinese characters, chinese, characters, taiwan, scooters	chinese, chinese characters, characters, presence chinese, asian	chinese, taiwan, taipei, traditional chinese, traditional
Thailand	river, bridge, cable, mekong, stayed	bangkok, soi, thai, thailand, bridge	river, chao phraya, chao, phraya, phraya river	river, bridge, water, stayed bridge, cable	river, phraya, chao, chao phraya, bangkok
United States	american, midwest, road, flat, north american	midwest, residential, alaska, united states, american	american, lawns, maintained lawns, american cars, homes	suburban, north, north american, houses, road	north american, american, north, english, alaska

Table 5: Qualitative geographic descriptors produced by different vision-language models across countries. Models generally reason with nationality-specific terms.

Thailand (Table 5) correlates with their highest misclassification rates (67% and 54% country-level error, respectively). Models rely mainly on generic regional descriptors when they lack distinctive vocabulary. We believe targeted data augmentation with country-specific visual-textual pairs in the pretraining corpora could yield improvements for underperforming regions.

5 Discussion

First, active exploration can improve geolocation performance, but the improvements are highly conditional (**RQ1**). When agents begin with a reasonably good initial hypothesis (i.e., the initial guess is "in the right neighborhood"), additional navigation tends to refine that hypothesis and substantially reduce Haversine distance (see Figure 4 and Appendix B). However, for many trials where the initial guess is far off, extra exploration often fails to recover and sometimes reinforces incorrect priors. This suggests that VLMs are often better at confirmation than recovery, and established mistakes in context largely influence outcomes.

Second, our analyses uncover a marked and systematic bias favoring developed-region locations (**RQ2**). Although continent and country accuracies remain relatively high, mean Haversine distances and failure modes are substantially worse for developing-region samples (Table 4 and Figure 5). This bias manifests both in poorer distance-based outcomes and in the vocabulary models that rely on (Table 5). Third, reasoning verbosity is inversely correlated with success in our task (Appendix C). Longer explanations and heavier action usage are not necessarily signs of better decision-making.

Fourth, model architecture and pretraining regimes strongly affect how efficiently agents use the limited action budget (**RQ3**). Gemini 3.0 Flash and GPT-5 Mini achieved the best overall accuracies and the lowest median distances while using substantially fewer actions on average (Table 3, Table 4). In contrast, models such as Llama 4 Scout used many rotations but few forward moves. This suggests that good static-image geolocation performance does not directly translate into effective embodied exploration across frontier VLMs.

6 Conclusion

We introduce GeoAgent, a geolocation guessing task for vision-language models that require embodied agentic web exploration. Our analyses reveal several behavioral caveats in current-generation models that are overlooked by benchmarks focusing on static image-based recognition: lack of self-correction when initial guesses are incorrect, poor efficiency tuning in models such as Claude Haiku 4.5 and Llama 4 Scout, and a significant developed/developing bias. Through our experiments, we argue for improved pre-training for embodied geospatial navigation and location predictions with alternate evaluation angles not studied in previous literature.

Limitations

There are several limitations to acknowledge. First, we limited agents to eight exploration steps to balance thoroughness with computational costs, though different action budgets may reveal different behavioral patterns in the agents, and some initial errors may be recoverable with extended exploration. Extended action limits (beyond 8) could yield in stronger statistical validity, but remain infeasible given current budgets, given diminishing returns beyond action 5 (Figure 4 and Appendix B). Second, our agent's action space (nine actions) may not capture the full range of human navigation strategies. Third, factors such as Street View image quality and coverage density have not been studied. We omit human baselines as prior work on GeoGuessr-like tasks has already established poor human performance on geolocalization. While our sample of 1,200 locations may not fully represent global diversity, we emphasize that GeoAgent is designed to evaluate agentic embodied navigation rather than serve as a comprehensive geographic benchmark.

Ethical Considerations

Geolocalization systems like GeoAgent raise legitimate concerns around privacy and potential misuse, like surveillance applications and unauthorized location inference from personal media. GeoAgent does not process user-generated content and operates exclusively on publicly available Google Street View imagery, which blurs all privately identifiable information, such as license plates and faces. However, we acknowledge that advances in agentic geolocalization could lower the barrier for misuse if applied to unconstrained image sources. Additionally, our dataset inherits the well-documented geographic coverage bias of Street View, which disproportionately represents developed regions with higher-quality imagery, potentially amplifying performance disparities across nations. We encourage future work to incorporate explicit misuse safeguards, such as restricted API access and geofencing, alongside more geographically balanced data collection efforts.

References

Fenghua Cheng, Jinxiang Wang, Sen Wang, Zi Huang, and Xue Li. 2025. Geoguess: Multimodal reasoning based on hierarchy of visual information in street view. *Preprint*, arXiv:2506.16633.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Preprint*, arXiv:2306.06070.

Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. 2025. A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp). *ArXiv*, abs/2505.02279.

Lutfi Eren Erdogan, Nicholas Lee, Siddharth Jha, Sehoon Kim, Ryan Tabrizi, Suhong Moon, Coleman Richard Charles Hooper, Gopala Anumanchipalli, Kurt Keutzer, and Amir Gholami. 2024. TinyAgent: Function calling at the edge. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 80–88, Miami, Florida, USA. Association for Computational Linguistics.

Izzeddin Gur, Hiroki Furuta, Austin V Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2024. A real-world webagent with planning, long context understanding, and program synthesis. In *The Twelfth International Conference on Learning Representations*.

Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. 2024. Pigeon: Predicting image geolocations. *Preprint*, arXiv:2307.05845.

Jingyuan Huang, Jen-tse Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. 2025. AI sees your Location—But with a bias toward the wealthy world. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18019–18039, Suzhou, China. Association for Computational Linguistics.

Neel Jay, Hieu Minh Jord Nguyen, Trung-Dung Hoang, and Jacob Haimes. 2025. Evaluating precise geolocation inference capabilities of vision language models. *ArXiv*, abs/2502.14412.

Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. 2025a. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. *Preprint*, arXiv:2506.14674.

Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. 2025b. From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models. *Preprint*, arXiv:2508.01608.

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 *Forty-second International Conference on Machine Learning*.

Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Uttama Nambi. 2025. Agentic reasoning and tool integration for llms via reinforcement learning. *ArXiv*, abs/2505.01441.

Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet, Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and Xiuying Chen. 2026. Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework. *Preprint*, arXiv:2502.13759.

Sudharshan Suresh, Nathaniel Chodosh, and Montiel Abello. 2018. Deepgeo: Photo localization with deep neural network. *Preprint*, arXiv:1810.03077.

Mohit Talreja, Joshua Diao, Jim Thannikary James, Radu Casapu, Tejas Santanam, Ethan Mendes, Alan Ritter, Wei Xu, and James Hays. 2026. Georc: A benchmark for geolocation reasoning chains. *Preprint*, arXiv:2601.21278.

Shruti Vyas, Chen Chen, and Mubarak Shah. 2022. Gama: Cross-view video geo-localization. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, page 440–456, Berlin, Heidelberg. Springer-Verlag.

Albatool Wazzan, Stephen MacNeil, and Richard Souvenir. 2024. Comparing traditional and llm-based search for image geolocation. *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. Webwalker: Benchmarking llms in web traversal. *ArXiv*, abs/2501.07572.

Chenhui Xu, Fuxun Yu, Michael J. Bianco, Jacob Kovarskiy, Raphael Tang, Qi Zhang, Zirui Xu, Will LeVine, Brandon Dubbs, Heming Liao, Cassandra Burgess, Suvam Bag, Jay Patravali, Rupanjali Kukal, Mikael Figueroa, Rishi Madhok, Nikolaos Karianakis, and Jinjun Xiong. 2025. Geo-r1: Unlocking vlm geospatial reasoning with cross-view reinforcement learning. *Preprint*, arXiv:2510.00072.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Qiang Yi and Lianlei Shan. 2025. Geolocsft: Efficient visual geolocation via supervised fine-tuning of multimodal foundation models. *Preprint*, arXiv:2506.01277.

Zheyuan Zhang, Runze Li, Tasnim Kabir, and Jordan Boyd-Graber. 2025. Navig: Natural language-guided analysis with vision language models for image geolocalization. *Preprint*, arXiv:2502.14638.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. 2024. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *The Twelfth International Conference on Learning Representations*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Appendices

This supplementary material presents additional details on the following aspects:

- **Appendix A:** Prompts
- **Appendix B:** Distance Convergence Across Actions
- **Appendix C:** Action-Outcome Correlation Analysis

A Prompts

Figure 6 presents the prompt template used across all evaluated VLMs. The prompt structure consists of four components: (1) task framing, establishing the GeoGuessr context, (2) dynamic state information including round number, remaining actions, and current heading, (3) contextual memory of up to five prior observations from the current round, and (4) the available action space.

We enforce structured JSON output to ensure consistent parsing across model architectures. Each response requires five fields: observations (salient visual cues), reasoning (geographic deduction), confidence level (four-point scale), selected action, and current location guess with city and country. The Google Geocoding API is then used to deduce the latitude and longitude for the guessed location to avoid any LLM-induced hallucination in coordinate calculations. Models are encouraged to keep the Observations and reasoning fields brief, as verbose outputs correlate negatively with task performance (see Appendix C).

GeoGuessr Agent Prompt

You are playing *GeoGuessr*. Guess the location from the image.

State: Round `{round_num}` | Actions left `{max_actions - count}` | Heading `{heading}`°

Context: Up to the last five observations from the current round, if available.

Available actions: `ROTATE_LEFT`, `ROTATE_RIGHT`, `ROTATE_BEHIND`, `MOVE_FORWARD`, `RETURN_START`, `GUESS`

Response format (JSON only):

```
{ "observations": "Brief key clues",
"reasoning": "Short deduction",
"confidence":
"low|medium|high|very_high",
"action": "ROTATE_*|MOVE_*|GUESS",
"guess": { "location": "Place, City,
Country", "country": "Country or null",
"city": "City or null" } }
```

Observations and reasoning are limited to 100 words each.

Figure 6: Prompt used by the GeoGuessr agent for location inference and action selection.

B Distance Convergence Across Actions

Figure 7 presents the mean Haversine distance to ground truth across sequential actions, stratified by final outcome correctness. For rounds culminating in correct city guesses (left panel), all models demonstrate monotonic convergence toward the target location, with GPT-5 Mini achieving sub-500 km distances by action 3. Notably, Llama 4 Scout and Claude Haiku 4.5 exhibit high initial uncertainty (>1500 km), but only the former converges effectively given sufficient exploration budget.

The right panel reveals that when final guesses are incorrect, models show minimal distance improvement regardless of exploration effort. Claude Haiku 4.5 and Gemma 3 27B maintain relatively stable (but incorrect) estimates, while GPT-5 Mini diverges after action 5. This asymmetry provides evidence for confirmation bias in sequential geospatial reasoning, where models struggle to abandon initial misconceptions even when presented with contradictory visual evidence.

C Action-Outcome Correlation Analysis

Table 6 reports Pearson and Spearman correlations between agent behavioral features and task outcomes across all model runs. We observe statistically significant positive correlations between reasoning length and all accuracy measures (continent: $\rho = 0.140$, country: $\rho = 0.137$, city: $\rho = 0.090$), alongside a negative correlation with distance ($\rho = -0.142$), suggesting that longer reasoning traces are associated with more accurate geolocalization rather than mere verbosity. Reasoning length shows the strongest individual association with points ($\rho = 0.241$), reinforcing that deliberate multi-step analysis contributes to task performance.

Total actions and screenshots taken also correlate positively with continent and country accuracy ($\rho = 0.124$ and $\rho = 0.118$ for total actions), though the relationship weakens considerably at the city level (Pearson $r = -0.009$, Spearman $\rho = 0.055$), indicating that additional exploration aids broad localization but offers diminishing returns for fine-grained disambiguation. Move actions show a consistent positive association across

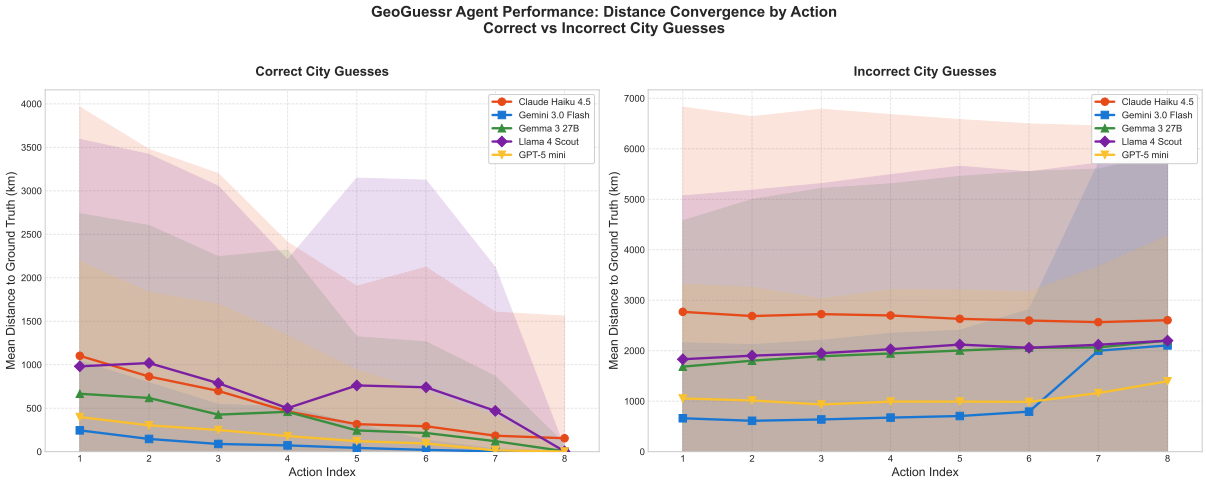

Figure 7: Aggregated mean distance improvement over actions per model.

all outcomes, with the strongest link to points ($\rho = 0.195$) and distance ($\rho = -0.108$), suggesting that forward movement is a more productive exploration strategy than rotation. This aligns with our earlier finding that Llama 4 Scout's rotation-heavy policy (6.14 rotations vs. 0.41 moves per round) underperforms relative to Gemini Flash 3.0's movement-dominant strategy (3.57 moves vs. 0.39 rotations). All correlations reject the null hypothesis at $\alpha = 0.05$, supporting the finding that increased exploration effort correlates with improved geospatial reasoning, though effect sizes remain modest ($|\rho| < 0.25$), indicating that action quality continues to matter more than action quantity.

D Multiview Baseline Implementation

Figure represents a sample from the dataset curated for evaluation of the VLM for the Multiview-Baseline. The multiview baseline is a non-agentic pipeline in which the VLM is provided with a single, static composite image for each location, and performs exactly one inference, without interactive navigation or iterative reasoning.

Composite Capture. We use a fully automated script, producing deterministic composite images for all 1200 locations. For each location, the street view panorama is first positioned to the target coordinates. Four screenshots are captured at offsets of 0°, 90°, 180°, and 270° from the initial heading. To capture the 8 view composite, the system additionally performs a `MOVE_FORWARD` action by selecting the navigation link with the closest heading alignment, after which four more screenshots are taken with identical offsets. The eight views are arranged into a labeled 4×2 grid composite (2560×776 px), with original spawn-point views occupying the top row and post-move views the bottom. A 4-view spawn-only composite is produced similarly. Both are saved as JPEG and accompanied by a manifest recording per-location metadata including coordinates, development status, and recognizability tier.

Offline Inference. The pre-captured composite is passed to each model alongside the prompt in Figure 8. The prompt omits state and action fields. All other prompt structure, output format requirements, temperature settings, and geocoding procedures are kept identical to the agentic condition. This ensures that performance differences between the agentic and multiview conditions are attributable to embodied navigation rather than prompt or scoring asymmetries.

Multiview Baseline Prompt

You are playing *GeoGuessr*. Guess the location from the image.

State: Round {N}

Response format (JSON only):

```
{ "observations": "Brief key clues,
"reasoning": "Short deduction",
"confidence":
"low|medium|high|very_high",
"guess": { "location": "Specific
place, City, Country", "country":
"Country or null", "city": "City or
null" } }
```

Observations and reasoning are limited to 100 words each.

Figure 8: Prompt used by the multiview baseline models for single-pass location inference. Unlike the agentic prompt (Figure 6), no action selection, remaining action count, or heading state is provided, reflecting the non-interactive nature of the baseline.

Figure 9: Examples of 4-view and 8-view composite images used in the multiview baseline. The top strip shows four spawn-point screenshots at cardinal headings (4-view). The bottom two rows show the full 8-view composite, with spawn-point views (blue labels) in the first row and post-`MOVE_FORWARD` views (green labels) in the second row, each captured at identical heading offsets.

Feature	Outcome	Pearson r	Pearson p	Spearman ρ	Rejects Null Hypothesis (H_0)?
Reasoning Length	Continent Correct	0.129	0.0000	0.140	YES
	Country Correct	0.119	0.0000	0.137	YES
	City Correct	0.070	0.0000	0.090	YES
	Distance Km	−0.130	0.0000	−0.142	YES
	Points	0.181	0.0000	0.241	YES
Screenshots Taken	Continent Correct	0.097	0.0000	0.128	YES
	Country Correct	0.081	0.0000	0.121	YES
	City Correct	−0.009	0.0213	0.057	YES
	Distance Km	−0.085	0.0000	−0.102	YES
	Points	0.121	0.0000	0.216	YES
Total Actions	Continent Correct	0.096	0.0000	0.124	YES
	Country Correct	0.081	0.0000	0.118	YES
	City Correct	−0.009	0.0167	0.055	YES
	Distance Km	−0.085	0.0000	−0.108	YES
	Points	0.120	0.0000	0.211	YES
Move Actions	Continent Correct	0.099	0.0000	0.122	YES
	Country Correct	0.086	0.0000	0.114	YES
	City Correct	0.014	0.0004	0.047	YES
	Distance Km	−0.099	0.0000	−0.108	YES
	Points	0.140	0.0000	0.195	YES

Table 6: Correlation analysis between agent behavior features and task outcomes. Pearson and Spearman correlation coefficients are reported along with hypothesis test results.

Model	Dev. Level	# Actions	# Moves	# Rotations	Mean Reas. Len.	Accuracy (%)			Mean Dist. (km)
						Continent	Country	City	
Random Baseline	Developed	-	-	-	-	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	9982.18 ± 306.7
	Developing	-	-	-	-	0.00 ± 0.0	0.00 ± 0.0	0.00 ± 0.0	10007.06 ± 255.4
Claude Haiku 4.5	Developed	6.29 ± 0.1	3.15 ± 0.1	2.04 ± 0.1	434.7 ± 7.3	68.23 ± 6.9	53.22 ± 4.4	17.93 ± 1.7	1782.64 ± 263.2
	Developing	5.40 ± 0.0	2.68 ± 0.1	1.64 ± 0.0	415.4 ± 8.9	61.53 ± 1.4	43.92 ± 4.9	14.08 ± 1.4	3712.72 ± 410.9
Gemma 3 27B	Developed	6.92 ± 0.1	2.70 ± 0.2	3.22 ± 0.2	220.0 ± 2.0	45.94 ± 2.7	36.35 ± 1.1	22.81 ± 0.5	1307.74 ± 315.4
	Developing	6.76 ± 0.0	3.89 ± 0.1	1.87 ± 0.1	221.7 ± 0.9	71.18 ± 1.8	56.56 ± 1.5	17.92 ± 3.1	2536.77 ± 114.8
Llama 4 Scout	Developed	7.57 ± 0.1	0.20 ± 0.0	6.37 ± 0.1	197.0 ± 3.1	64.53 ± 2.0	49.07 ± 2.5	14.75 ± 2.1	1622.49 ± 107.6
	Developing	7.48 ± 0.1	0.44 ± 0.0	6.04 ± 0.1	205.4 ± 1.7	73.89 ± 3.3	61.06 ± 1.7	14.69 ± 2.5	2472.46 ± 363.1
GPT-5 mini	Developed	4.95 ± 0.5	2.46 ± 0.4	1.49 ± 0.1	183.4 ± 1.4	91.69 ± 2.0	83.72 ± 2.1	34.42 ± 2.4	645.95 ± 163.7
	Developing	4.20 ± 0.3	2.14 ± 0.3	1.07 ± 0.1	182.0 ± 1.9	84.62 ± 1.0	76.44 ± 1.8	34.93 ± 1.1	1316.75 ± 291.6

Table 7: Agent behavior and accuracy breakdown across developed and developing regions. Models generally make similar exploration actions across both region types, but the final outcomes vary widely. Most models secure higher continent, country, and city guessing accuracies, while Gemma 3 27B reverses this trend. Importantly, all tested VLMs show acute developed region bias in mean Haversine distance, resulting in 50-103% better numbers than the equivalent developing region result (lower is better).

Country	Top Words
Argentina	spanish, andean, colombian, colonial, white, argentine, argentina, altitude, high altitude, urban
Australia	australian, modern, australia, melbourne, red, brick, white, tropical, trees, vegetation
Bangladesh	bengali, tropical, india, indian, vegetation, bengali script, bangladesh, tropical vegetation, kerala, script
Brazil	brazilian, brazil, tropical, portuguese, vegetation, tropical vegetation, modern, urban, portuguese signage, portuguese language
Cambodia	cambodia, rural, tropical, vegetation, red, soil, tropical vegetation, laterite, red laterite, laos
Canada	canadian, north american, american, prairie, canada, rural, midwest, trees, flat, road
Chile	arid, spanish, andean, peru, terrain, colonial, mexican, vegetation, palm, peruvian
China	elevated pedestrian, highway, modern, elevated, wet, overpass, highway overpass, urban, lighting wet, lighting
Egypt	orthodox, pollution, chinese, egyptian, georgia, caucasus, tbilisi, georgian, religious, fence religious
Estonia	european, birch, rural, nordic, road, wooden, russia, birch trees, baltic, trees
Eswatini	rural, vegetation, soil, road, red, landscape, highland, brazil, rwanda, terrain
Fiji	tropical, palm, palm trees, vegetation, white, road, island, trees, rural, caribbean
Finland	finland, nordic, finnish, birch, forest, birch trees, sweden, road, snow, finland sweden
France	paris, french, haussmann, parisian, shopping, haussmann style, arcade, pedestrian shopping, modern, shopping street
Germany	german, european, urban, berlin, modern, germany, brick, courtyard, radiator, german city
India	indian, india, hindi, tropical, typical indian, devanagari, auto, urban, devanagari script, script
Japan	japanese, japan, japanese residential, modern, typical japanese, distinctly japanese, suburban, urban, japanese suburban, modern japanese
Norway	nordic, road, birch, finland, sweden, norwegian, scandinavian, white, german, norway
Switzerland	swiss, german, alpine, european, switzerland, modern, green shutters, green, french, tunnel
Taiwan	chinese, taiwan, urban, chinese characters, taiwanese, tropical, characters, taipei, modern, chinese city
Thailand	thai, monument, thai script, tropical, temple, thailand, bangkok, tropical vegetation, vegetation, alley
United Kingdom	scottish, uk, london, scotland, british, victorian, modern, white, rendered, red
United States	midwest, american, road, landscape, terrain, rural, trees, flat, vegetation, suburban
Vietnam	vietnamese, vietnam, tropical, chi minh, minh, minh city, ho chi, ho, chi, vietnamese city

Table 8: TF-IDF qualitative geographic descriptors for Claude-Agent.

Country	Top Words
Argentina	spanish, spanish speaking, presence spanish, spanish text, latin american, latin, south, american, speaking country, speaking
Australia	australian, english, palm, palm trees, australia, modern, city, skyline, city skyline, cars
Bangladesh	bengali, bangladesh, script, bengali script, tuk, tropical, rickshaws, presence bengali, south asian, palm
Brazil	portuguese, portuguese speaking, brazil, presence portuguese, cars, portuguese text, latin american, speaking, speaking country, spanish
Cambodia	dirt, khmer, tropical, cambodia, dirt road, khmer script, southeast, shows dirt, wall, dirt path
Canada	north, north american, typical north, american, suburban, canada, suggest north, houses, vehicles, cars
China	walkway, walkway overpass, overpass, wet, chinese, characters, china, chinese characters, city, metal railing
Egypt	arabic, arabic script, desert, desert like, like environment, script, presence arabic, middle eastern, middle, straight
France	french, france, european, likely france, presence french, french signage, buildings french, claire, city, modern
India	india, tuk, hindi, script, english, indian, devanagari, tuks, tuk tuks, likely india
Japan	japanese, japanese characters, characters, japan, presence japanese, likely japan, characters sign, characters signs, location japan, japan urban
Norway	european, norwegian, houses, nordic, norway, red, european style, suburban, rural, brick
Switzerland	european, german, european style, green shutters, green, shutters, european city, tram, green dome, city
Taiwan	chinese, chinese characters, characters, presence chinese, taiwan, scooters, asian, characters signs, east asian, chinese speaking
Thailand	thai, murky, river, thailand, brown murky, murky water, thai script, water, river brown, presence thai
United Kingdom	uk, houses, british style, british, european, semi detached, detached, brick, detached houses, suburban
United States	north, suburban, north american, typical north, american, houses, states, united states, united, typical
Vietnam	vietnamese, vietnam, motorbikes, presence vietnamese, likely vietnam, southeast, vietnamese text, motorcycles, asian, vietnamese characters

Table 9: TF-IDF qualitative geographic descriptors for Groq-Agent.

Country	Top Words
Argentina	spanish, argentina, argentine, buenos aires, buenos, aires, ushuaia, spanish language, uruguay, american
Australia	australian, melbourne, australia, sydney, laneway, australian style, english, darwin, tropical, inner
Bangladesh	bengali, bangladesh, bengali script, tropical, dhaka, rickshaws, rickshaw, script, cycle, india
Brazil	portuguese, brazilian, brazil, tropical, portuguese language, brazilian style, indicate brazil, portuguese signage, curitiba, brazilian city
Canada	prairie, canadian, north american, american, north, canada, saskatchewan, alberta, canadian prairie, flat prairie
France	paris, french, central paris, france, parisian, haussmann, haussmann style, lib, french language, le
Germany	german, berlin, european, germany, bahn, german language, eu, brick, beer garden, courtyard
India	india, indian, devanagari, hindi, tropical, south asian, rickshaw, auto, indian city, english
Japan	japanese, tokyo, kei, japan, residential, japanese text, utility, japanese style, typical japanese, indicate japan
Norway	norway, norwegian, birch, scandinavian, sweden, norway sweden, wooden, scandinavia, nordic, oslo
Switzerland	swiss, switzerland, alpine, european, zurich, german, swiss style, parking, french, shutters
Taiwan	taiwan, taipei, chinese, traditional chinese, traditional, characters, scooters, chinese characters, taiwanese, taiwan urban
Thailand	thai, bangkok, soi, thailand, boat, bridge, thai script, river, tropical, stayed bridge
United Kingdom	uk, scotland, uk style, dash, pebble dash, pebble, london, central london, scottish, georgian
United States	midwestern, american, north american, north, united states, english, deciduous, ranch, united, town
Vietnam	vietnamese, vietnam, minh, chi minh, chi, ho chi, hanoi, ho, minh city, motorbikes

Table 10: TF-IDF qualitative geographic descriptors for OpenAI-Agent.