GeoEvolve: Recursive Skill Evolution for Image Geolocalization

Anonymous authors
Paper under double-blind review

Abstract

Image geolocalization with large vision-language models (LVLMs) can benefit from external evidence gathered through tools when visual reasoning alone is insufficient to resolve geographic ambiguity. However, unconstrained tool use often introduces distracting or conflicting signals. We introduce **GeoEvolve**, a framework that recursively evolves `GeoSkill`, a single geolocalization skill encoding an explicit orchestration policy. **GeoEvolve** induces an initial policy from cold-start trajectories collected under minimally constrained tool use, and then enters an on-policy loop that generates new trajectories under its current guidance. At each iteration, **GeoEvolve** assigns process-level credit to effective and harmful decisions within a trajectory batch, and uses this feedback to preserve, clarify, or revise rules for hypothesis formation, tool selection, evidence verification, and stopping. On GeoEvo-Test and Im2GPS3K, `GeoSkill` consistently improves over unconstrained tool use and remains effective across LVLMs of different scales and families. Distilling skill-guided trajectories into a task-specific model yields further gains and sets a new state-of-the-art among compact models in image geolocalization. The code, data, and skills will be publicly released.

1 Introduction

Image geolocalization, the task of inferring the geographic location of an image from visual evidence, has attracted growing interest from online gaming platforms[1] and the research community (Wilson et al., 2024; Li et al., 2026). With the emergence of large vision-language models (LVLMs), recent approaches have moved beyond relying solely on visual recognition and reasoning (Li et al., 2024; Song et al., 2025; Yerramilli et al., 2025), and increasingly leverage external tools to gather additional evidence (Zhou et al., 2024; Jia et al., 2024; 2025; Wang et al., 2025b; Ji et al., 2026) when the image itself is insufficient to resolve geographic ambiguity. This turns geolocalization from a *one-shot prediction problem* into an *iterative investigation process*, in which the model can gather, interpret, and verify evidence before making a prediction.

However, effective tool use is not simply a matter of providing an LVLM with more tools (Repantis et al., 2026). Different images often require substantially different investigation strategies. A distinctive landmark may be resolved through a small number of targeted actions, whereas an ambiguous street scene may require multiple rounds of retrieval, search, and evidence verification. This investigation involves decisions about which *hypotheses* to pursue, which *tools* to invoke, how to interpret the resulting *evidence*, and when to *stop*. The effectiveness of tool-augmented geolocalization consequently depends heavily on *how these tools are orchestrated throughout the investigation.*

Existing tool-augmented approaches (Tao et al., 2026) typically rely on the LVLM to coordinate tool use within each individual trajectory. While this can produce effective behavior within a particular investigation, the underlying orchestration strategy remains implicit in the trajectory. Recurring successes and failures are therefore difficult to consolidate into guidance for future investigations or reuse with another model. *The orchestration strategy itself can instead become the object of learning*: it can be represented outside the LVLM, refined through accumulated execution experience, and shared across models equipped with the same tools.

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

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

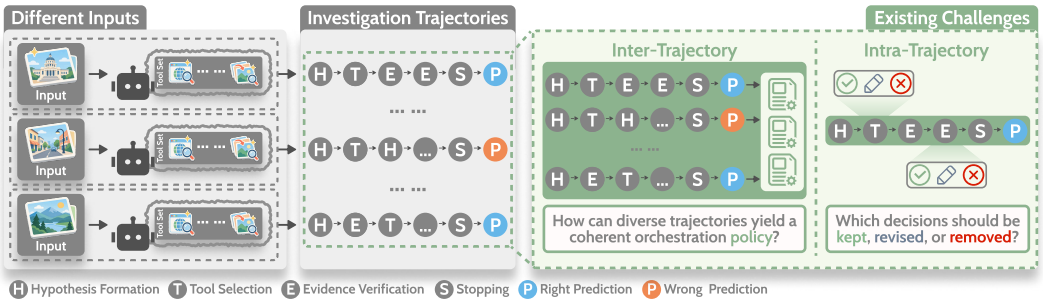


Figure 1: Motivation for **GeoEvolve**: learning an explicit orchestration policy from diverse investigation trajectories presents complementary inter- and intra-trajectory challenges.

To address these challenges, we introduce **GeoEvolve**, a framework for recursive skill evolution in image geolocalization. We represent the orchestration strategy as a geolocalization skill[2], termed `GeoSkill`. Rather than hard-coding a fixed workflow or modifying the LVLM, `GeoSkill` guides *hypothesis formation*, *tool selection*, *evidence verification*, and *stopping* throughout the investigation. This explicit representation makes the strategy inspectable and reusable across investigations.

The key to **GeoEvolve** is to treat investigation trajectories as experience for refining the skill. A trajectory contains *more than a final location prediction*: individual decisions along the investigation can determine whether the subsequent evidence becomes more informative or the investigation becomes unnecessarily costly. **GeoEvolve** therefore evaluates decisions within a trajectory batch and identifies behaviors that contribute to effective investigations as well as those that lead to ineffective or unnecessary actions. These *process-level signals* are then used to revise the corresponding skill rules, allowing the skill to accumulate experience across investigations rather than relearning how to use tools from scratch for each image.

GeoEvolve begins by inducing an initial `GeoSkill` from cold-start trajectories generated by a frozen source LVLM under minimally constrained tool use. It then repeatedly executes the current `GeoSkill` to generate on-policy trajectories and uses the resulting experience for further refinement. To determine which parts of an investigation should be retained or changed, **GeoEvolve** assigns *process-level credit* to consequential decisions within each trajectory batch. The evolution model uses this feedback to preserve well-supported rules, clarify rules that were not followed, and revise rules that remain ineffective. All updates are stored in the external skill while the source LVLM remains frozen. Once evolved, the same `GeoSkill` can directly guide other tool-augmented LVLMs with the same tool interfaces, without repeating the evolution process for each model. Beyond direct skill reuse, the trajectories generated under `GeoSkill` can also serve as training data for task-specific model adaptation. We further apply supervised fine-tuning (SFT) to these skill-guided trajectories, distilling the induced behavior into a task-specific model.

Our contributions are summarized as follows:

- **Training in Textual Policy Space.** We formulate skill evolution as a training process in which `GeoSkill` serves as a textual policy checkpoint that accumulates orchestration knowledge across evolution rounds. Stored outside the source model parameters, this checkpoint can be directly reused by other tool-augmented LVLMs with compatible tool interfaces.
- **Process-Level On-Policy Evolution.** We propose **GeoEvolve**, which initializes `GeoSkill` from cold-start trajectories and recursively refines it using on-policy experience. Process-level credit identifies consequential decisions and distinguishes execution failures from policy deficiencies, enabling targeted preservation, clarification, and revision of skill rules.
- **Cross-Model Reuse and Skill-Guided Supervision.** We find that a single `GeoSkill` evolved solely from `Qwen3.5-9B` experience can transfer to both the larger `Qwen3.5-27B` and a different-family `GLM-5.3-Flash` without re-evolution, demonstrating that the learned orchestration strategy extends beyond its source model. Moreover, `GeoSkill`-guided trajectories provide effective supervision for task-specific model training: distilling these trajectories into `GeoEvolve-9B` consistently outperforms SFT on tool-only trajectories.

[2] `https://github.com/anthropics/skills`

2 Related Work

2.1 Image Geolocalization

Image geolocalization requires interpreting diverse environmental cues, from distinctive landmarks to subtle signals (*e.g.*, vegetation, architecture, and weather). Early approaches primarily rely on *classification* (Weyand et al., 2016; Seo et al., 2018; Theiner et al., 2022; Clark et al., 2023), *retrieval* (Hays & Efros, 2008; Torii et al., 2015; Vo et al., 2017; Berton et al., 2021; Vivanco Cepeda et al., 2023; Xu et al., 2024), or *generation* (Izbicki et al., 2020; Dufour et al., 2025; Wang et al., 2025c) paradigms to map visual features to geographic coordinates. While effective within constrained settings, these data-driven methods often lack the high-level semantic reasoning required for robust open-world generalization.

LVLMs have shifted geolocalization toward knowledge-driven reasoning, with general-purpose models commonly adapted through SFT on curated instruction data and chain-of-thought supervision (Li et al., 2024; Song et al., 2025; Campos et al., 2025; Dou et al., 2024; Zhang et al., 2025). However, these models remain dependent on parametric knowledge, risking hallucinated reasoning and incorrect geographic associations (Hu et al., 2024; Wei et al., 2024; Huang et al., 2025). Recent work addresses these limitations through retrieval-augmented generation (Zhou et al., 2024; Jia et al., 2024; 2025), reinforcement learning (Li et al., 2025; Wang et al., 2025a), and multi-agent collaboration (Han et al., 2025; Zheng et al., 2025b). More recently, agentic approaches enable LVLMs to decompose scenes into verifiable cues and use external tools to gather evidence and iteratively refine geographic hypotheses (Wang et al., 2025b; Ji et al., 2026; Jia et al., 2026). Despite this progress, tool orchestration remains implicit within individual trajectories, leaving how investigation experience can be consolidated into an explicit and reusable orchestration policy underexplored.

2.2 Agentic Reasoning and Skill Evolution

Building on structured reasoning paradigms (*e.g.*, chain-of-thought (Wei et al., 2022), tree-of-thought (Yao et al., 2023), and graph-based reasoning (Besta et al., 2024)), *ReAct* (Yao et al., 2022) interleaves reasoning with external actions, inspiring agents that interact through APIs (Schick et al., 2023) and multi-agent collaboration (Hong et al., 2023; Shen et al., 2024). Recent approaches improve controllability through structured workflows and programmatic routines (Zhang et al., 2024; Hu et al., 2025; Zhu et al., 2025b). However, rigid execution graphs may limit adaptability (Topsakal & Akinci, 2023; Duan & Wang, 2024), whereas fully autonomous agents can incur inefficient trial-and-error interactions (Zhuang et al., 2023; Song et al., 2024; Wang et al., 2024a; Ding et al., 2026). Agent skills offer a middle ground by distilling interaction experience into reusable procedures that guide future reasoning and action (Wang et al., 2023; Zhao et al., 2024; Wang et al., 2024b; Zheng et al., 2025a). Recent methods (Zhang et al., 2026; Ouyang et al., 2026; Xia et al., 2026) further organize skills into external libraries and enable their autonomous evolution using successful and failed trajectories (Zhang et al., 2026; Ouyang et al., 2026; Xia et al., 2026). However, trajectory-level outcomes provide limited guidance for refining reasoning-intensive skills. We therefore use process-level credit to evolve a geolocalization skill from one LVLM's experience and reuse the same skill across tool-augmented LVLMs.

3 Methodology

We present **GeoEvolve**, a framework that recursively evolves `GeoSkill`, an explicit orchestration policy for tool-augmented image geolocalization. Figure 2 summarizes the overall framework. We cast skill evolution as training in textual policy space: trajectories from a frozen source LVLM are used to iteratively revise and save `GeoSkill` instead of updating model parameters. The resulting skill can be applied unchanged to other frozen LVLMs sharing the same tool interfaces.

3.1 Problem Formulation

Let $\mathcal{M}_\theta$ denote a frozen LVLM, instantiated as a source model $\mathcal{M}_{\theta_s}$ for skill evolution and as a target model $\mathcal{M}_{\theta_t}$ for cross-model deployment. We further denote the fixed evolution model by $\mathcal{E}$ and the external tool set by $\mathcal{T}$. At evolution round r, $\mathcal{S}^{(r)}$ denotes the current skill checkpoint. Given

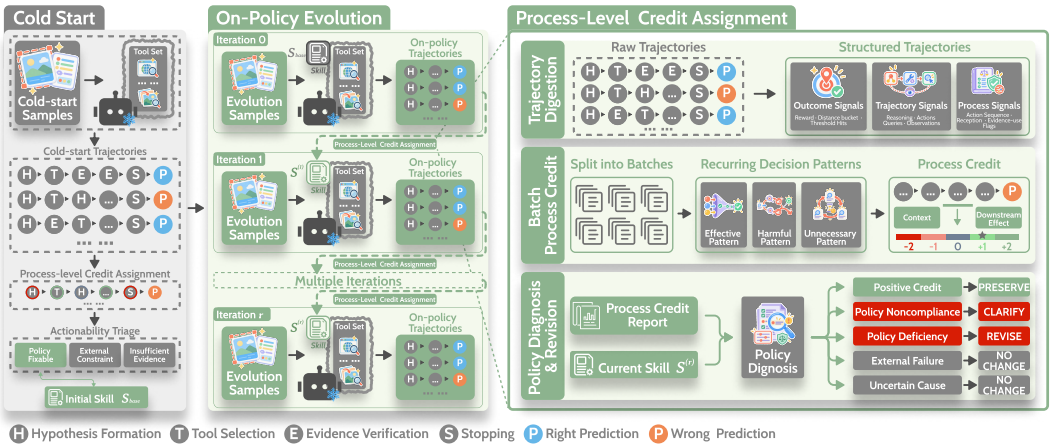


Figure 2: Overview of **GeoEvolve**. A frozen source LVLM first generates cold-start trajectories for skill initialization and then produces on-policy trajectories under the current skill for recursive refinement. Process-level credit guides both stages, and the resulting `GeoSkill` can be applied unchanged to other LVLMs sharing the same tool interfaces.

an image I, the agent generates an investigation trajectory

$$\tau = \text{Agent}\left(\mathcal{M}_\theta, I, \mathcal{T}; \mathcal{S}^{(r)}\right) = (z_1, \ldots, z_T, \hat{y}), \tag{1}$$

where $\hat{y}$ is the predicted location and z_t is the decision at step t. We represent each decision as $z_t = (r_t, a_t, q_t, o_t, \pi_t)$, where r_t, a_t, q_t, o_t, and π_t denote the reasoning, selected action, tool query, returned observation, and subsequent interpretation, respectively. For actions that do not invoke a tool, q_t and o_t are empty.

The skill $\mathcal{S}^{(r)}$ is a natural-language policy provided in the agent context. It guides hypothesis formation, tool selection, evidence verification, and stopping without modifying the LVLM parameters or tool implementations. Given a data distribution $\mathcal{D}$, skill evolution aims to improve the expected geolocalization utility

$$\max_{\mathcal{S}} \; \mathbb{E}_{(I,y)\sim\mathcal{D}} \left[\mathcal{U}\left(\text{Agent}(\mathcal{M}_{\theta_s}, I, \mathcal{T}; \mathcal{S}), y\right)\right], \tag{2}$$

where $\mathcal{U}$ measures the geolocalization quality of the resulting prediction.

Without updating model parameters, $\mathcal{E}$ revises and saves the skill as a sequence of textual policy checkpoints, $\mathcal{S}^{(0)} \to \mathcal{S}^{(1)} \to \cdots \to \mathcal{S}^{(R)} = \mathcal{S}^\star$. The initial checkpoint is constructed from skill-free trajectories, while subsequent checkpoints incorporate process feedback from source-model trajectories collected under the preceding skill. After evolution, $\mathcal{S}^\star$ is applied unchanged as $\text{Agent}(\mathcal{M}_{\theta_t}, I, \mathcal{T}; \mathcal{S}^\star)$, requiring neither target-model trajectories nor further skill updates.

3.2 Recursive Skill Evolution

GeoEvolve evolves the orchestration policy through cold-start initialization and recursive on-policy refinement. All trajectories used for evolution are generated by the frozen source LVLM $\mathcal{M}_{\theta_s}$, while $\mathcal{E}$ evaluates the resulting processes and revises the skill. The source LVLM, evolution model, and external tools remain fixed throughout this process. We use disjoint subsets $\mathcal{D}_{\text{cold}}$, $\mathcal{D}_{\text{on}}$, and $\mathcal{D}_{\text{dev}}$ for initialization, on-policy evolution, and checkpoint selection, respectively. The development set provides no feedback for skill revision.

Process-Level Feedback. A trajectory outcome provides only coarse supervision and cannot reveal the contribution of individual decisions. An incorrect prediction may contain useful intermediate decisions, while a correct prediction may rely on unsupported or unnecessary actions. We therefore compare trajectory batches to identify recurring patterns and their downstream effects.

For a trajectory batch $\mathcal{B}_k = \{(\tau_i, y_i)\}_{i\in\mathcal{I}_k}$, we summarize each outcome as $u_i = (\rho_i, b_i, \mathbf{h}_i)$, where ρ_i is a distance-based score, b_i is its error bucket, and $\mathbf{h}_i$ records correctness at the predefined

geographic thresholds. Each trajectory is converted into a process digest $\bar{\tau}_i$ containing its reasoning, tool interactions, returned observations, and final prediction. Ground-truth coordinates and place names are excluded, so $\mathcal{E}$ receives only $\bar{\mathcal{B}}_k = \{(\bar{\tau}_i, u_i)\}_{i \in \mathcal{I}_k}$.

By comparing trajectories within the batch, $\mathcal{E}$ identifies recurring patterns in hypothesis formation, tool selection, evidence verification, and stopping. For each pattern, it records its description $p_{k,j}$, decision stage $s_{k,j}$, qualitative credit $\gamma_{k,j}$, supporting trajectories $\mathcal{I}_{k,j}$, evidence summary $e_{k,j}$, and confidence $\kappa_{k,j}$:

$$\begin{aligned} \mathcal{F}_k^{\text{proc}} &= \text{Eval}_{\text{proc}}\left(\bar{\mathcal{B}}_k; \mathcal{S}_k\right) \\ &= \left\{(p_{k,j}, s_{k,j}, \gamma_{k,j}, \mathcal{I}_{k,j}, e_{k,j}, \kappa_{k,j})\right\}_{j=1}^{J_k}, \end{aligned} \tag{3}$$

where $\mathcal{S}_k$ is the skill used to generate the batch and is empty for cold-start trajectories.

We use the five-level credit scale $\gamma_{k,j} \in \{-2, -1, 0, +1, +2\}$, ranging from strong negative to strong positive contributions. Credit reflects the available evidence, downstream influence, and necessity of the identified decision pattern. Patterns with uncertain effects are excluded from skill updates. Their empirical support is measured by $n_{k,j} = |\mathcal{I}_{k,j}|$ and $\alpha_{k,j} = n_{k,j}/|\mathcal{I}_k|$. Recurring patterns are consolidated across batches, with strong evidence producing firm rules, moderate evidence producing conditional guidance, and weak evidence retained only as hypotheses or omitted. The detailed construction of process-level feedback is provided in Appendix A.2.

Cold-Start Initialization. For each image $I_i^{\text{cold}} \in \mathcal{D}_{\text{cold}}$, the source LVLM generates a trajectory without task-specific orchestration guidance: $\tau_i^{\text{cold}} = \text{Agent}(\mathcal{M}_{\theta_s}, I_i^{\text{cold}}, \mathcal{T}; \varnothing)$. The task prompt and tool interfaces remain fixed.

We apply process-level evaluation to the resulting trajectories and consolidate recurring patterns across batches. Before constructing the skill, $\mathcal{E}$ removes tool-or-environment failures and unresolved cases that cannot be addressed through the orchestration policy. Let $\mathcal{F}_{\text{cold}}^{\text{act}}$ denote the remaining actionable feedback. The initial checkpoint is then synthesized as $\mathcal{S}^{(0)} = \text{Initialize}(\mathcal{F}_{\text{cold}}^{\text{act}})$, with each rule weighted by the support of its source pattern.

On-Policy Refinement. We partition $\mathcal{D}_{\text{on}}$ into disjoint batches $\{\mathcal{D}_{\text{on}}^{(r)}\}_{r=1}^{R}$ and process them sequentially. At round r, the current checkpoint guides trajectory generation for the next batch: $\tau_i^{(r)} = \text{Agent}(\mathcal{M}_{\theta_s}, I_i^{(r)}, \mathcal{T}; \mathcal{S}^{(r-1)})$. The resulting experience therefore reflects how the source LVLM behaves under the latest skill.

Process evaluation produces $\mathcal{F}_{\text{on}}^{(r)} = \text{Eval}_{\text{proc}}(\bar{\mathcal{B}}_{\text{on}}^{(r)}; \mathcal{S}^{(r-1)})$. The evolution model $\mathcal{E}$ compares each credited pattern with the guidance already encoded in $\mathcal{S}^{(r-1)}$. Positive patterns are preserved, while negative patterns are attributed to policy non-compliance, policy deficiency, tool-or-environment failure, or insufficient evidence. Non-compliance indicates that suitable guidance exists but was not followed, so the corresponding rule may be clarified. Policy deficiency indicates that the guidance is missing, ambiguous, or ineffective, allowing a rule to be added, revised, or removed. The remaining cases do not modify the skill.

Based on these diagnoses, $\mathcal{E}$ constructs an update plan $\mathcal{P}^{(r)} = \text{Diagnose}(\mathcal{S}^{(r-1)}, \mathcal{F}_{\text{on}}^{(r)})$ and produces a candidate checkpoint $\widetilde{\mathcal{S}}^{(r)} = \text{Revise}(\mathcal{S}^{(r-1)}, \mathcal{P}^{(r)})$. Each round applies only a small number of feedback-supported edits, limiting policy drift and preserving unrelated guidance.

Checkpoint Selection. A revision that improves the current on-policy batch may not generalize. At each scheduled gate, the candidate is evaluated with the source LVLM on the held-out set $\mathcal{D}_{\text{dev}}$:

$$\mathcal{S}^{(r)} = \begin{cases} \widetilde{\mathcal{S}}^{(r)}, & \text{Gate}\left(\widetilde{\mathcal{S}}^{(r)}, \mathcal{S}^{(r-1)}; \mathcal{M}_{\theta_s}, \mathcal{D}_{\text{dev}}\right) = 1, \\ \mathcal{S}^{(r-1)}, & \text{otherwise.} \end{cases} \tag{4}$$

A candidate is accepted only when it remains non-inferior to $\mathcal{S}^{(0)}$ and improves over the latest accepted checkpoint. Several revisions may be accumulated before evaluation, with a failed gate discarding all pending changes. The complete cold-start and on-policy evolution procedure, including checkpoint selection and rollback, is summarized in Appendix A.1. The final validated `SKILL.md` is retained as $\mathcal{S}^{\star}$.

4 EXPERIMENTS

We evaluate **GeoEvolve** by evolving a single `GeoSkill` from trajectories generated by a frozen source LVLM and applying the resulting skill at inference time. We compare skill-equipped LVLMs with existing geolocalization methods and conduct controlled studies of tool access, skill evolution, cross-model transfer, on-policy scaling, and skill-guided trajectory distillation.

4.1 EXPERIMENTAL SETUP

Datasets and splits. We randomly sample 600 images from MP-16 (Larson et al., 2017) and divide them equally into a cold-start subset for skill initialization and an on-policy subset for recursive evolution. A separate development split is used only for checkpoint selection. We construct GEO-EVO-TEST by sampling another 1,000 images from the remaining MP-16 data and evaluate on both GEO-EVO-TEST and IM2GPS3K (Vo et al., 2017). All MP-16-derived splits are mutually disjoint, and evaluation images are excluded from the image-retrieval database.

Models and evolution protocol. We use `Qwen3.5-9B`[3] as the frozen source LVLM $\mathcal{M}_{\theta_s}$ for generating all cold-start and on-policy trajectories. `GLM-5.3-Flash`[4] serves as the evolution model $\mathcal{E}$, performing process-level credit assignment and skill revision from sanitized trajectory digests and outcome summaries without access to ground-truth coordinates or place names.

The source LVLM first generates skill-free trajectories for constructing $\mathcal{S}^{(0)}$ and subsequently generates on-policy trajectories under the current checkpoint. Candidate checkpoints are evaluated on the held-out development split, and the final accepted skill $\mathcal{S}^{\star}$ is fixed for benchmark evaluation. We apply the same $\mathcal{S}^{\star}$ unchanged to `Qwen3.5-9B`, `Qwen3.5-27B`, and `GLM-5.3-Flash`, without collecting additional evolution trajectories or performing model-specific skill updates. All LVLM parameters remain frozen except in the skill-distillation study.

Tool configuration. The tool set $\mathcal{T}$ consists of `Web_Search` and `Image_Retrieval`, which provide textual evidence and visually similar geotagged images, respectively. We implement `web_search` with the Tavily Search API and `image_retrieval` with a G3 encoder (Jia et al., 2024) and FAISS over 4.1M geotagged social-media images, excluding images that overlap with the evaluation sets. A more detailed description of the two tools is provided in Appendix B.2. All ablations are conducted on GEO-EVO-TEST using fixed decoding settings and identical tool interfaces.

Skill-guided distillation. To examine whether skill-guided trajectories provide effective supervision, we use `GLM-5.3-Flash` as the teacher to generate 5K trajectories under $\mathcal{S}^{\star}$ and fine-tune `Qwen3.5-9B`, yielding `GeoEvolve-9B`. As a control, the same teacher generates 5K trajectories with identical tools but without `GeoSkill`. At inference time, both models retain tool access, while `GeoEvolve-9B` also receives $\mathcal{S}^{\star}$, matching its training configuration.

Metrics. We report accuracy within 1, 25, 200, 750, and 2,500 kilometers based on the Haversine distance between predicted and reference coordinates. Missing or invalid predictions are counted as failures at all thresholds.

4.2 MAIN RESULTS

We report the main comparison on GEO-EVO-TEST and IM2GPS3K in Table 1 and analyze the results from three perspectives. Overall, the results show that the evolved skill improves agentic geolocalization beyond existing tool use, although its effectiveness depends on the base model, and that skill-guided behavior can be further strengthened through supervised fine-tuning.

Comparison with Existing Methods. We isolate the effect of `GeoSkill` by comparing each backbone without tools, with tools, and with both tools and the evolved skill. On GEO-EVO-TEST, `Qwen3.5-9B` improves monotonically across the three settings, revealing two distinct bottlenecks: tool access addresses missing geographic evidence, while `GeoSkill` improves how that evidence is acquired and integrated. For the stronger `GLM-5.3-Flash`, tools mainly improve fine-grained localization but may disrupt reliable coarse-level predictions; adding `GeoSkill` preserves the fine-grained gains while recovering the degraded performance. This shows that `GeoSkill` regulates

[3] https://huggingface.co/Qwen/Qwen3.5-9B

[4] https://huggingface.co/zai-org/GLM-5.3-Flash

Table 1: Performance comparison on GEO-EVO-TEST and IM2GPS3K (Vo et al., 2017).

Method	GEO-EVO-TEST (% @ km)					IM2GPS3K (Vo et al., 2017) (% @ km)				
	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
I. Non-LVLM Geolocalization Methods										
ISNs‡ (Muller-Budack et al., 2018)	9.30	24.10	33.70	45.50	62.30	10.50	28.00	36.60	49.70	66.00
GeoCLIP‡ (Vivanco Cepeda et al., 2023)	11.24	28.31	41.61	60.53	78.62	14.11	34.47	50.65	69.67	83.82
Translocator† (Pramanick et al., 2022)	–	–	–	–	–	11.80	31.10	46.70	58.90	80.10
RFM-YFCC (Dufour et al., 2025)	4.50	24.00	36.00	54.60	78.10	5.41	29.70	44.71	61.83	79.55
II. LVLMs w/o Tools										
InternVL3-8B (Zhu et al., 2025a)	3.02	16.21	27.49	43.10	58.81	6.44	25.69	34.57	49.38	61.66
Qwen3.5-9B (Team, 2026)	6.56	19.06	31.45	46.93	64.24	8.81	28.99	41.98	59.66	77.20
Qwen3.5-27B (Team, 2026)	9.06	22.11	32.69	50.49	70.12	12.06	34.21	46.82	63.47	78.48
Kimi-K2.6 (Moonshot AI, 2026)	18.72	40.94	52.55	65.47	78.58	21.54	52.14	64.45	76.89	87.89
GLM-5.3-Flash (Z.ai (Zhipu AI), 2026)	17.28	41.82	54.19	<u>70.35</u>	<u>83.84</u>	20.50	<u>52.63</u>	**66.99**	<u>80.58</u>	**90.83**
GeoReasoner-7B (Li et al., 2024)	4.60	21.00	27.20	40.90	58.10	7.67	26.94	36.63	52.27	65.39
GLOBE-7B (Li et al., 2025)	6.82	26.80	42.38	61.10	79.53	9.84	40.18	56.19	71.45	82.38
III. LVLMs w/ Tools										
Qwen3.5-9B (Team, 2026)	14.61	31.01	43.70	51.45	68.80	–	–	–	–	–
GLM-5.3-Flash (Z.ai (Zhipu AI), 2026)	<u>24.95</u>	<u>47.39</u>	<u>55.08</u>	68.75	83.51	–	–	–	–	–
G3 (Jia et al., 2024)	16.90	34.00	44.40	58.30	75.10	14.31	35.87	49.42	66.93	81.78
GeoVista (Wang et al., 2025b)	5.59	22.47	30.32	45.38	62.69	–	–	–	–	–
SpotAgent (Jia et al., 2026)	10.70	31.00	46.30	62.00	76.00	14.12	40.36	57.80	73.43	85.75
Qwen3.5-9B + `GeoSkill` (**Ours**)	20.40	36.32	47.62	57.95	74.53	20.10	38.52	57.88	74.61	85.10
GLM-5.3-Flash + `GeoSkill` (**Ours**)	**25.34**	**47.51**	**55.89**	**72.32**	**84.56**	**24.97**	**52.78**	<u>66.29</u>	**81.24**	<u>90.70</u>
`GeoEvolve-9B` + `GeoSkill` (**Ours**)	22.87 (+2.47)	41.92 (+5.60)	51.75 (+4.13)	63.63 (+5.68)	78.05 (+3.52)	<u>22.54</u> (+2.44)	43.00 (+4.48)	59.08 (+1.20)	74.72 (+0.11)	86.30 (+1.20)

† denotes methods that are not publicly available. ‡ denotes training overlap with GEO-EVO-TEST.
The same `GeoSkill`, evolved from `Qwen3.5-9B` trajectories, is applied unchanged across LVLMs. For `GeoEvolve-9B + GeoSkill`, parenthesized values denote percentage-point improvements over `Qwen3.5-9B + GeoSkill`. Best and second-best results are shown in **bold** and <u>underline</u>, respectively.

rather than merely encourages tool use. The different trends across models and benchmarks further suggest that `GeoSkill` transfers a common investigation strategy without enforcing identical behavior: its benefit depends on how each LVLM interprets and executes the same guidance.

Cross-Model Skill Transfer. We evaluate `GeoSkill` with two different LVLM backbones. `Qwen3.5-9B` + `GeoSkill` (Ours) uses the same backbone that generates the trajectories for skill evolution, together with the evolved `GeoSkill`. In contrast, `GLM-5.3-Flash` + `GeoSkill` (Ours) applies the same evolved `GeoSkill` to a different backbone without any modification or re-evolution of the skill. The consistent improvements across the two configurations and different geographic scales indicate that `GeoSkill` encodes an orchestration strategy that can be transferred across LVLMs. The magnitude of the gains, however, varies across backbones, reflecting differences in how each model follows and executes the shared guidance. We further analyze the effect of the base model and the transferability of `GeoSkill` in the ablation study.

Skill-Guided Model Training. We further distill the behavior induced by `GeoSkill` into `GeoEvolve-9B`, a task-specific model trained on skill-guided trajectories. With the same `GeoSkill` applied at inference time, `GeoEvolve-9B` + `GeoSkill` consistently improves over `Qwen3.5-9B` + `GeoSkill` across both benchmarks and geographic scales. This demonstrates that the behavior induced by `GeoSkill` can be further leveraged through model training.

We additionally analyze inference-time efficiency in Appendix C, focusing on interaction cost, tool usage, and runtime across different skill configurations.

4.3 ABLATION STUDIES

To disentangle the contributions of external tools, skill initialization, and recursive evolution, we conduct four controlled studies covering tool access and evolution stages, cross-model transfer, on-policy scaling, and skill-guided fine-tuning. All ablations are conducted on GEO-EVO-TEST using identical evaluation samples, decoding settings, and tool interfaces.

Contributions of Tools and Skill Evolution. Table 2 disentangles the effects of tools and the two stages of skill evolution. `Image_Retrieval` provides stronger evidence than `Web_Search`, while combining both tools performs worse than image retrieval alone. This shows that tools are not automatically complementary and may introduce distracting or conflicting evidence without explicit orchestration. The cold-start skill reverses this degradation across all thresholds, demonstrating that an initial policy helps coordinate multiple evidence sources. On-policy evolution provides further

Table 2: Ablation of tools and skill evolution on GEO-EVO-TEST with `Qwen3.5-9B`.

Setting	Tool Access		Skill Evolution		Accuracy (% @ km)				
	Web Search	Image Retrieval	Cold Start	On-Policy	Street 1	City 25	Region 200	Country 750	Continent 2500
No Tools	–	–	–	–	6.56	19.06	31.45	46.93	64.24
Web Search Only	✓	–	–	–	7.70	21.41	34.11	45.16	61.94
Image Retrieval Only	–	✓	–	–	17.32	32.60	44.12	52.73	69.68
Both Tools	✓	✓	–	–	14.61	31.01	43.70	51.47	68.81
+ Cold-Start	✓	✓	✓	–	18.94	33.91	45.76	54.49	70.28
+ On-Policy Only	✓	✓	–	✓	<u>19.12</u>	<u>34.96</u>	<u>47.51</u>	<u>55.52</u>	<u>71.94</u>
+ `GeoSkill`	✓	✓	✓	✓	**20.40**	**36.32**	**47.62**	**57.95**	**74.53**

On-Policy Only evolves the skill without cold-start initialization, whereas *Full* `GeoSkill` uses both stages. Best and second-best accuracy results are shown in **bold** and <u>underline</u>, respectively.

Table 3: Effect of `GeoSkill` across different base models on GEO-EVO-TEST.

Base Model	Tools	`GeoSkill`	Accuracy (% @ km)				
			Street 1	City 25	Region 200	Country 750	Continent 2500
Qwen3.5-9B	–	–	6.56	19.06	31.45	46.93	64.24
	✓	–	<u>14.61</u>	<u>31.01</u>	<u>43.70</u>	<u>51.45</u>	<u>68.80</u>
	✓	✓	**20.40**	**36.32**	**47.62**	**57.95**	**74.53**
Qwen3.5-27B	–	–	9.06	22.11	32.69	50.49	70.12
	✓	–	<u>18.38</u>	<u>35.72</u>	<u>47.27</u>	<u>58.64</u>	<u>75.49</u>
	✓	✓	**21.06**	**37.16**	**48.54**	**60.04**	**76.79**
GLM-5.3-Flash	–	–	17.28	41.82	54.19	<u>70.35</u>	<u>83.84</u>
	✓	–	<u>24.95</u>	<u>47.39</u>	<u>55.08</u>	68.75	83.51
	✓	✓	**25.34**	**47.51**	**55.89**	**72.32**	**84.56**

Tools denotes access to both `Web Search` and `Image Retrieval`. Each `GeoSkill` is evolved separately for its corresponding base model. Best and second-best results within each base model are shown in **bold** and <u>underline</u>, respectively.

gains, while the full `GeoSkill` consistently outperforms evolution without cold-start initialization, confirming that cold-start experience provides a policy foundation.

Effectiveness across Base Models. We evaluate the same `GeoSkill`, evolved from `Qwen3.5-9B` trajectories, on three LVLMs with different scales and model families in Table 3. Across all backbones, combining tools with `GeoSkill` consistently outperforms using either the LVLM alone or the LVLM with tools, showing that the learned orchestration policy transfers beyond its source model. The larger gains for `Qwen3.5-9B` suggest that explicit orchestration can compensate for limited base-model capacity. For `GLM-5.3-Flash`, unconstrained tool use improves fine-grained localization but can disrupt reliable coarse predictions, whereas `GeoSkill` preserves the benefits of external evidence across geographic scales. These results show that a fixed skill can transfer across LVLMs, although its realized benefit depends on how each model executes the same guidance.

Scaling with On-Policy Trajectories. Figure 3 explains why **`GeoEvolve`** uses a compact on-policy evolution set. Performance improves as on-policy experience increases and reaches its overall optimum with 192 trajectories, but further scaling provides no sustained benefit. Because skill evolution consolidates recurring behaviors into a textual policy, additional trajectories may increasingly repeat existing evidence or introduce conflicting revisions after the main orchestration patterns have been covered. Evolution therefore depends more on the diversity and causal value of its experience than on volume alone, supporting the use of a modest on-policy set.

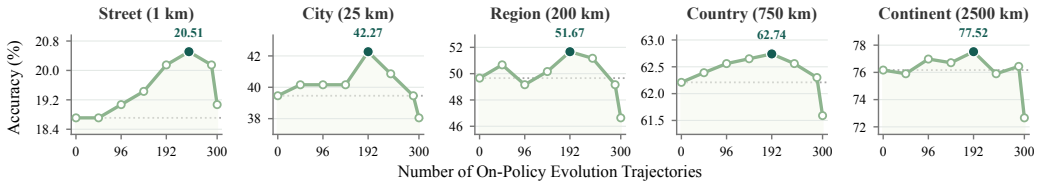


Figure 3: Scaling with on-policy trajectories on GEO-EVO-TEST. Zero denotes the cold-start skill.

Distilling Skill-Guided Behavior. To evaluate whether `GeoSkill` can also benefit model training, Table 4 compares SFT on tool-only and `GeoSkill`-guided trajectories. Tool-only trajectory SFT already improves over direct tool use, showing that agentic trajectories provide useful training supervision. More importantly, `GeoEvolve-9B + GeoSkill` consistently outperforms SFT on tool-only trajectories across all geographic scales, suggesting that `GeoSkill`-guided trajectories provide stronger task-specific supervision. Its gains over `Qwen3.5-9B + GeoSkill` further

Table 4: Evaluation of skill guidance and trajectory fine-tuning on GEO-EVO-TEST.

Setting	Distilled Trajectories	Inference w/ Tools	Inference w/ `GeoSkill`	Accuracy (% @ km) Street 1	City 25	Region 200	Country 750	Continent 2500
Tool-Free	–	–	–	6.56	19.06	31.45	46.93	64.24
Tool-Augmented	–	✓	–	14.61	31.01	43.70	51.45	68.80
`GeoSkill`	–	✓	✓	<u>20.40</u>	36.32	47.62	57.95	74.53
SFT on Tool-Only Trajectories	Tool-only	✓	–	19.78	<u>39.20</u>	<u>48.89</u>	<u>60.10</u>	<u>76.15</u>
`GeoEvolve-9B`	`GeoSkill`-guided	✓	✓	**22.87**	**41.92**	**51.75**	**63.63**	**78.05**

Tools include Web Search and Image Retrieval; trajectory labels denote SFT data, and only `GeoEvolve-9B` retains `GeoSkill` at inference. Best and second-best results are shown in **bold** and <u>underline</u>, respectively.

confirm that fine-tuning contributes beyond inference-time guidance alone. These results demonstrate that `GeoSkill` provides both an effective orchestration policy and a source of training data for building a stronger task-specific model. We use a consistent SFT configuration across all settings; detailed hyperparameters are provided in Appendix B.1.

5 DISCUSSION

From Textual Guidance to Model Training. Skill evolution can be viewed as training in textual policy space. In conventional model training, data provide experience, an objective supplies supervision, and an optimizer updates parameters. In **`GeoEvolve`**, investigation trajectories provide data, process-level credit supplies supervision, and the evolution model revises `SKILL.md`. Each accepted skill initializes the next round, with checkpoint selection serving a role similar to validation. The resulting checkpoint takes a different form: numerical parameters in model training and an explicit natural-language policy in skill evolution.

The same approach could extend beyond image geolocalization to tasks where agents make repeated tool-use decisions and learn reusable strategies from their outcomes. The resulting skill can be inspected, revised, and shared across compatible agents. Skill-guided trajectories can also support model training, linking textual policy evolution with parameter optimization.

Limitations and Future Directions. Although the evolved `GeoSkill` transfers across the LVLMs studied here, all evolution trajectories are generated by a single source model, which may bias the learned policy toward patterns especially useful to `Qwen3.5-9B`. It remains to be studied whether trajectories from multiple source models can produce a skill that transfers more uniformly across models with different capabilities.

The quality of skill evolution depends on the causal judgments of the evolution model. Although process-level feedback provides more informative supervision than assigning the final outcome to every decision, evaluators may still misidentify the actions responsible for success or failure, motivating more reliable evaluation through multiple evaluators or counterfactual comparisons.

Finally, our scaling results show that additional on-policy trajectories do not guarantee further improvement. As the skill grows, new feedback may become redundant or conflict with existing guidance, suggesting the need for more selective experience collection.

6 CONCLUSION

We introduced **`GeoEvolve`**, a framework that recursively evolves `GeoSkill` for tool-augmented image geolocalization while keeping the underlying LVLM fixed. **`GeoEvolve`** initializes `GeoSkill` from cold-start trajectories and iteratively refines it through process-level credit derived from on-policy experience. Experiments on MP16-EVO-TEST and IM2GPS3K demonstrate that the evolved `GeoSkill` improves geolocalization performance across different LVLM backbones and geographic scales. We further show that the behavior induced by `GeoSkill` can be distilled into training trajectories for a task-specific model, yielding further gains in geolocalization performance. Overall, **`GeoEvolve`** provides a complementary paradigm for improving LVLM-based task performance through recursive skill evolution, enabling skill transfer across model backbones and skill-guided model training.

AI USE STATEMENT

In this work, we used generative AI tools to polish author-written text and improve the clarity of the manuscript. Generative AI models were also used as components of **GeoEvolve** to generate investigation trajectories, provide process-level feedback, and revise `GeoSkill`, as described in the paper. The authors reviewed and revised all AI-assisted text, verified the reported experimental results, and take responsibility for the final content of this work, including its text, claims, and artifacts.

ETHICS STATEMENT

Image geolocalization can reveal where a photograph was taken even when its location is not explicitly provided, creating potential privacy and surveillance risks. Our experiments use existing benchmarks to evaluate geolocalization performance and do not seek to identify or track individuals. When releasing research artifacts, we will respect the source datasets' usage terms and review images and location metadata for sensitive personal information.

REPRODUCIBILITY STATEMENT

The Method and Experiments sections describe the skill evolution procedure, dataset splits, tool settings, and evaluation protocol. We will release the code, data, and evolved skills upon acceptance to facilitate independent reproduction of our results.

REFERENCES

Gabriele Berton, Carlo Masone, Valerio Paolicelli, and Barbara Caputo. Viewpoint invariant dense matching for visual geolocalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12169–12178, 2021.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Ron Campos, Ashmal Vayani, Parth Parag Kulkarni, Rohit Gupta, Aritra Dutta, and Mubarak Shah. Gaea: A geolocation aware conversational model. *arXiv preprint arXiv:2503.16423*, 2025.

Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23182–23190, 2023.

Wenxuan Ding, Nicholas Tomlin, and Greg Durrett. Calibrate-then-act: Cost-aware exploration in llm agents. *arXiv preprint arXiv:2602.16699*, 2026.

Zhiyang Dou, Zipeng Wang, Xumeng Han, Guorong Li, Zhipei Huang, and Zhenjun Han. Gaga: Towards interactive global geolocation assistant. *arXiv preprint arXiv:2412.08907*, 2024.

Zhihua Duan and Jialin Wang. Exploration of llm multi-agent application implementation based on langgraph+ crewai. *arXiv preprint arXiv:2411.18241*, 2024.

Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. Around the world in 80 timesteps: A generative approach to global visual geolocation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23016–23026, 2025.

Xiao Han, Chen Zhu, Hengshu Zhu, and Xiangyu Zhao. 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*, pp. 814–825, 2025.

James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2008.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*, 2023.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6953–6975, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Mike Izbicki, Evangelos E Papalexakis, and Vassilis J Tsotras. Exploiting the earth's spherical geometry to geolocate images. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pp. 3–19. Springer, 2020.

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

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

Pengyue Jia, Yiding Liu, Xiaopeng Li, Xiangyu Zhao, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, and Dawei Yin. G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. *Advances in Neural Information Processing Systems*, 37:53198–53221, 2024.

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

Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE MultiMedia*, 24 (1):93–96, 2017.

Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Forty-first International Conference on Machine Learning*, 2024.

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

Ling Li, Yutian Jiang, Ziqian Mo, Qihan Yu, Na Di, Yao Zhou, Xixuan Hao, Hill Zhang, Xinhu Zheng, Xinlei He, et al. A survey on visual geolocalization: Taxonomy, progress, and prospects. 2026.

Moonshot AI. Kimi k2.6 model card. Hugging Face, 2026. URL `https://huggingface.co/moonshotai/Kimi-K2.6`.

Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 563–579, 2018.

Siru Ouyang, Jun Yan, I Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, et al. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, volume 2026, pp. 94327–94354, 2026.

Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. In *European Conference on Computer Vision*, pp. 196–215. Springer, 2022.

Vyzantinos Repantis, Ameya Gawde, Harshvardhan Singh, and Joey Blackwell II. How many tools should an llm agent see? a chance-corrected answer. *arXiv preprint arXiv:2605.24660*, 2026.

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

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

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. Small llms are weak tool learners: A multi-llm agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16658–16680, 2024.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7584–7600, 2024.

Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet, Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and Xiuying Chen. Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework. *arXiv preprint arXiv:2502.13759*, 2025.

Xijia Tao, Teng Yihua, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. Mmsearch-plus: Benchmarking provenance-aware search for multimodal browsing agents. In *International Conference on Learning Representations*, volume 2026, pp. 70385–70422, 2026.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable semantic photo geolocation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 750–760, 2022.

Oguzhan Topsakal and Tahir Cetin Akinci. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International conference on applied engineering and natural sciences*, volume 1, pp. 1050–1056, 2023.

Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1808–1817, 2015.

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

Nam Vo, Nathan Jacobs, and James Hays. Revisiting IM2GPS in the Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2621–2630, 2017.

Boshi Wang, Hao Fang, Jason Eisner, Benjamin Van Durme, and Yu Su. Llms in the imaginarium: tool learning through simulated trial and error. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10583–10604, 2024a.

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

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

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

Zhangyu Wang, Jielu Zhang, Zhongliang Zhou, Qian Cao, Nemin Wu, Zeping Liu, Lan Mu, Yang Song, Yiqun Xie, Ni Lao, et al. Locdiffusion: Identifying locations on earth by diffusing in the hilbert space. *arXiv preprint arXiv:2503.18142*, 2025c.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. Measuring and reducing llm hallucination without gold-standard answers. *arXiv preprint arXiv:2402.10412*, 2024.

Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 37–55. Springer, 2016.

Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Image and object geo-localization. *International Journal of Computer Vision*, 132(4):1350–1392, 2024.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

Shixiong Xu, Chenghao Zhang, Lubin Fan, Gaofeng Meng, Shiming Xiang, and Jieping Ye. Addressclip: Empowering vision-language models for city-wide image address localization. In *European Conference on Computer Vision*, pp. 76–92. Springer, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Sahiti Yerramilli, Nilay Pande, Rynaa Grover, and Jayant Sravan Tamarapalli. Geochain: Multimodal chain-of-thought for geographic reasoning. In *EMNLP (Findings)*, pp. 23624–23639, 2025.

Z.ai (Zhipu AI). Glm-5.3-flash model card. Hugging Face, 2026. URL `https://huggingface.co/zai-org/GLM-5.3-Flash`.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*, 2024.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, volume 2026, pp. 86069–86100, 2026.

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

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

Boyuan Zheng, Michael Y Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. Skillweaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025a.

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

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Runchuan Zhu, Bowen Jiang, Lingrui Mei, Fangkai Yang, Lu Wang, Haoxiang Gao, Fengshuo Bai, Pu Zhao, Qingwei Lin, Saravan Rajmohan, et al. Adaptflow: Adaptive workflow optimization via meta-learning. *arXiv preprint arXiv:2508.08053*, 2025b.

Yuchen Zhuang, Xiang Chen, Tong Yu, Saayan Mitra, Victor Bursztyn, Ryan A Rossi, Somdeb Sarkhel, and Chao Zhang. Toolchain*: Efficient action space navigation in large language models with a* search. *arXiv preprint arXiv:2310.13227*, 2023.

A Detailed Algorithms

A.1 Skill Evolution Procedure

Algorithm A.1: Cold-Start Initialization and Recursive On-Policy Skill Evolution

Require: Frozen rollout model $\mathcal{M}_\theta$; evolution model E; fixed tools $\mathcal{T}$; cold-start data $\mathcal{D}_{\text{cold}}$; on-policy batches $\{\mathcal{D}_{\text{on}}^{(r)}\}_{r=1}^{R}$; development set $\mathcal{D}_{\text{dev}}$; gate interval G
Ensure: Final accepted skill checkpoint $\mathcal{S}_{\text{acc}}$

Stage I: Cold-start initialization

$\mathcal{B}_{\text{cold}} \leftarrow \varnothing$
for all $(I_i, y_i) \in \mathcal{D}_{\text{cold}}$ **do**
  $\tau_i \leftarrow \textsc{Agent}(\mathcal{M}_\theta, I_i, \mathcal{T}; \varnothing)$
  $\mathcal{B}_{\text{cold}} \leftarrow \mathcal{B}_{\text{cold}} \cup \{(\tau_i, y_i)\}$
end for
$\mathcal{F}_{\text{cold}}^{\text{proc}} \leftarrow \textsc{ProcessCredit}(\mathcal{B}_{\text{cold}}, \varnothing, E)$
$(\mathcal{F}_{\text{cold}}^{\text{act}}, \mathcal{F}_{\text{cold}}^{\text{excl}}) \leftarrow \textsc{Triage}(\mathcal{F}_{\text{cold}}^{\text{proc}}, E)$
$\mathcal{S}^{(0)} \leftarrow \textsc{Initialize}(\mathcal{F}_{\text{cold}}^{\text{act}}, E)$
$\mathcal{S}_{\text{base}} \leftarrow \mathcal{S}^{(0)}$
$\mathcal{S}_{\text{acc}} \leftarrow \mathcal{S}^{(0)}$
$\mathcal{S}_{\text{pend}} \leftarrow \mathcal{S}^{(0)}$
$V_{\text{base}} \leftarrow \textsc{Evaluate}(\mathcal{S}_{\text{base}}, \mathcal{D}_{\text{dev}})$
$V_{\text{acc}} \leftarrow V_{\text{base}}$

Stage II: Recursive on-policy refinement

for $r = 1, \ldots, R$ **do**
  $\mathcal{B}_{\text{on}}^{(r)} \leftarrow \varnothing$
  for all $(I_i, y_i) \in \mathcal{D}_{\text{on}}^{(r)}$ **do**
    $\tau_i^{(r)} \leftarrow \textsc{Agent}(\mathcal{M}_\theta, I_i, \mathcal{T}; \mathcal{S}_{\text{pend}})$
    $\mathcal{B}_{\text{on}}^{(r)} \leftarrow \mathcal{B}_{\text{on}}^{(r)} \cup \{(\tau_i^{(r)}, y_i)\}$
  end for
  $\mathcal{F}_{\text{on}}^{(r)} \leftarrow \textsc{ProcessCredit}(\mathcal{B}_{\text{on}}^{(r)}, \mathcal{S}_{\text{pend}}, E)$
  $\mathcal{P}^{(r)} \leftarrow \textsc{Diagnose}(\mathcal{S}_{\text{pend}}, \mathcal{F}_{\text{on}}^{(r)}, E)$
  if $\mathcal{P}^{(r)}$ contains a supported edit **then**
    $\widetilde{\mathcal{S}}^{(r)} \leftarrow \textsc{Revise}(\mathcal{S}_{\text{pend}}, \mathcal{P}^{(r)}, E)$
  else
    $\widetilde{\mathcal{S}}^{(r)} \leftarrow \mathcal{S}_{\text{pend}}$
  end if
  $\mathcal{S}_{\text{pend}} \leftarrow \widetilde{\mathcal{S}}^{(r)}$
  if $r \bmod G = 0$ **or** $r = R$ **then**
    $V_{\text{cand}} \leftarrow \textsc{Evaluate}(\mathcal{S}_{\text{pend}}, \mathcal{D}_{\text{dev}})$
    if $\textsc{Gate}(V_{\text{cand}}, V_{\text{base}}, V_{\text{acc}}) = 1$ **then**
      $\mathcal{S}_{\text{acc}} \leftarrow \mathcal{S}_{\text{pend}}$
      $V_{\text{acc}} \leftarrow V_{\text{cand}}$
    else
      $\mathcal{S}_{\text{pend}} \leftarrow \mathcal{S}_{\text{acc}}$ ▷ Roll back pending revisions
    end if
  end if
end for
return $\mathcal{S}_{\text{acc}}$

A.2 PROCESS-LEVEL CREDIT ASSIGNMENT

Algorithm A.2: Process-Level Credit Assignment

Require: Trajectory batch $\mathcal{B}$; current skill $\mathcal{S}$; evolution model E
Ensure: Process feedback $\mathcal{F}^{\text{proc}}$

$\bar{\mathcal{B}} \leftarrow \varnothing$
for all $(\tau_i, y_i) \in \mathcal{B}$ **do**
  $\hat{y}_i \leftarrow \text{FINALPREDICTION}(\tau_i)$
  $u_i \leftarrow \text{OUTCOME}(\hat{y}_i, y_i)$
  $\bar{\tau}_i \leftarrow \text{DIGEST}(\tau_i)$
  $\bar{\mathcal{B}} \leftarrow \bar{\mathcal{B}} \cup \{(\bar{\tau}_i, u_i)\}$
end for
$\mathcal{F}^{\text{proc}} \leftarrow \text{EVALPROC}(\bar{\mathcal{B}}, \mathcal{S}, E)$
$\mathcal{F}^{\text{proc}} \leftarrow \text{ANNOTATESUPPORT}(\mathcal{F}^{\text{proc}}, \bar{\mathcal{B}})$
return $\mathcal{F}^{\text{proc}}$

B IMPLEMENTATION DETAILS

B.1 SFT CONFIGURATION

We perform supervised fine-tuning on skill-guided trajectories to distill the evolved behavior into a task-specific model. The 5K SFT data are randomly sampled from MP16 Larson et al. (2017) and are disjoint from the image retrieval database as well as from all evolution, development, and test splits, so that no SFT image is used as retrieval evidence or as an evaluation sample. The tool-only control set is generated by the same teacher on the identical 5K queries, differing only in whether `GeoSkill` is provided, so that the two SFT settings differ solely in supervision signal. All SFT experiments use the same training configuration, with the key hyperparameters summarized in Table 5, including the learning rate, batch size, training epochs, warmup ratio, and maximum sequence length.

Table 5: Hyperparameter settings for supervised fine-tuning.

Hyper Params	Value
Tuning Strategy	LoRA
LoRA Rank r	16
LoRA Alpha α	32
Learning Rate	1e-4
LR Scheduler	cosine
Warmup Ratio	0.03
Total Batch Size	16
Training Epochs	3
Training Samples	5K
Weight Decay	0.1
Optimizer	AdamW
Adam Beta1	0.9
Adam Beta2	0.95
Gradient Norm Clip	1.0
Model Max Length	32768
Precision	bf16

Table 6: **Inference-time efficiency on GEO-EVO-TEST with `Qwen3.5-9B`.** The settings follow the ablation of tools and skill evolution in Table 2. We report interaction cost, tool invocation patterns, end-to-end latency, and street-level accuracy.

Setting	Interaction		Tool Invocation			Latency (s)			Street
	Rounds	Calls	Use Rate	IR	WS	Total	Median	Model	1km
Tool access only (no `GeoSkill`)									
No Tools	1.00	0.00	0%	–	–	27.55	13.86	27.55	6.56
Web Search Only	1.54	0.54	32%	–	0.54	16.15	8.99	14.90	7.70
Image Retrieval Only	1.80	0.80	77%	0.80	–	17.98	16.29	6.40	17.32
Both Tools	2.14	1.14	70%	0.71	0.43	23.66	17.37	12.24	14.61
Tool access with skill evolution									
+ Cold-Start	3.56	2.56	89%	1.21	1.35	42.49	29.01	17.30	18.94
+ On-Policy Only	<u>3.17</u>	<u>2.17</u>	89%	1.08	<u>1.09</u>	<u>39.30</u>	<u>27.34</u>	18.78	<u>19.12</u>
+ `GeoSkill`	**2.79**	**1.79**	**96%**	1.17	**0.61**	**38.08**	**25.65**	**15.44**	**20.40**
Δ vs. + On-Policy Only	−12.0%	−17.5%	+7 pt	–	−44.0%	−3.1%	−6.2%	−17.8%	+1.28 pt
Δ vs. + Cold-Start	−21.6%	−30.1%	+7 pt	–	−54.8%	−10.4%	−11.6%	−10.8%	+1.46 pt

IR and *WS* denote the average numbers of `Image_Retrieval` and `Web_Search` calls per image, and *Use Rate* is the fraction of images with at least one tool call. *Street 1km* denotes the accuracy (%) within 1 km under the same evaluation protocol as the main ablation. Efficiency values are means over the full run; median latency is additionally reported due to the heavy-tailed latency distribution. Within the skill-evolution settings, best and second-best results are shown in **bold** and <u>underline</u>, respectively.

Table 7: **Efficiency along the skill-evolution recipe on GEO-EVO-TEST.** On-policy evolution initialized from cold-start reduces interaction rounds, tool calls, and web-search usage while maintaining a high tool-use rate, improving runtime, and attaining the best street-level accuracy.

Metric	+ Cold-Start	+ On-Policy Only	+ `GeoSkill`
Rounds / image	3.56	<u>3.17</u>	**2.79**
Tool calls / image	2.56	<u>2.17</u>	**1.79**
`Image_Retrieval`	1.21	1.08	1.17
`Web_Search`	1.35	<u>1.09</u>	**0.61**
Tool-use rate	89%	89%	**96%**
Max rounds observed	24	<u>18</u>	**13**
Total latency (s)	42.49	<u>39.30</u>	**38.08**
median	29.01	<u>27.34</u>	**25.65**
p90	89.41	<u>71.49</u>	**63.32**
Model latency (s)	<u>17.30</u>	18.78	**15.44**
Completed runs	90%	90%	**96%**
Accuracy @ 1km (%)	18.94	<u>19.12</u>	**20.40**

All settings use both tools and correspond to the three skill-evolution settings in Table 6. Best and second-best results are shown in **bold** and <u>underline</u>, respectively.

B.2 WEB SEARCH AND IMAGE RETRIEVAL

Web Search. We implement web search using the Tavily Search API[5], which provides web search results and their associated page content to the LVLM. The search tool is used to retrieve external textual evidence during image geolocalization.

Image Retrieval. For image retrieval, we use the visual tower of G3 Jia et al. (2024) to encode the query image and retrieve visually similar images from the MP16 dataset Larson et al. (2017). The retrieval database is constructed from images that are disjoint from all training, evolution, development, and test splits used in our experiments. In particular, no image appearing in the retrieval database is used as an input or evaluation sample in these splits, preventing direct overlap between the retrieval corpus and the datasets used for skill evolution and evaluation. The retrieved images are provided to the LVLM as visual evidence during the investigation.

C EFFICIENCY AND RUNTIME ANALYSIS

We further analyze inference-time efficiency on GEO-EVO-TEST with a `Qwen3.5-9B` backbone, using the same settings as the ablation of tools and skill evolution in Table 2. Table 6 reports

[5] `https://www.tavily.com/`

Table 8: **Cost per successful street-level localization on GEO-EVO-TEST.** Interaction cost and latency are normalized by the number of images localized within 1 km.

Setting	Street 1km (%)	Calls / correct	Latency (s) / correct
+ Cold-Start	18.94	13.52	224.3
+ On-Policy Only	<u>19.12</u>	<u>11.35</u>	<u>205.5</u>
+ `GeoSkill`	**20.40**	**8.77**	**186.7**

Normalized cost is computed as the mean tool calls or total latency per image divided by the street-level accuracy. Best and second-best results are shown in **bold** and <u>underline</u>, respectively.

interaction cost, tool invocation patterns, end-to-end latency, and street-level accuracy. The results show a consistent reduction in interaction cost along the evolution recipe. Starting from the cold-start skill, on-policy evolution reduces the average number of rounds from 3.56 to 2.79 and tool calls from 2.56 to 1.79 per image, while street-level accuracy improves from 18.94% to 20.40%. This reduction is primarily driven by web search, whose invocation count decreases from 1.35 to 0.61, while image-retrieval usage remains relatively stable. Meanwhile, the final configuration triggers at least one tool call on 96% of images, compared with 89% for both the cold-start-only and on-policy-only settings. Thus, the reduced interaction cost does not simply result from suppressing tool use, but from more selective evidence acquisition.

The reduction in interaction cost is also reflected in runtime. As shown in Table 7, total latency decreases from 42.49s for the cold-start skill to 38.08s for the full `GeoSkill`. The median latency decreases from 29.01s to 25.65s, while the p90 latency decreases from 89.41s to 63.32s, indicating improvements in both typical and tail latency. Model latency is also lower for the final configuration, from 17.30s and 18.78s to 15.44s, suggesting that the runtime reduction is not solely due to reduced tool-service time. Finally, Table 8 normalizes interaction cost and latency by successful localizations within 1 km. The full `GeoSkill` requires 8.77 tool calls and 186.7s per correctly localized image, compared with 13.52 calls and 224.3s for the cold-start skill, indicating that its efficiency gains are preserved, and even amplified, when localization accuracy is taken into account.