Perceive to Hypothesize, Verify to Ground: an Agentic Reasoning Framework for Open-World Geo-Localization

Anonymous ACL submission

Abstract

Open-world geo-localization requires models to reason over ambiguous visual cues through multi-step reasoning and external knowledge grounding. While recent large vision-language models exhibit strong multimodal reasoning capabilities, existing approaches still suffer from perceptual hallucination and context drift due to the lack of explicit evidence-grounded verification. In this work, we reformulate geo-localization as a human-like **perceive-then-verify** reasoning problem and propose **GeoPAVE** (Geo Perception-and-Verification-Engine), a bi-level agentic framework that contains perception-based hypothesis generation via single-pass rollouts and verification-based evidence grounding for decision actions: *support*, *refute*, and *refine*. To support rigorous evaluation, we further introduce **PAVED**, a novel dataset derived from real-world user check-in data, equipped with comprehensive reasoning trajectories featuring multi-hop queries, multi-round tool invocations, and structured perception-verification traces. Extensive experiments verify the effectiveness of our method across multiple real-world datasets. The dataset and code are available at `https://anonymous.4open.science/r/GeoPAVE-E390/`.

1 Introduction

Open-world geo-localization aims to infer the geographic coordinates of a street-view image around the world, without relying on a retrieval gallery. Therefore, this task requests models to integrate fine-grained visual evidence with broad geographic, cultural, and semantic knowledge (Yang et al., 2026; Jia et al., 2025). This moves past simple single-shot perception, requiring iterative reasoning, multi-hop retrieval, and fact verification with open-world exploration over external knowledge sources (Cai et al., 2026; Lymperaiou et al., 2026). Geo-localization supports a wide range of

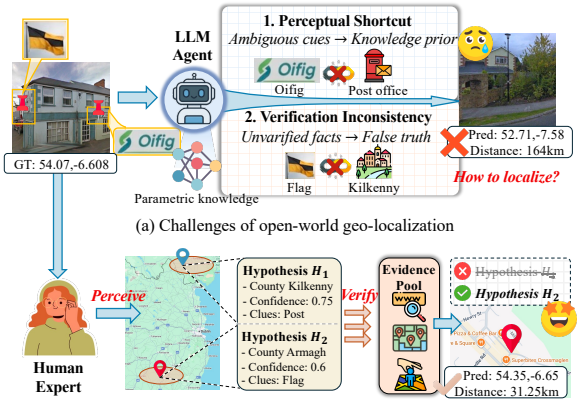

Figure 1: Challenges of geo-localization.

real-world applications, such as autonomous driving, criminal investigation (Campos et al., 2025), and an important testbed for open-world multimodal reasoning and knowledge-grounded intelligence (Geng et al., 2026).

Traditional geo-localization research has focused on visual perception through classification approach (Weyand et al., 2016; Pramanick et al., 2022) and feature retrieval (Vivanco Cepeda et al., 2023; Haas et al., 2024). However, their effectiveness largely depends on large-scale, predefined, and closed-world reference datasets. With the advancement of deep learning, recent work has begun to leverage large language models equipped with world knowledge to reason over visual cues. By enhancing visual reasoning abilities through supervised fine-tuning (Li et al., 2024; Wang et al., 2025a) or reinforcement learning (Wang et al., 2025b; Jia et al., 2026; Liu et al., 2026), geo-localization is gradually shifting from pure pattern recognition toward perception-driven reasoning.

Though promising, recent LVLM post-training paradigms (Wang et al., 2025b; Li et al., 2025) focus on extending reasoning depth, yet they still struggle to maintain reliable reasoning when per-

ceptual evidence is ambiguous and intermediate reasoning steps lack reliable auditing. These limitations reveal two core challenges (see Figure 1) inherent to perception-driven reasoning:

1) Perceptual Shortcut: When visual evidence is partial or ambiguous, LVLMs maps tend to map salient local cues to to high-frequency geographic regions based on blurred text or architectural styles. Although test-time scaling techniques, such as parallel rollouts (Ji et al., 2026), expand the search space, they increase inference costs without fundamentally resolving the inherent unreliability of perception-driven reasoning in single pass (Tran and Kiela, 2026).

2) Verification Inconsistency: Existing geo-localization research largely overlooks the importance of evidence-grounded verification. While agentic RL has demonstrated the utility of external tools, unconstrained tool invocation in open environments may introduces lengthy contexts that mix useful evidence with irrelevant retrieval results, undermining the consistency between initial visual cues and final conclusions. For example, when a blurred storefront retrieves similar POIs from multiple cities, the model may follow an unverified search result and converge to a plausible but visually inconsistent location.

To tackle these issues, we revisit the intrinsic nature of ambiguous reasoning in open-world geo-localization. We argue that the key lies in *simulating human-like cognitive behavior*. In geographic discovery games such as GeoGuessr (Talreja et al., 2026; Cheng et al., 2025), expert players rarely rely on a single-shot intuition, but instead operate in a continuous **perceive-then-verify** loop. They first perceive multi-granularity visual cues to form and mentally rank a set of plausible hypotheses, continuously cross-checking these hypotheses against external geographic and map-based context to confirm, reject, or refine their beliefs.

Inspired by the human-like reasoning process, we first reformulate geo-localization as a perception-verification paradigm, and then propose **GeoPAVE**, a bi-level agentic framework that decomposes geo-localization into human-like perception and verification stages. Our GeoPAVE consists of 1) an *Agentic Perception Module* that generates uncertainty-aware hypotheses within a single reasoning pass, while leveraging uncertainty quantification to determine verification actions; and 2) a *Verification Module* that grounds candidate hypotheses through strategic tool invocation, enabling evidence-based hypothesis refinement. By adaptively adjusting hypotheses with verifiable external evidence GeoPAVE effectively suppresses hallucinations while preserving contextual consistency. To facilitate rigorous evaluation, we construct **PAVED**, a high-quality dataset sourced from human check-ins, containing rich tool-use evidence-based hypothesis verification reasoning trajectories generated from GeoPAVE. Our contributions can be summarized as follows:

- We revisit open-world geo-localization based on human cognitive reasoning, reformulating it as a **perceive-then-verify** paradigm and proposing a perception–verification agentic reasoning framework **GeoPAVE**. This framework naturally produces high-quality reasoning trajectories with rich evidence-based verification.
- We introduce an agentic perception module that condenses candidate-space exploration into a single reasoning and uncertainty estimation, together with a structured verification module that maps uncertainty signals to specialized tool-based verification skills, performing tool-augmented evidence retrieval and iterative hypothesis refinement.
- Extensive experiments across multiple benchmarks demonstrate that GeoPAVE achieves strong and competitive performance against representative retrieval-based, agentic, and closed-source LVLM baselines. In addition, we present PAVED (Perception-and-Verification-Engine-Dataset), comprising streetview images collected from human check-ins, paired with verification procedures and toolcall trajectories.

2 Preliminaries

2.1 Problem Statement

Given an image $\mathcal{I}$, geo-localization aims to predict its coordinate $\hat{\mathbf{p}} = (\text{lat}, \text{lon})$. We formulate it as a **perception-verification problem**: a vision-language model first produces candidate hypotheses $\mathcal{H} = \{h_i\}_{i=1}^{k}$ with confidence scores and uncertainty tags, while external tools retrieve evidence $\mathcal{E} = \{\mathbf{e}_j\}_{j=1}^{m}$ to refine them. The final coordinate is selected by evidence-conditioned scoring over the refined hypotheses.

2.2 Related Works

Open-World Geo-Localization. Open-World geo-localization aims to predict the precise geographic location of an image across the globe.

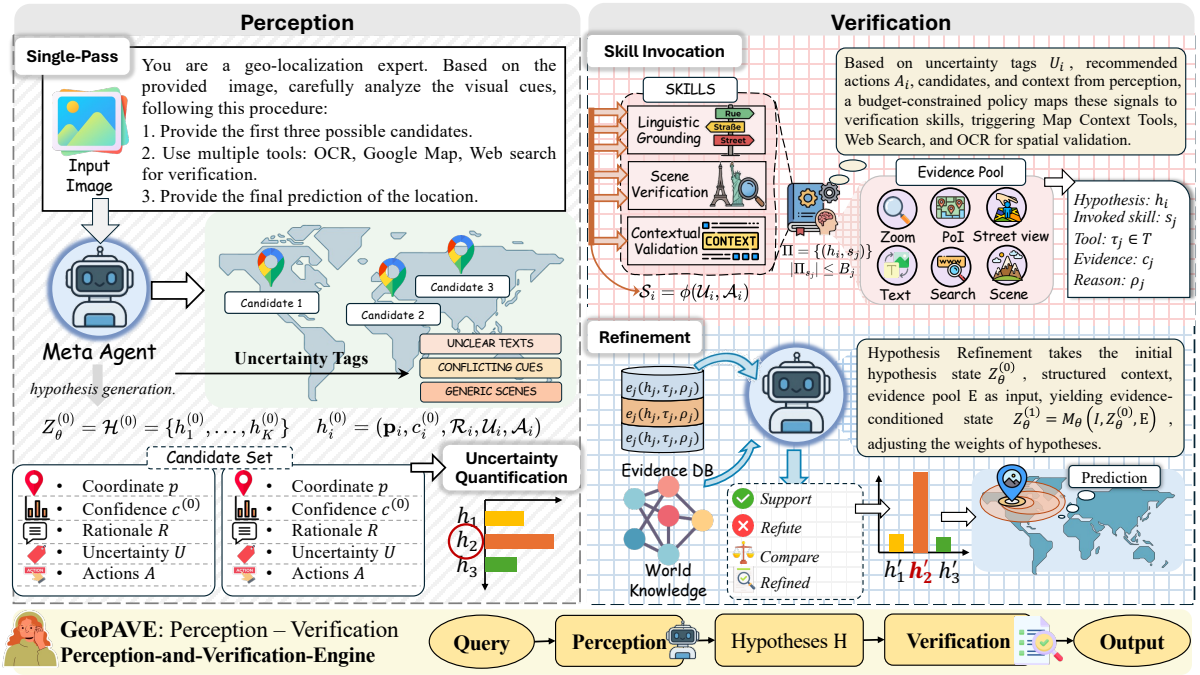

Figure 2: Overview of GeoPAVE.

Traditional computer vision methods tackle this through classification by discretizing the Earth into predefined cells (Seo et al., 2018; Pramanick et al., 2022; Ghasemi et al., 2025) or retrieval by matching queries against large-scale visual databases (Shatwell et al., 2025; Haas et al., 2024; Jia et al., 2024). However, these closed-world approaches heavily rely on massive datasets and struggle to generalize to open-world conditions. Recently, large vision-language models (LVLMs) have addressed this limitation by leveraging multimodal reasoning and extensive world knowledge to directly infer locations (Li et al., 2024; Yu et al., 2026; Li et al., 2025). While their explicit reasoning improves interpretability, the absence of verification mechanisms limits the reliability of these reasoning outcomes.

Agentic Reasoning System. Recent progress in large language models (LLMs) has been driven by integrating external tools, enabling agentic multimodal systems to ground perception through environment interactions and extend beyond parametric knowledge (Schick et al., 2023; Khaki et al., 2026; Qin et al., 2024). To optimize tool use and long-horizon decision-making, further studies explore agentic reinforcement learning (Li et al., 2026; Qin et al., 2025) and test-time scaling via parallel or multi-agent exploration (Zhu et al., 2025; Qiao et al., 2025). In the geo-localization domain, researchers employ multi-agent debate (Han et al., 2025a; Zheng et al., 2025) and agentic RL (Ji et al., 2026; Jia et al., 2026) to improve reasoning depth. However, these methods demand expensive computational resources, as their performance gains heavily rely on increased reasoning trajectories. In contrast, our proposed **GeoPAVE** introduces structured bi-level perception-verification stage, achieving higher reliability.

3 Methodology

3.1 Overview

GeoPAVE constructs a bi-level verification architecture to address the challenges of ambiguous visual reasoning in geo-localization, which decouples hypothesis generation from evidence-based validation. GeoPAVE grounds predictions in verifiable external evidence through structured tool invocation, providing a **hypothesis-verification** process: a perception Module that performs single-pass perception to explore the hypothesis space, and a verification module that implements non-parametric refinement through cross-modal evidence aggregation from heterogeneous knowledge sources. The pipeline can be shown in Figure 2.

3.2 Perception Module

Single-Pass Candidate Generation. For input image $\mathcal{I}$, the perception agent generates a first-pass

hypothesis state $Z_\theta^{(0)} = \mathcal{H}^{(0)} = \{h_1^{(0)}, \ldots, h_K^{(0)}\}$, where $K \in \{1, 2, 3\}$. This process produces multiple potential outcomes in a single inference pass. Each structured hypothesis $h_i^{(0)} = (\mathbf{p}_i, c_i^{(0)}, \mathcal{R}_i, \mathcal{U}_i, \mathcal{A}_i)$ consists of a predicted coordinate $\mathbf{p}_i = (\text{lat}_i, \text{lon}_i)$, an initial confidence score $c_i^{(0)} \in [0, 1]$, a natural language rationale $\mathcal{R}_i$, uncertainty tags $\mathcal{U}_i \subseteq \mathcal{U}$ from a predefined vocabulary, and verification actions and tools $\mathcal{A}_i$.

To enable reliable test-time scaling, we model uncertainty as structured tags that guide hypothesis verification. Given an initial hypothesis confidence $c_i^{(0)}$, uncertainty signals $\mathcal{U}_i$ derived from conflicting cues trigger confidence calibration $c_i^{(0)} \leftarrow f(c_i^{(0)}, \mathcal{U}_i)$ suppressing overconfident predictions before verification.

Uncertainty Quantification. GeoPAVE represents uncertainty as explicit routing signals for tool-call process. If visual cues, OCR-like text snippets, or candidate semantics conflict, the perception module lowers the adjusted confidence and records the unresolved ambiguity through fields such as `CONFLICTING_CUES` and uncertainty tags $\mathcal{U}$. This structured context describe image-internal observations and verification needs, which indicates that visual features and textual semantics diverge) to mitigate overconfident predictions on ambiguous data (detailed information in Appendix D.2). The tag set activates verification skills through a mapping $\phi : \mathcal{U} \to \mathcal{S}$ in the verification stage, without skill-based learning policies.

3.3 Verification Module

Conditional Skill Invocation. GeoPAVE employs formulates verification as conditional skill invocation, mapping epistemic signals to verification operations. Specifically, a verification skill $s \in \mathcal{S}$ is defined as a tuple $s = (\mathcal{U}_s, t_s, \psi_s)$, where $\mathcal{U}_s$ denotes uncertainty tags that activate the skill, $t_s \in \mathcal{T}$ denotes the external knowledge source, and $\psi_s : \mathcal{E} \times \mathcal{H} \to \mathbb{R}$ quantifies the mutual information $I(\mathbf{p}; \mathbf{e} \mid \mathcal{I}, \mathcal{E}_t)$ between the evidence item $\mathbf{e}$ and the true location $\mathbf{p}$ conditioned on the image and prior evidence, where $\mathcal{H}$ denotes the hypothesis set and $\mathcal{E}$ denotes the evidence pool. GeoPAVE instantiates three core skills:

1) *Linguistic Grounding* extracts script and language features from street signs to produce geographic constraints.
2) *Scene Verification* queries search engines with visual scenes to retrieve metadata.
3) *Contextual Validation* checks candidate locations against elevation, waterbody, topographic, and nearby-POI evidence from geographic databases to assess spatial consistency.

Figure 3: PAVED trace construction pipeline.

The skill selection policy $\pi : \mathbf{s}_t \to \mathcal{S}$ is rule-based, mapping uncertainty tags to verification skills and thereby enabling adaptive verification without policy learning. Each invoked skill produces a structured evidence record $\mathbf{e}_j = (i_j, \tau_j, o_j, y_j, \rho_j, \Delta_j, v_j)$, where i_j indexes the target hypothesis in $\mathcal{H}$, $\tau_j \in \mathcal{T}$ denotes the evidence source, o_j is the tool observation or retrieved content, ρ_j provides a compact reason code, $y_j \in \{\textsc{Support}, \textsc{Refute}, \textsc{Refine}\}$ denotes the verification decision, Δ_j is the rule-based score update induced by this evidence, and $v_j \in \{0, 1\}$ indicates whether the evidence triggers a hard veto. The outputs of all invoked skills form the evidence pool $\mathcal{E} = \{\mathbf{e}_j\}_{j=1}^m$, which serves as the grounded interface between tool execution and hypothesis refinement.

Specifically, SUPPORT increases the confidence of a hypothesis when the retrieved evidence is consistent with the candidate location, REFUTE decreases the confidence or rejects a hypothesis when severe geographic contradictions are observed, and REFINE updates the candidate location or its supporting context when newly discovered textual or spatial evidence can be successfully grounded.

Hypothesis Refinement. After evidence collection, GeoPAVE performs a second reasoning pass to reassess the initial hypotheses under the grounded evidence pool $\mathcal{E}$. Specifically, it takes the initial hypothesis state $Z_\theta^{(0)} = \{(h_i^{(0)}, c_i^{(0)})\}_{i=1}^K$, together with the image $\mathcal{I}$ and the evidence pool

Dataset	GC	RL	HR	DV	NE	CoT	Tool
Im2GPS3k	✓	✗	✗	✗	✓	✗	✗
MAPBench	✓	○	○	○	✓	○	✓
EarthWhere	✓	○	○	○	✓	○	○
GeoRC	○	✓	○	✗	○	✓	✗
GeoBench	✓	✓	✓	✓	✓	○	✓
PAVED	✓	✓	✓	✓	✓	✓	✓

GC: global coverage; RL: reasonable localizability; HR: high-resolution imagery; DV: data variety; NE: nuanced evaluation; CoT: paired reasoning trajectory; Tool: explicit tool-call trajectory. ✓: yes, ○: partial, ✗: no.

Table 1: Comparison of geolocalization benchmarks. PAVED combines globally distributed human check-in imagery, localizability filtering, high-resolution visual evidence, distance- and address-level evaluation, and auditable tool-grounded reasoning trajectories.

$\mathcal{E}$, as input to the refinement agent $\mathcal{M}_\theta$, as defined in Eq. 1. The agent re-evaluates each candidate hypothesis according to supportive, refuting, and refinement evidence, producing an updated hypothesis state $Z_\theta^{(1)}$. In this process, supportive evidence increases the candidate confidence, refuting evidence decreases it, and severe contradictions trigger a veto that excludes the candidate from final selection.

$$Z_\theta^{(1)} = \mathcal{M}_\theta(\mathcal{I}, Z_\theta^{(0)}, \mathcal{E}) = \{(h_i^{(1)}, c_i^{(1)})\}_{i=1}^K. \quad (1)$$

The final hypothesis is then deterministically selected as the highest-scoring valid candidate: $\hat{h} = \arg\max_{h_i \in Z_\theta^{(1)}} c_i^{(1)}$. By anchoring this second reasoning pass explicitly to auditable evidence, the model effectively reduces context drift and preserves reasoning consistency in ambiguous, long-horizon visual scenarios.

4 Datasets

4.1 Data Collection.

We construct **PAVED** (Perception-and-Verification-Engine-Dataset) from the Foursquare check-in corpus (Noulas et al., 2011), containing 2,524 manually filtered street-view images collected from worldwide human check-ins. PAVED focuses on accessible places with clear semantic cues, such as commercial areas, public streets, and residential districts, better reflecting realistic open-world geo-localization scenarios. As summarized in Table 1, PAVED complements existing benchmarks by jointly supporting global coverage, localizability, high-resolution imagery, nuanced evaluation, and paired tool-grounded reasoning trajectories. Additional details on geographic distribution, coverage, and sampling are provided in the Appendix B.1.

4.2 Chain-of-Thought

Structured CoT. We combine PAVED with agentic perception-verification reasoning. Unlike conventional datasets that merely provide final coordinates and perception-driven reasoning traces, **PAVED** features structured Chain-of-Thought (CoT) trajectories that explicitly decouple reasoning into hypothesis generation, evidence verification, and candidates refinement. Given an image $\mathcal{I}$ and its ground-truth coordinate $\mathbf{p}^*$, we formulate an answer-conditioned expert trajectory as Eq. 2:

$$\Gamma^* = \{Z^{(0)}, \mathcal{E}, Z^{(1)}, \mathbf{p}^*\}, \quad (2)$$

where the initial hypothesis state $Z^{(0)} = \{h_i^{(0)}\}_{i=1}^K$ comprises candidates formalized as $h_i^{(0)} = (\mathbf{p}_i, c_i^{(0)}, \mathcal{R}_i, \mathcal{U}_i, \mathcal{A}_i)$, capturing the coordinate, initial confidence, perception rationale, uncertainty, and recommended verification actions, respectively. The trajectory progresses through $\mathcal{E}$, representing the sequence of tool actions and observations, ultimately yielding the refined hypothesis state $Z^{(1)}$.

Trajectory Curation. The generation follows a rigorous two-stage curation workflow: 1) The base GeoPAVE framework generates a preliminary structured trace that logs candidate generation, tool interactions, evidence evaluations, and the final convergence state; 2) a stronger reviewer model is introduced to audit the initial trace utilizing rejection sampling to enforce tool-call validity, evidence consistency, and prediction precision ($d(\hat{\mathbf{p}}, \mathbf{p}^*) < 5$km), as shown in Figure 3. By formatting the retained trajectories with structured tags (e.g., <think>) to separate internal reasoning from external observations, we provide a high-quality dataset with transparent reasoning traces, offering a valuable resource for analyzing verifiable agentic geo-localization.

5 Experiments

To examine the effectiveness of GeoPAVE and bi-level architecture, we organize our experiments around three research questions:

Model	PAVED					MAPBench-V2					IM2GPS3K				
	500m	2km	25km	200km	750km	500m	2km	25km	200km	750km	500m	2km	25km	200km	750km
GeoCLIP	2.58	7.49	24.52	37.08	59.31	0.88	1.59	6.01	12.39	32.29	8.94	18.35	32.17	49.45	66.53
GeoReasoner	0.99	5.23	31.42	36.57	50.71	0.00	1.46	11.76	21.44	37.09	2.04	10.98	25.19	36.24	52.82
GAEA	1.86	7.25	34.39	41.64	61.69	0.42	5.26	25.45	34.84	56.24	-	-	32.61	51.42	-
GeoVista	1.55	6.74	39.22	44.97	63.31	0.25	4.17	22.86	27.36	33.54	3.47	13.55	28.06	35.90	44.61
GLOBE	0.67	2.85	14.86	20.80	38.99	1.20	10.67	13.25	26.54	21.77	2.10	8.91	35.12	50.24	64.61
Map With Thinking 1 rollout	3.69	9.65	40.53	47.38	70.55	13.73	23.20	49.52	57.95	76.60	7.39	19.81	41.09	56.45	74.62
Map With Thinking 4 rollouts	9.09	17.89	51.03	58.07	78.30	**22.15**	**27.08**	40.26	53.43	62.48	7.94	16.52	29.70	38.44	48.18
Gemini-3.1-Pro	11.10	17.00	57.50	63.40	77.20	12.37	22.51	53.48	76.22	82.72	-	-	-	-	-
SmileGeo	-	-	-	-	-	-	-	-	-	-	-	-	47.77	-	-
GPT-5.5	5.31	13.16	39.42	47.28	67.51	7.68	18.48	34.63	43.68	58.70	**19.19**	**35.37**	**56.52**	**69.44**	**81.38**
GeoPAVE (Ours)	**16.87**	**26.89**	**59.96**	**69.35**	**85.86**	13.81	24.83	**63.19**	**70.42**	**86.87**	15.92	30.51	52.30	66.18	81.24

Table 2: Comparison of geolocation accuracy on PAVED, MAPBench-V2, and IM2GPS3K under different error thresholds (values in %). **Bold** and underline denote the best and second-best results, respectively.

- **RQ1:** Can GeoPAVE achieve fine-grained geo-localization in ambiguous visual geo-localization under lower cost?
- **RQ2:** How do the individual components of GeoPAVE's bi-level architecture contribute to the overall prediction accuracy?
- **RQ3:** Does hypothesis-verification effectively mitigate *perceptual shortcuts* and prevent *contextual inconsistency* during long-horizon ambiguous visual reasoning of geo-localization?

5.1 Experiment Setup

Benchmark. We evaluate GeoPAVE on three complementary benchmarks. **PAVED** is our 2,524-image benchmark constructed from human check-ins, covering worldwide street-view images with visual cues suitable for evidence-based geographic reasoning. **IM2GPS3K** (Vo et al., 2017) is a standard public global geo-localization benchmark containing 2,997 Flickr-sourced geotagged images. **MAPBench-V2** (Ji et al., 2026) is a China-centered benchmark covering urban POI and street-view scenes across multiple cities, and we use its 2.4K-image test split for evaluation. Detailed descriptions of these public datasets are provided in the appendix B.2. Together, these datasets test global open-world localization, legacy Internet-photo localization, and China-specific fine-grained urban geo-localization.

Baselines. We compare GeoPAVE with representative retrieval-based, fine-tuned, reasoning-based, and tool-augmented geo-localization systems. The retrieval baseline is GeoCLIP (Vivanco Cepeda et al., 2023). Fine-tuned or geolocation-specialized LVLM baselines include GAEA (Campos et al., 2025), GeoReasoner (Li et al., 2024), and GLOBE (Li et al., 2025). Agentic and web-augmented baselines include GeoVista (Wang et al., 2025b) and Thinking with Map (Ji et al., 2026). We also consider a strong closed-source LVLMs such as GPT-5.5 and Gemini-3.1-Pro. To ensure a fair comparison with Thinking with Map (Ji et al., 2026), we consider both its original single-rollout setting and test-time scaling approaches as baselines. Specifically, we evaluate the 4-rollout setting by incorporating Gemini-3.0-Flash to provide a comprehensive assessment. More details are in Appendix B.3.

Metrics. We evaluate geo-localization quality from final-coordinate accuracy and verification-trace quality. For coordinate accuracy, we report distance-threshold accuracy $\mathrm{Acc}@r$, where a prediction is correct if its haversine distance to the ground-truth coordinate is within r. r is set as $500\mathrm{m}$, $2\mathrm{km}$, $25\mathrm{km}$, $200\mathrm{km}$, and $750\mathrm{km}$ respectively, covering street-level to regional localization.

Beyond final accuracy, we introduce three trace-based metrics to examine candidate diversity and evidence-grounded correction. Let p_i^* denote the ground-truth coordinate of the i-th image, $\mathcal{H}_i^{(0)} = \{p_{ij}\}_{j=1}^{K_i}$ denote the initial candidate coordinates, $\tilde{p}_i$ denote the initial prediction, and $\hat{p}_i$ denote the

Model	Accuracy (%)				
	500m	2km	25km	200km	750km
P-Only	12.72	24.29	56.63	66.25	85.38
V-Only	14.22	22.28	55.84	65.43	84.61
V w/o OCR	15.16	24.31	58.91	68.54	85.32
V w/o Web	11.87	22.86	56.43	69.00	85.06
V w/o Map	13.58	25.36	57.52	67.82	84.31
GeoPAVE	**16.87**	**26.89**	**59.96**	**69.35**	**85.86**

Table 3: Ablation study of different GeoPAVE configurations on the PAVED dataset. **P** denotes the Perception module, and **V** denotes the Verification module.

final prediction.

Candidate Oracle Recall (COR@r). COR@r measures whether the initial hypothesis set covers the correct geographic region:

$$\mathrm{COR@}r = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\min_{j} d(p_{ij}, p_i^*) \le r\right]. \quad (3)$$

Correction Recall (CR@r). CR@r evaluates whether verification recovers initially failed samples. Given $\mathcal{S}_\tau = \{i \mid e_i^{(0)} > \tau\}$, where $e_i^{(0)} = d(\tilde{p}_i, p_i^*)$, we compute:

$$\mathrm{CR@}r \mid \tau = \frac{1}{|\mathcal{S}_\tau|}\sum_{i\in\mathcal{S}_\tau} \mathbf{1}\left[d(\hat{p}_i, p_i^*) \le r\right]. \quad (4)$$

Log Correction Gain (LCG). LCG measures the logarithmic error reduction from the initial prediction to the final prediction:

$$\mathrm{LCG} = \frac{1}{N}\sum_{i=1}^{N} \log_{10}\frac{d(\tilde{p}_i, p_i^*)+1}{d(\hat{p}_i, p_i^*)+1}. \quad (5)$$

Settings. For closed-source models, we perform inference using their official APIs. For open-sourced models, we perform inference using four NVIDIA A6000 GPUs. All models use the same Nominatim map-service interface to ensure fair comparison under a shared geocoding backend. GeoPAVE is implemented with Gemini-3.0-Flash. More implementation details are provided in the Appendix C.

5.2 Overall Performance (RQ1)

We compare GeoPAVE with representative retrieval-based, fine-tuned, agentic, and closed-source LVLM baselines on PAVED, MAPBench-V2, and IM2GPS3K, as summarized in Table 2. On PAVED, GeoPAVE achieves the best performance across all thresholds, improving Acc@500m from 11.10% to 16.87% and Acc@2km from 17.89% to 26.89%. The consistent gains from fine-grained to coarse-grained thresholds indicate that the perceive-then-verify design improves fine-grained geo-localization, which is especially important for PAVED, where many samples require text cuies, POI, and contextual cues rather than relying on a single visual prior.

The remaining results highlight where GeoPAVE is most effective. The stronger performance of Map With Thinking 4 rollouts on MapBenchV2 can be attributed to the large-scale agentic RL training on the such data, along with test-time rollouts enlarging the search space for reasoning . However, GeoPAVE outperforms it from Acc@25km onward, suggesting that while our method is slightly weaker in exact fine-grained geo-localization,it remains effective at coarse-grained scnarios. For the Flickr-based benchmark, GeoPAVE slightly trails GPT-5.5, which is expected since many images are indoor scenarios, which favor larger closed-source LVLMs with stronger general visual understanding and broader world knowledge. These results suggest that GeoPAVE is most effective when visual observations can be connected to verifiable local evidence, while remaining competitive under less constrained image distributions.

5.3 Ablation Study (RQ2)

Effect of perception and verification. We first evaluate the contribution of GeoPAVE's two-stage design. As shown in Table 3, both *P-only* and *V-only* underperform the full model. Removing verification reduces Acc@500m from 16.87% to 12.72%, showing that initial hypotheses still require evidence-based refinement. Verification-only also degrades performance, indicating that verification is less effective without uncertainty-aware candidate generation. These results confirm that perception and verification are complementary in GeoPAVE.

Effect of verification skills. We further ablate three verification skills: Linguistic Grounding (LG; OCR-based script/language cues), Scene Verification (SV; search-engine evidence), and Contextual Validation (CV; map/POI evidence). As shown in Table 3, removing SV causes the largest drop, reducing Acc@500m from 16.87% to 11.87%, which indicates the importance of search-based scene

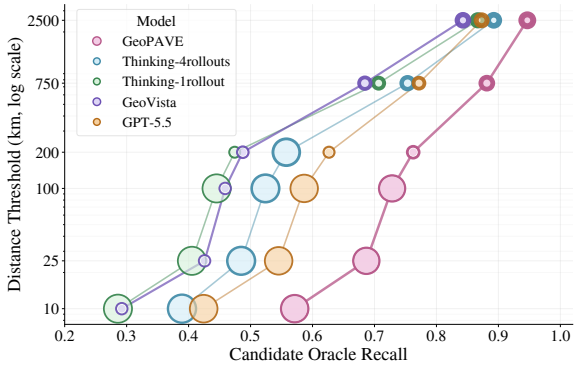

Figure 4: Diagnostic Analysis of Candidate Coverage and Evidence-Grounded Correction.

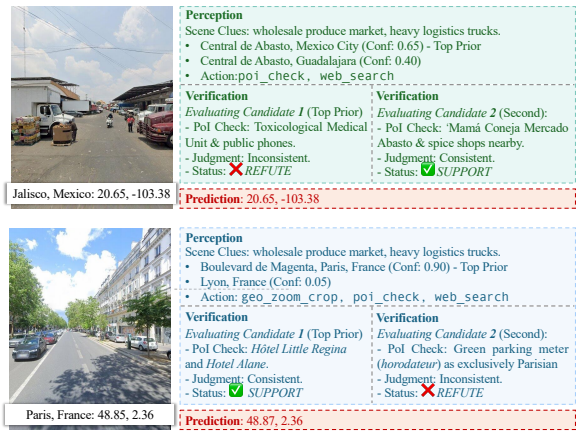

Figure 5: Case Study: Evidence-Grounded Support, Refutation, and Hypothesis Refinement.

grounding. Removing CV decreases Acc@500m to 13.58% and Acc@750km to 84.31%, showing that map-side context helps filter spatially inconsistent hypotheses. Removing LG leads to a milder but consistent decline, suggesting that OCR-derived script and language cues provide useful textual anchors. These results show that the three skills contribute complementary evidence for verification.

5.4 Quantitative Analysis (RQ3)

To examine candidate refinement and contextual consistency, we visualize tool-augmented models in Figure 4 with three diagnostic metrics: Candidate Oracle Recall (COR) for initial hypothesis coverage, Correction Recall (CR) for recovery from failed reasoning, and Log Correction Gain (LCG) for relative error reduction. Unlike Acc@r, these metrics expose the correction process behind final predictions. Figure 4 shows that baselines mainly improve at coarse thresholds, while GeoPAVE achieves higher COR and stronger fine-grained localization, especially within 10 km. Its larger LCG bubbles and thicker CR contours indicate more effective correction of intermediate errors. These results suggest that GeoPAVE mitigates context drift by grounding hypothesis refinement in evidence-supported candidate regions.

5.5 Case Study

Figure 5 presents two examples to better understand the perception-verification process of GeoPAVE. In the first case, the perception stage recognizes the scene as a wholesale market with logistics trucks, and keeps both Mexico City and Guadalajara as plausible "Central de Abasto" candidates. Although the top prior points to Mexico City, the correct Guadalajara region remains in the candidate set, showing that the initial hypotheses provide useful coverage under visual ambiguity. The verification stage then checks nearby POI evidence. The Mexico City candidate is rejected because its local context is inconsistent with the observed market scene, while the Guadalajara candidate is supported by nearby market-related POIs, leading to the final correction.

The Paris case shows a complementary situation. The perception stage already proposes Boulevard de Magenta as the strongest hypothesis, and verification confirms it with nearby hotel and street-level evidence, while the Lyon alternative is ruled out due to inconsistent contextual cues. Together, these cases show that perception preserves plausible candidates and verification uses external evidence to correct or confirm them, reducing reliance on a single perceptual prior. Details are provided in Appendix D.3.

6 Conclusion and Future Works

We reformulate open-world geo-localization as a **perceive-then-verify** reasoning problem and propose **GeoPAVE**, a bi-level agentic framework that combines perception-based hypothesis generation with evidence-grounded verification, reducing AI shortcuts and context inconsistency under ambiguous visual cues and demonstrating the effectiveness of GeoPAVE in both fine-grained geo-localization.

Future work will extend GeoPAVE and PAVED into a scalable data generation and training pipeline, using verified reasoning traces to train stronger geo-localization agents. We will further explore agentic reinforcement learning in interactive environments, where models can actively gather evidence and improve geographic reasoning through feedback.

Limitations

While GeoPAVE demonstrates strong performance across multiple benchmarks, several limitations remain. First, the current verification module relies on a rule-based uncertainty-to-skill routing policy, which may not generalize optimally to edge cases where uncertainty signals are ambiguous or underspecified. Second, GeoPAVE's inference pipeline involves multiple sequential tool invocations, which introduces additional latency compared to single-pass reasoning approaches; future work could explore more efficient verification scheduling strategies. Third, although PAVED provides globally distributed coverage, its images are sourced from Foursquare check-ins and may overrepresent commercially active urban areas, potentially limiting evaluation diversity in rural or less-documented regions. Finally, the trajectory curation process relies on a stronger reviewer model to filter low-quality traces, and the quality of PAVED annotations is thus partially dependent on the capabilities of this external model.

Ethical Consideration

PAVED is derived from real-world check-in locations and street-view imagery, which raises privacy, licensing, and misuse concerns. We remove user identifiers, timestamps, and other check-in metadata that could reveal individual mobility patterns. When raw imagery cannot be redistributed under third-party licenses, we release only derived annotations, metadata, or retrieval scripts. Image geolocalization systems may be misused for surveillance, doxxing, or sensitive-location identification; therefore, GeoPAVE and PAVED are intended for benchmark research and should not be used to identify private individuals or infer sensitive personal locations.

AI-assisted technologies were used solely for language polishing and grammar correction to improve the clarity and readability of the manuscript. They were not used to generate the core ideas, experimental design, results, analyses, or references. The authors carefully reviewed all AI-assisted edits and take full responsibility for the integrity of the work, including the validity of the experimental results and the accuracy of all citations.

References

Zhixi Cai, Fucai Ke, Kevin Leo, Sukai Huang, Maria Garcia de la Banda, Peter J Stuckey, and Hamid Rezatofighi. 2026. Mata: A trainable hierarchical automaton system for multi-agent visual reasoning. *arXiv preprint arXiv:2601.19204*.

Ron Campos, Ashmal Vayani, Parth Parag Kulkarni, Rohit Gupta, Aritra Dutta, and Mubarak Shah. 2025. Gaea: A geolocation aware conversational model. *arXiv e-prints*, pages arXiv–2503.

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

Xinyu Geng, Yanjing Xiao, Yuyang Zhang, Hanwen Wang, Xinyan Liu, Rui Min, Tianqing Fang, and Yi R Fung. 2026. Geobrowse: A geolocation benchmark for agentic tool use with expert-annotated reasoning traces. *arXiv preprint arXiv:2604.04017*.

Narges Ghasemi, Amir Ziashahabi, Salman Avestimehr, and Cyrus Shahabi. 2025. Geotoken: Hierarchical geolocalization of images via next token prediction. *arXiv preprint arXiv:2511.01082*.

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.

Xiao Han, Chen Zhu, Hengshu Zhu, and Xiangyu Zhao. 2025a. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 814–825.

Xiao Han, Chen Zhu, Hengshu Zhu, and Xiangyu Zhao. 2025b. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 814–825.

Yuxiang Ji, Yong Wang, Ziyu Ma, Yiming Hu, Hailang Huang, Xuecai Hu, Guanhua Chen, Liaoni Wu, and Xiangxiang Chu. 2026. Thinking with map: Reinforced parallel map-augmented agent for geolocalization. *arXiv preprint arXiv:2601.05432*.

Furong Jia, Ling Dai, Wenjin Deng, Fan Zhang, Chen Hu, Daxin Jiang, and Yu Liu. 2026. Spotagent: Grounding visual geo-localization in large vision-language models through agentic reasoning. *arXiv preprint arXiv:2602.09463*.

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.

Advances in neural information processing systems, 37:53198–53221.

Pengyue Jia, Yingyi Zhang, Xiangyu Zhao, and Sharon Li. 2025. Geoarena: An open platform for benchmarking large vision-language models on worldwide image geolocalization. *arXiv preprint arXiv:2509.04334*.

Saeed Khaki, Ashudeep Singh, Nima Safaei, and Kamal Ginotra. 2026. Vistira: Closing the image-text modality gap in visual math reasoning via structured tool integration. *arXiv preprint arXiv:2601.14440*.

Ling Li, Yu Ye, Yao Zhou, Bingchuan Jiang, and Wei Zeng. 2024. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. *arXiv preprint arXiv:2406.18572*.

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

Renda Li, Hailang Huang, Fei Wei, Feng Xiong, Yong Wang, and Xiangxiang Chu. 2026. Adacurl: Adaptive curriculum reinforcement learning with invalid sample mitigation and historical revisiting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23123–23131.

Jinkun Liu, Haohan Chi, Lingfeng Zhang, Yifan Xie, YuAn Wang, Long Chen, Hangjun Ye, Xiaoshuai Hao, and Wenbo Ding. 2026. Thinking in text and images: Interleaved vision–language reasoning traces for long-horizon robot manipulation. *arXiv preprint arXiv:2605.00438*.

Maria Lymperaiou, Vasileios Karampinis, Giorgos Filandrianos, Angelos Vlachos, Chrysoula Zerva, and Athanasios Voulodimos. 2026. Reasoning or pattern matching? probing large vision-language models with visual puzzles. *arXiv preprint arXiv:2601.13705*.

Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 570–573.

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. In *European Conference on Computer Vision*, pages 196–215. Springer.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, and 1 others. 2025. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents. *arXiv preprint arXiv:2509.13309*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pages 9695–9717.

Yulei Qin, Xiaoyu Tan, Zhengbao He, Gang Li, Haojia Lin, Zongyi Li, Zihan Xu, Yuchen Shi, Siqi Cai, Renting Rui, and 1 others. 2025. Learn the ropes, then trust the wins: self-imitation with progressive exploration for agentic reinforcement learning. *arXiv preprint arXiv:2509.22601*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.

Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. 2018. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551.

David G Shatwell, Ishan Rajendrakumar Dave, Sirnam Swetha, and Mubarak Shah. 2025. Gt-loc: Unifying when and where in images through a joint embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11.

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. *arXiv preprint arXiv:2601.21278*.

Dat Tran and Douwe Kiela. 2026. Single-agent llms outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets. *arXiv preprint arXiv:2604.02460*.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701.

Nam Vo, Nathan Jacobs, and James Hays. 2017. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 2621–2630.

Chun Wang, Xiaojun Ye, Xiaoran Pan, Zihao Pan, Haofan Wang, and Yiren Song. 2025a. Gre suite: Geo-localization inference via fine-tuned vision-language models and enhanced reasoning chains. *arXiv preprint arXiv:2505.18700*.

Yikun Wang, Zuyan Liu, Ziyi Wang, Han Hu, Pengfei Liu, and Yongming Rao. 2025b. Geovista: Web-augmented agentic visual reasoning for geolocalization. *arXiv preprint arXiv:2511.15705*.

Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet-photo geolocation with convolutional neural networks. In *European conference on computer vision*, pages 37–55. Springer.

Chenjie Yang, Yutian Jiang, and Chenyu Wu. 2026. Skill-conditioned visual geolocation for vision-language. *arXiv preprint arXiv:2604.09025*.

Bo Yu, Fengze Yang, Yiming Liu, Chao Wang, Xuewen Luo, Taozhe Li, Ruimin Ke, Xiaofan Zhou, and Chenxi Liu. 2026. Locatability-guided adaptive reasoning for image geo-localization with vision-language models. *arXiv preprint arXiv:2603.13628*.

Heng Zheng, Yuling Shi, Xiaodong Gu, Haochen You, Zijian Zhang, Lubin Gan, Hao Zhang, Wenjun Huang, and Jin Huang. 2025. Graphgeo: Multi-agent debate framework for visual geo-localization with heterogeneous graph neural networks. *arXiv preprint arXiv:2511.00908*.

King Zhu, Hanhao Li, Siwei Wu, Tianshun Xing, Dehua Ma, Xiangru Tang, Minghao Liu, Jian Yang, Jiaheng Liu, Yuchen Eleanor Jiang, and 1 others. 2025. Scaling test-time compute for llm agents. *arXiv preprint arXiv:2506.12928*.

Appendix

A Experiment Results

To provide a more fine-grained evaluation of GeoPAVE on PAVED, we further report Acc@10km and Acc@100km in addition to the main thresholds. These two intermediate thresholds measure city-level and regional-level localization accuracy, respectively, and help clarify whether the gains of GeoPAVE remain consistent beyond street-level prediction. As shown in Table 4, GeoPAVE achieves 47.76% at Acc@10km and 65.31% at Acc@100km, outperforming all baselines at both thresholds. Compared with the strongest closed-source baseline, Gemini-3.1-Pro, GeoPAVE improves Acc@10km by 2.26 percentage points and Acc@100km by 4.61 percentage points. These results suggest that evidence-grounded verification not only improves fine-grained localization, but also provides stable benefits for broader city- and region-level geographic reasoning.

B Datasets

To ensure distributional diversity and rigorously evaluate the open-world geolocalization capabilities of agentic models, our evaluation spans three distinct datasets. These datasets were carefully selected and curated to assess models across legacy internet photos, region-specific urban environments, and human-centric reasoning scenarios.

B.1 PAVED

PAVED (Perception-and-Verification-Dataset) consists of 2,524 manually filtered street-view images collected from worldwide human check-ins, with hypothesis locations sourced from the Foursquare check-in corpus (Noulas et al., 2011). Each image is anchored to a real point of interest visited by users and corresponds to accessible, semantically meaningful urban scenes, such as commercial streets, local neighborhoods, storefronts, road signs, and surrounding POIs. We retrieve high-resolution imagery through Google Maps and manually filter out non-localizable samples, including indoor scenes, close-up objects, visually uninformative views, and overly iconic global landmarks. As a result, PAVED focuses on realistic geo-localization cases where models must infer locations from rich but often ambiguous street-level semantic cues.

As shown in Table 1, PAVED differs from existing geo-localization benchmarks in both data construction and reasoning supervision. While prior datasets often emphasize either global coverage, retrieval-style evaluation, or final-coordinate prediction, PAVED jointly provides global human-centric imagery, reasonable localizability, high-resolution visual evidence, data diversity, and nuanced distance/address-level evaluation. More importantly, PAVED is paired with structured Chain-of-Thought annotations that decompose each sample into hypothesis generation and tool-augmented verification. These trajectories record how external evidence is retrieved, evaluated, and used to refine hypothesis locations, enabling PAVED to evaluate whether agentic models can actively verify hypotheses rather than rely only on parametric priors.

Unlike existing benchmarks that rely on randomly scraped coordinates or easily recognizable landmarks, PAVED is constructed from the Foursquare check-in corpus (Noulas et al., 2011), ensuring that every image is anchored to genuine human activity and accessible, semantically meaningful places (e.g., commercial districts, local neighborhoods). All corresponding high-resolution imagery is retrieved via Google Maps to support reliable visual clue extraction and grounding. To maintain meaningful localization difficulty, we conducted localizability filtering to remove non-localizable images (e.g., indoor rooms, close-up objects) and easily recognizable global landmarks. Most importantly, PAVED is accompanied by Structured Chain-of-Thought (Structured CoT) annotations that decouple the reasoning process into a hypothesis-generation state and a tool-augmented verification state. This unique structure provides a robust testbed to evaluate whether agentic models can actively retrieve external information to validate or refine their hypotheses, rather than relying solely on parametric priors.

B.2 Details of Benchmarks

IM2GPS3K. As a standard and widely recognized global benchmark (Vo et al., 2017), IM2GPS3K contains 2,997 geotagged images sourced from Flickr. This dataset serves to evaluate general, open-world localization capabilities

Model	Performance on GeoPAVE (%)						
	500m	2km	10km	25km	100km	200km	750km
GeoCLIP	2.58	7.49	16.36	24.52	32.21	37.08	59.31
GeoReasoner	0.99	5.23	21.35	31.42	34.07	36.57	50.71
GAEA	1.86	7.25	24.60	34.39	38.19	41.64	61.69
GeoVista	1.55	6.74	26.78	39.22	42.31	44.97	63.31
GLOBE	0.67	2.85	10.34	14.86	18.23	20.80	38.99
Thinking	3.69	9.65	28.97	40.53	44.46	47.38	70.55
Thinking*4+verifier	9.09	17.89	38.46	51.03	55.29	58.07	78.30
Gemini 3.1 pro	11.10	17.00	45.50	57.50	60.70	63.40	77.20
GPT 5.5	5.31	13.16	30.06	39.42	43.55	47.28	67.51
GeoPAVE	**16.87**	**26.89**	**47.76**	**59.96**	**65.31**	**69.35**	**85.86**

Table 4: Comparison of geolocation accuracy under different error thresholds (values in %). Best results are in bold.

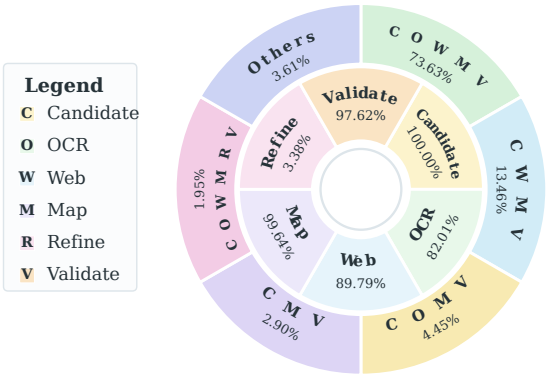

Figure 6: Tool Invocation Patterns in PAVED Verification Traces.

on legacy internet photos. Because these images are largely crowd-sourced and vary significantly in quality, visual composition, and geographic context, they pose a traditional yet robust challenge for testing whether models can extract meaningful location cues from unstructured and often noisy visual data.

MAPBench-V2. To test fine-grained, region-specific reasoning and robustness, we utilize the 2.4K-image test split of MAPBench-V2 (Ji et al., 2026). This benchmark is heavily China-centered, featuring a diverse array of urban points of interest (POIs) and street-view scenes across multiple cities. Evaluating on MAPBench-V2 allows us to assess a model's ability to interpret specific regional infrastructure, localized textual cues (e.g., Chinese storefront signage), and distinct architectural styles, thereby testing the limits of its geographic knowledge and fine-scale urban geolocalization.

B.3 Baselines

Open-source Models

- **GeoCLIP (Vivanco Cepeda et al., 2023):** A CLIP-inspired image-to-GPS retrieval approach that models the Earth as a continuous space. It employs positional encoding with random Fourier features and constructs a hierarchical representation to effectively align image embeddings with corresponding GPS locations for worldwide geolocalization.
- **GeoReasoner (Li et al., 2024):** A novel framework leveraging large vision-language models (LVLMs) augmented with human inference knowledge derived from real geo-localization games. To ensure data quality, it first curates a dataset of highly locatable street views, followed by specialized fine-tuning through dedicated reasoning and location-tuning stages to enhance geographical deduction.
- **GAEA (Campos et al., 2025):** A conversational visual-language model explicitly designed for geolocation awareness. It is trained on GAEA-1.4M, a synthesized large-scale dataset built upon OpenStreetMap attributes, enabling nuanced location-specific reasoning and dialogue beyond pure coordinate prediction.
- **GeoVista (Wang et al., 2025b):** A web-augmented agentic model designed for geolocalization that integrates explicit tool invocations, such as image zooming and web searching, directly into the visual reasoning loop. It is optimized through supervised fine-tuning followed by reinforcement learning.

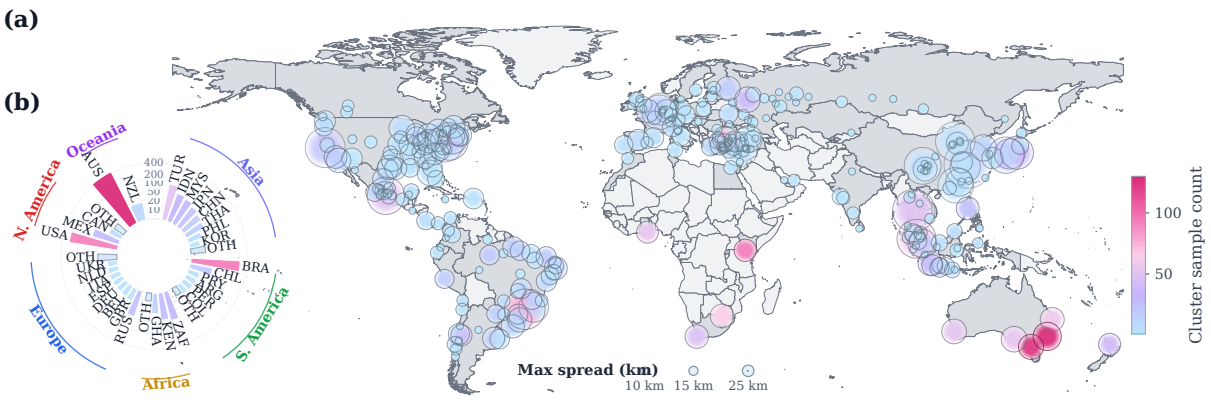

Figure 7: Geological Distribution of PAVED. Our dataset provides high-resolution, globally annotated imagery sourced from real human check-ins, designed to evaluate the general perception and verification abilities of agentic models.

- **smileGeo (Han et al., 2025b):** A collaborative swarm intelligence framework that leverages multiple Internet-enabled LVLM agents to associate images with precise geographic locations. It facilitates inter-agent communication to synthesize models' inherent knowledge with retrieved web information, employing a dynamic learning strategy to optimize interactions and reduce redundant communication.

- **Map With Thinking (Ji et al., 2026):** A map-augmented reasoning agent that utilizes reinforcement learning to gather geographic evidence. To comprehensively evaluate its capacity, we benchmark it under two distinct inference settings:
  - **1 rollout:** The base setting where the RL-finetuned agent is directly utilized to perform a single-pass inference for coordinate prediction.
  - **4 rollouts:** A parallel test-time scaling framework where four agents generate independent hypothesis locations. In our implementation, considering the trade-off between reasoning performance and computational overhead, we adapted the original pipeline by deploying Gemini 3.0 Flash as the strict verifier (in place of the original Qwen3-VL-235B-A22B) to synthesize the gathered evidence and output the most geographically plausible final coordinate.

Closed-source Models

- **Gemini 3.1 Pro:** State-of-the-art native multimodal foundation model developed by Google. Thanks to its massive context windows and advanced high-resolution visual processing capabilities, Gemini 3.1 Pro excels at extracting fine-grained geographic clues (e.g., street signs, architectural styles, and vegetation), serving as a heavy-weight baseline offering deep, multi-step spatial reasoning and extensive zero-shot world knowledge.

- **GPT-5.5:** The latest proprietary large multimodal model introduced by OpenAI, representing the current frontier of closed-source visual reasoning. It demonstrates unprecedented zero-shot generic visual comprehension and cross-lingual text-in-image reading capabilities. For geolocalization, it leverages powerful chain-of-thought (CoT) reasoning to synthesize diverse environmental clues and accurately deduce global coordinates even in unconstrained or ambiguous street views.

C Evaluation Metric Details

This section supplements the evaluation metrics defined in Section 5.1. We provide implementation details for coordinate parsing, distance computation, candidate-based diagnostics, and aggregation.

Coordinate accuracy. For each image I_i, a method outputs a final coordinate $\hat{p}_i = (\widehat{\mathrm{lat}}_i, \widehat{\mathrm{lon}}_i)$. If a response contains multiple coordinates, we use the coordinate explicitly marked as the final answer. For GeoPAVE, this corresponds to the final prediction after verification. For baselines, we parse the final latitude–longitude pair using a deterministic coordinate parser.

We compute the geographic error using the haversine distance with Earth radius $R = 6371.0088$ km. Distance-threshold accuracy is de-

fined as:

$$\mathrm{Acc@}r = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}\left[d(\hat{p}_i, p_i^*) \leq r\right].$$

We report $\mathrm{Acc@}r$, where a prediction is considered correct if its haversine distance to the ground-truth coordinate is within radius r. We evaluate $r \in \{0.5, 2, 10, 25, 100, 200, 750\}$ km, and present representative thresholds in the main tables due to space constraints. Predictions with missing, invalid, or unparsable coordinates are counted as incorrect for all thresholds.

Candidate-based diagnostics. For GeoPAVE, the initial candidate set $H_i^{(0)}$ is the set of hypotheses generated by the perception module before verification. The initial prediction $\tilde{p}_i$ is the top-ranked candidate before verification, and the final prediction $\hat{p}_i$ is the coordinate selected after evidence-grounded refinement. For methods that output only one coordinate, the initial candidate set is treated as a singleton.

We use the same definitions of Candidate Oracle Recall (COR), Correction Recall (CR), and Log Correction Gain (LCG) as in Section 5.1. For CR, we set $\tau = r$ unless otherwise specified, meaning that $\mathrm{CR@}r$ measures the fraction of samples that are initially wrong at threshold r but become correct after verification. If no sample fails initially at threshold r, CR is reported as N/A. For LCG, invalid predictions are assigned a clipped maximum error of 20,015 km, approximately half of Earth's circumference.

D Implementation Details

D.1 Perception

The perception stage first produces a coarse fallback and three candidate locations, each with confidence, uncertainty tags, and recommended verification actions. We expose two image-level tools, Zoom-Crop and OCR, whose definitions are shown in Figure 8 and Figure 9. The controller maps uncertainty tags to tool actions through the routing policy in Figure 10, and the OCR prompt is provided in Figure 11.

D.2 Verification

The verification stage grounds each perception candidate with OCR, Web, and Map evidence. We define the verification tools in Figure 12, Figure 13, and Figure 14. Map evidence is exposed through three compact sub-tools rather than raw map-service outputs. Tool invocation and refinement follow Figure 15. All tool outputs are serialized into the compact interface in Figure 16 before being sent to the verification refiner prompt in Figure 17.

D.3 Qualitative Results

Figure 18 shows that our proposed model (GeoPAVE) produces reasoning trajectories with improved coherence and interpretability. In particular, the model keeps ambiguous visual cues as explicit hypotheses, records uncertainty when decisive text is absent, and uses external evidence to prevent visually plausible but over-specific candidate coordinates from being accepted.

We present a comparative example in Figure 18 to illustrate the difference in reasoning capabilities. Given a suburban street view, **GPT-5.5** suffers from a severe cue-level error; it incorrectly identifies the scene as having right-hand traffic and misinterprets the road infrastructure as European, ultimately predicting Luxembourg with a distance error of over 8,700 km. **Gemini-3.1-Pro** demonstrates stronger perception by correctly identifying left-hand traffic, the distinctive interlocking zig-zag bricks, and the green palisade fencing, narrowing the location down to South Africa. However, it relies entirely on its internal parametric memory, leading to an incorrect city-level prediction (Pretoria).

In contrast, **GeoPAVE** exhibits a structured reasoning trajectory. It first extracts fine-grained visual evidence, including the yellow road edge lines and interlocking brick median paving. More importantly, instead of making an immediate point prediction, our model formulates multiple hypotheses (Johannesburg vs. Pretoria) and leverages external tool-use (e.g., web queries for "yellow road edge lines left hand drive countries") to evaluate candidate consistency. When overly specific retrieved candidates conflict with the hilly sports-field scene, the evidence gate rejects them and falls back to the broader Johannesburg hypothesis.

Tool: `Zoom-Crop`

Description: Crops a complete local region for close inspection when a visible sign, storefront, plate, address panel, or landmark detail is present but not readable at full-image scale.
Arguments:

- `bbox_2d` (array of numbers): Bounding box in normalized image coordinates. *(Required)*
- `label` (string): Short label for the target region. *(Optional)*

Figure 8: Perception-stage Zoom-Crop tool definition.

Tool: `OCR`

Description: Reads legible text from the original image or a returned crop, and converts it into compact place-like and auxiliary text evidence.
Arguments:

- `image_id` (string): The original image or the crop identifier returned by `Zoom-Crop`. *(Required)*

Figure 9: Perception-stage OCR tool definition.

Algorithm: Perception Tool Routing

Require: Image I, top candidate c_1, uncertainty tags T, recommended actions A, need-zoom flag z, zoom targets Z.
$T_{\text{text}} \leftarrow \{\texttt{TEXT_UNCLEAR}, \texttt{SYMBOLS_UNCLEAR}\}$.
$T_{\text{text}} \leftarrow T_{\text{text}} \cup \{\texttt{NEED_ZOOM}\}$.
$Z_{\text{text}} \leftarrow \{\texttt{sign_text}, \texttt{license_plate}, \texttt{storefront}\}$.
$T_{\text{map}} \leftarrow \{\texttt{CONTEXT_MISSING}, \texttt{TOPOGRAPHY_UNCLEAR}, \texttt{LANDMARK_ID_UNCLEAR}\}$.
$T_{\text{map}} \leftarrow T_{\text{map}} \cup \{\texttt{WATERBODY_DOMINANT}, \texttt{CONFLICTING_CUES}\}$.
$r_{\text{text}} \leftarrow z$ or $T \cap T_{\text{text}} \neq \emptyset$ or $Z \cap Z_{\text{text}} \neq \emptyset$.
$r_{\text{text}} \leftarrow r_{\text{text}}$ or $\texttt{OCR} \in A$ or $\texttt{ZOOM} \in A$.
if no visible text, sign, storefront, plate, or landmark detail exists **then**
    Skip image tools.
else if visible text is already readable **then**
    Store it as OCR memory.
else if r_{text} **then**
    Crop the complete text-bearing object if a close-up is needed.
    Run OCR on the crop or original image.
end if
if OCR memory contains a place, brand, road, or venue, or visual/text cues create an unresolved location conflict **then**
    Record a short reasoning Web query or POI query for Verification.
end if
if $T \cap T_{\text{map}} \neq \emptyset$ **then**
    Record a Map Context request for Verification.
end if

Figure 10: Perception-stage tool routing algorithm.

Prompt: OCR Module

```
You are an OCR module.
Read ALL legible text from the image.

Return ONLY a valid JSON object with this schema (no markdown, no extra text):
{
  "place_like_texts": ["...", "..."],
  "other_texts": ["...", "..."]
}

Rules:
- place_like_texts: ONLY snippets that look like names of places/venues/brands/organizations.
- other_texts: any other readable text (optional).
- Prefer short readable snippets over empty output when at least 3 consecutive visible
    characters can be read.
- If a snippet may be useful for geolocation but you are unsure whether it is a
    place/brand/organization, include it in other_texts instead of dropping it.
- If nothing readable, return empty lists.
- Do NOT invent text you cannot see.
```

Figure 11: OCR module prompt used in the perception stage.

Tool: `Text Evidence`

Description: Reuses OCR memory as image-grounded anchors for candidate validation and address refinement.

Arguments:

- `ocr_texts` (array of strings): Place-like and auxiliary snippets extracted from the image. *(Required)*

Figure 12: Verification-stage text evidence tool definition.

Tool: `Web Evidence`

Description: Executes perception-authored reasoning, OCR, and POI queries, then returns compact search evidence for each candidate.

Arguments:

- `queries` (array of strings): Web or POI-oriented queries authored during perception. *(Required)*

Figure 13: Verification-stage web evidence tool definition.

Tool: `Map Context`

Description: Provides compact map-side evidence for a candidate location, including local POI context, terrain plausibility, and spatial feature consistency. The tool returns structured support/refute signals rather than raw map-service outputs.

Arguments:

- `coordinate` (pair of floats): Candidate latitude and longitude. *(Required)*
- `radius_m` (integer): Local validation radius; default is 500m. *(Optional)*
- `keyword` (string): Optional POI name or category from perception queries. *(Optional)*
- `expected_terrain` (string): Predicted terrain tag, e.g., coastal, flat, hilly, or alpine. *(Optional)*
- `expected_features` (array of strings): Visual features to verify, such as coastline, river, forest, peak, bridge, station, or park. *(Optional)*

Core Capabilities:

- **Neighborhood POI Probe:** counts nearby POIs and returns nearest POI names, types, and distances within the local radius.
- **Topographic Consistency Check:** compares external elevation evidence with the terrain/elevation tag predicted from the image.
- **Spatial Feature Cross-Check:** checks local map-feature counts and ring-style spatial evidence to detect contradictions with the visual scene.

Figure 14: Verification-stage map context tool definition.

Algorithm: Verification Tool Routing

Require: Initial hypotheses $Z^{(0)} = \{h_i^{(0)}\}_{i=1}^{K}$, perception output P, OCR memory O, tool availability G, budgets B.
Ensure: Evidence pool $\mathcal{E}$, refined hypotheses $Z^{(1)}$, and final prediction $\hat{h}$.

Part I: Skill Invocation

Initialize evidence pool $\mathcal{E} \leftarrow \emptyset$ and query bank $Q \leftarrow \emptyset$.
Append to Q: P.`reasoning_web_queries`, P.`ocr_web_queries`, P.`poi_search_queries`, and candidate-level Web/POI queries from $Z^{(0)}$.
Deduplicate Q, reject generic or malformed queries, and retain at most B_Q admitted queries.
Let U_1 and $c_1^{(0)}$ denote the top-hypothesis uncertainty tags and confidence.
if Web Evidence is available and Q contains admitted Web queries **then**
  if the Web circuit is open or the text-only path has no OCR anchor **then**
    Skip Web Evidence and record the skip reason.
  else
    Invoke Web Evidence on the top budgeted Web queries and add compact snippets/entities to $\mathcal{E}$.
  end if
end if
Select up to B_V validation hypotheses from $Z^{(0)}$ by confidence, semantic specificity, and uncertainty; include low-confidence or conflict-marked candidates.
for each selected hypothesis $h_i^{(0)}$ **do**
  Add Text Evidence from OCR memory O to $\mathcal{E}_i$.
  Invoke Map Context for $h_i^{(0)}$ when map evidence is available or U_i contains context, topography, landmark, waterbody, or conflicting-cue uncertainty.
  Add Neighborhood POI Probe within 500m, matched POI detail when an anchor matches, and topographic consistency evidence to $\mathcal{E}_i$.
  Serialize $h_i^{(0)}$, image summary, query bank, OCR memory, Web Evidence, and Map Context as $\mathrm{CONTEXT_JSON}_i$.
end for

Part II: Refinement

if no low-confidence, conflict, query, or external-evidence signal is present **then**
  Skip refinement and keep the current evidence-gated ranking.
end if
for each serialized context $\mathrm{CONTEXT_JSON}_i$ **do**
  $j_i \leftarrow \mathrm{VerificationRefiner}(\mathrm{CONTEXT_JSON}_i)$.
  Parse j_i as {`confidence`, `consistent`, `reason_short`, `refined_address`, `conflicting_cues`, `new_candidates`}.
  if j_i.`conflicting_cues` $\neq \emptyset$ **then**
    Cap j_i.`confidence` < 0.90.
  end if
  Add a self-consistency evidence item to $\mathcal{E}$ with decision SUPPORT, REFUTE, or NODATA.
  if j_i.`consistent` and j_i.`confidence` ≥ 0.90 **then**
    Accept $h_i^{(0)}$ as validated; apply `refined_address` only as an address refinement for the same hypothesis.
    return $Z^{(1)}$ and $\hat{h} = h_i^{(1)}$.
  else
    Keep $h_i^{(0)}$ unresolved and collect at most B_S admissible proposals from `refined_address` and `new_candidates`.
  end if
end for
for each collected proposal q_s **do**
  Land q_s as a new candidate, then repeat Map Context serialization and VerificationRefiner scoring.
  if the spawned candidate reaches confidence ≥ 0.90 **then**
    Insert it into $Z^{(1)}$ and return it as $\hat{h}$.
  end if
end for
Select $\hat{h} = \arg\max_{h_i \in Z^{(1)}} c_i^{(1)}$ among evidence-gated candidates; if none is validated, keep the best unresolved score below 0.90 and expose remaining conflicts.

Figure 15: Verification-stage tool routing and refinement algorithm.

Interface: Tool-Evidence Serialization

```
Tool outputs are serialized as CONTEXT_JSON:
{
  "task": "Judge whether one candidate is self-consistent with image reasoning and tool
    evidence. Return JSON only.",
  "candidate": {...},
  "image_summary": {
    "text_presence": "...",
    "scene_clues": ["..."],
    "expected_elevation_tag": "...",
    "top_level_conflicting_cues": ["..."]
  },
  "query_bank": [{"q": "...", "src": "...", "web": true, "poi": false}],
  "ocr_texts": ["..."],
  "web_search": [{"query": "...", "title": "...", "url": "...", "snippet": "..."}],
  "web_entities": ["..."],
  "nearby": {"count": 0, "radius_m": 500, "pois": [...]},
  "poi_details": [{...}],
  "elevation_m": 0.0
}
```

Figure 16: Tool-evidence serialization interface passed to the verification refiner.

Prompt: Verification Refiner

```
Return ONLY a valid JSON object with this exact shape:
{"confidence":0.0,"consistent":false,"reason_short":"<=160 chars","refined_address":"","
    conflicting_cues":[],"new_candidates":[]}
Rules: confidence is this perception candidate's consistency after merging image reasoning/
    memory with new external evidence. If evidence supports refining the address for the SAME
     perception candidate, set refined_address. If confidence stays <0.90, propose a next
    candidate only when the evidence clearly supports switching; otherwise leave
    new_candidates empty and the controller will advance to the next perception candidate. If
     conflicting cues remain, confidence MUST be < 0.90. Use only provided evidence; do not
    invent street names or POIs. Do not force new candidates to be near the fallback; judge
    semantic consistency from the evidence.

CONTEXT_JSON:
{CONTEXT_JSON}
```

Figure 17: Verification refiner prompt.

Johannesburg, South Africa

GPT-5.5 — **Luxembourg, Luxembourg**

Visual premise: Calls the scene "Luxembourg City" with right-hand traffic, European road furniture, Luxembourg-style curbs/paving, and a fenced sports field.

Place prior: Maps the tree-lined sports-ground scene to Limpertsberg/Belair near Avenue de la Faïencerie and Val Sainte-Croix.

Failure mode: The wrong traffic-direction cue compounds into a European location prior.

Final: Luxembourg, Luxembourg; error 8706.31 km 😭.

GeoPAVE — **Johannesburg, South Africa**

Perception: Extracts left-hand driving, yellow road edge lines, interlocking brick median paving, green sports-field fencing, and hilly Highveld-like terrain; no decisive text is readable.

Routing: Marks `TEXT_UNCLEAR` and `GENERIC_SCENE`, routing the case from direct prediction to Web, POI, and map-context verification.

Hypotheses: Maintains Johannesburg (0.60), Pretoria (0.30), and Durban (0.10), preserving candidate diversity under ambiguity.

Tools: Issues the cue-aligned Web queries and candidate-specific POI/map checks described in the method.

Evidence pool: Serializes Johannesburg evidence at Sandhurst/Sandton and Pretoria evidence at Pretoria Central together with nearby POI and terrain context.

Refiner: Re-scores the candidate set against the evidence pool; both landed candidates are downgraded to 0.30 consistency.

Decision: Applies REFUTE to the precise candidates and keeps the broader Johannesburg fallback because dense CBD contexts conflict with the hilly sports-field scene.

Final: Falls back to the broader Johannesburg hypothesis; error 8.24 km 😊.

Gemini-3.1-Pro — **Pretoria, South Africa**

Perception: Correctly identifies left-side driving, interlocking zig-zag bricks, green palisade-style fencing, curved streetlights, and rolling Highveld terrain.

Scene layout: Uses tall curved streetlights and the general urban/suburban road layout as South African context.

Ambiguity: Notes that the terrain suggests either Pretoria or Johannesburg, but does not verify the city-level choice.

Final: Stops after single-pass reasoning and selects Pretoria; error 50.24 km 😑.

Figure 18: Reasoning comparison of three different models (GeoPAVE, Gemini-3.1-Pro, and GPT-5.5) on the same input image. Reliable visual cues identified by the models are marked in blue, while severe visual hallucinations are marked in red. GeoPAVE preserves uncertainty, verifies candidate consistency, and avoids accepting over-specific coordinates that conflict with the visual scene.