GeoVista: Web-Augmented Agentic Visual Reasoning

Anonymous Authors[1]

Abstract

Agentic visual reasoning has advanced multimodal understanding through multi-step tool use, yet current methods restrict these tools to image manipulation and never reach for external knowledge, leaving open-source models brittle on tasks that require both. We study this setting through geolocalization, a task that inherently couples fine-grained visual inspection with open-world knowledge retrieval. We introduce **GeoVista**, a 7B agentic model that interleaves image zoom-in and web-search tool calls within a reasoning loop, and, to move beyond the low resolution of prior geolocalization benchmarks, curate **GeoBench**, a benchmark of high-resolution photos, panoramas, and satellite imagery whose geographic cues are only resolvable through fine-grained visual search. We train GeoVista in two stages: a cold-start SFT on synthesized agentic trajectories to instill the tool-use format, followed by RL with **CasPO** (Cascaded Policy Optimization). Instead of treating each prediction as a pass-or-fail event, CasPO cascades credit along the country, province, and city hierarchy, yielding a dense, multi-scale learning signal. Beyond accuracy, CasPO reshapes reasoning behavior: fewer tool-call errors, more targeted searches, and emergent self-reflection. GeoVista achieves state-of-the-art among open-source models and matches proprietary systems such as Gemini-2.5-flash and GPT-5 across most metrics, despite being an order of magnitude smaller.

1. Introduction

Recent Vision-Language Models (VLMs) (Wang et al., 2024a; Wu et al., 2024; Chen et al., 2025; OpenAI et al., 2024; Team, 2025; Liu et al., 2023) have evolved multimodal reasoning from one-turn queries into agentic loops that "think with images" by interleaving textual chain-of-thought (Wei et al., 2022) with image-centric tools (Shao et al., 2024a; Hu et al., 2024) to tackle queries requiring fine-grained perception (Wu & Xie, 2023). The OpenAI o3 model (OpenAI, 2025b), alongside a line of closed-source milestones (Gao et al., 2025; Zhang et al., 2025b; Wang et al., 2025b; Xie et al., 2025) and open-source follow-ups (Su et al., 2025; Zheng et al., 2025; Lai et al., 2025; Zhang et al., 2025c) built via reinforcement learning, showcases dynamic tool invocation as a path toward human-like visual reasoning. Yet these works almost exclusively couple CoT with image-manipulation tools and never reach for external, open-world knowledge, leaving open-source models brittle on queries that demand both fine-grained perception *and* knowledge retrieval.

We study this setting through *geolocalization*, a task that natively couples both capabilities: pinpointing the origin of an image requires extracting subtle visual clues from high-resolution inputs (road signs, architectural styles, vegetation) and grounding them against open-world knowledge retrieved via web search (Mühlbacher et al., 2024; Zhou et al., 2024a; Pang et al., 2025; Li et al., 2025c; Zhang et al., 2025a; FutureSearch et al., 2025). Prior geolocalization benchmarks, however, rely on low-resolution imagery and underexploit the V* (Wu & Xie, 2023)-style fine-grained visual search that modern VLMs are capable of. We therefore curate **GeoBench**, a benchmark of high-resolution photos, panoramas, and satellite images with global coverage, filtered to remove non-localizable samples and easily recognizable landmarks, and equipped with level-wise metrics for nuanced evaluation.

We introduce **GeoVista**, a 7B agentic model that seamlessly integrates image zoom-in and web-search within a dynamic reasoning loop. As illustrated in Fig. 1, GeoVista actively decides when and how to invoke each tool, interleaving perceptual grounding with external information retrieval to validate and refine hypotheses. This reproduces the "thinking-with-tools" behavior of closed-source systems like o3 within an open-source 7B backbone.

GeoVista is trained in two stages. A **cold-start SFT** stage instills the tool-use format: we prompt closed-source VLMs for tool-invocation proposals with rationales, execute them to obtain observations, and serialize rationales,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

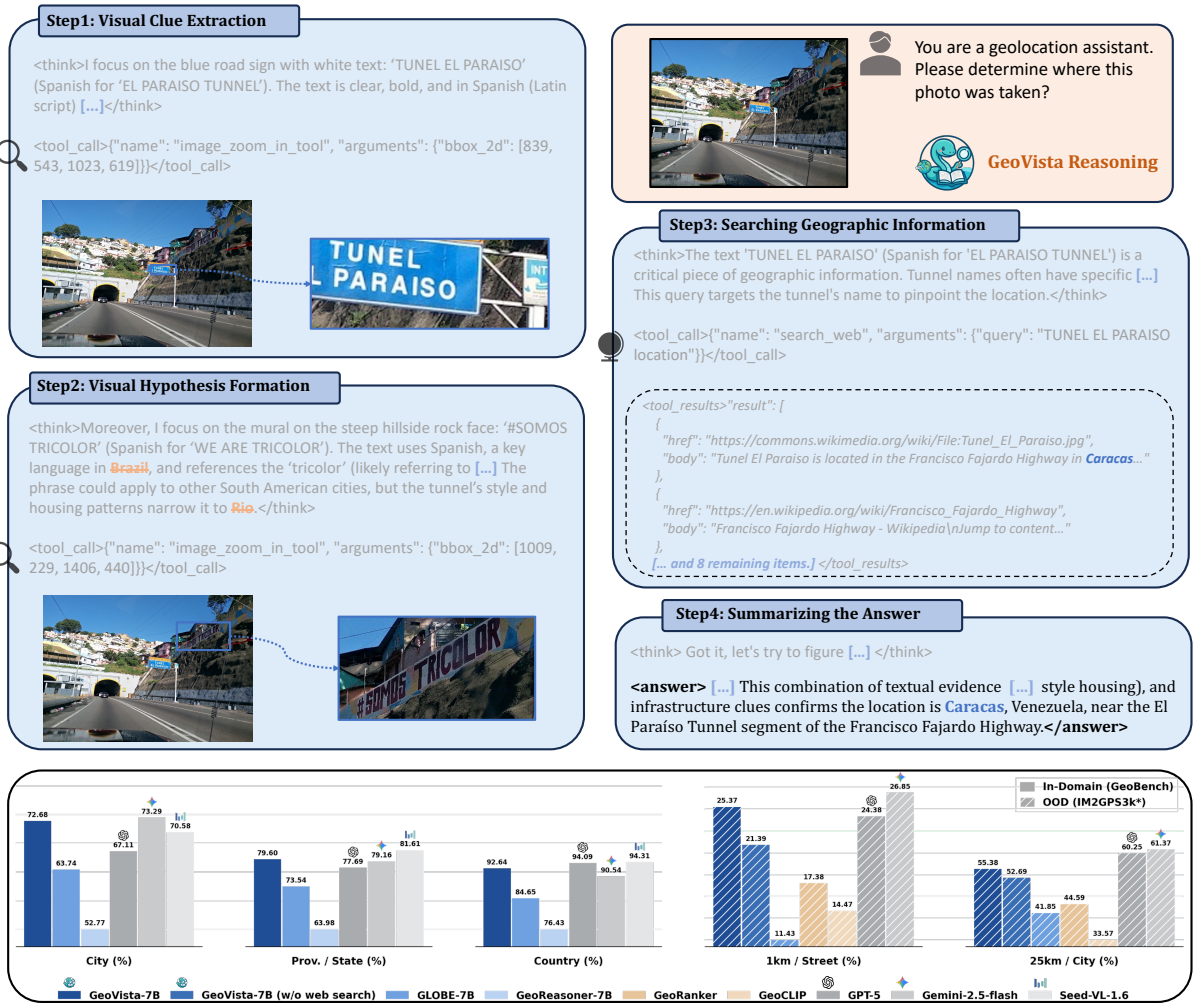


Figure 1. **Agentic reasoning and performance of GeoVista.** *(Top)* GeoVista seamlessly integrates two tool invocations, image zoom-in and web search, within a dynamic reasoning loop to validate or refine hypotheses. *(Bottom)* GeoVista achieves competitive performance against closed-source models on geolocalization benchmarks with orders of magnitude fewer parameters.

calls, and observations into multi-turn trajectories for supervised fine-tuning. We then enter a reinforcement-learning stage (DeepSeek-AI et al., 2025), where standard outcome-based RL collapses each prediction into a single binary reward, which provides an unproductively sparse signal for policies traversing long, tool-augmented trajectories. We propose **Cascaded Policy Optimization (CasPO)**, instantiated on top of GRPO (Shao et al., 2024b), which cascades credit along the natural country→province→city hierarchy and turns a single ground-truth label into a dense, multi-scale learning signal with no additional annotation. Crucially, the effect of CasPO goes beyond accuracy: it *reshapes the model's reasoning behavior*, reducing tool-call errors faster during training, producing more targeted web searches, and inducing emergent self-reflection.

Our contributions are summarized as follows:

- We propose **GeoVista**, a 7B open-source agentic VLM that unifies image zoom-in and web-search tool calls within a single dynamic reasoning loop. GeoVista actively decides when and how to invoke each tool, interleaving perceptual grounding with external knowledge retrieval to validate and refine hypotheses. This brings the o3-style "thinking-with-tools" behavior of proprietary systems into an open-source 7B backbone.
- We propose **CasPO**, a hierarchical RL method that turns sparse outcome rewards into dense, multi-scale supervision along the country, province, and city hierarchy. Paired with our full training pipeline (trajectory curation + cold-start SFT + RL), CasPO not only improves accuracy but also elicits emergent behavioral changes, including disciplined tool use, sharper search, and self-reflection, which we analyze in depth later in discussions.
- We curate **GeoBench**, a high-resolution agentic visual-

reasoning testbed, and conduct extensive experiments showing that GeoVista achieves state-of-the-art among open-source models and matches closed-source systems such as Gemini-2.5-flash and GPT-5 on most metrics with orders of magnitude fewer parameters.

2. Related Work

2.1. Thinking with Images

Research on thinking with images evolved from treating images as inputs to using visual intermediates for reasoning. Visual CoT (Shao et al., 2024a) introduced localized intermediate steps (e.g., boxes/regions) to guide attention; Visual Sketchpad (Hu et al., 2024) provided an editable canvas to draw/crop/annotate during inference; and Visual Planning argued for chains composed purely of images, replacing text with sequences of visual states. OpenAI o3 (OpenAI, 2025b) marked a watershed by productizing tool-mediated visual reasoning inside the chain (zoom, crop, rotate), triggering open replications.

After the emergence of OpenAI o3 (OpenAI, 2025b), Thyme (Zhang et al., 2025c) extends this paradigm with a code-executing visual sandbox that emits and runs image operators; mini-o3 (Lai et al., 2025) trains an agent to alternate "think–act" cycles with iterative region selection and overturn masking, scaling to deep multi-turn search; OpenThinkIMG (Su et al., 2025) unifies detectors, OCR, and drawing tools under a standardized controller with RL-learned tool policies; and DeepEyes (Zheng et al., 2025) shows purely RL-induced zoom behaviors without SFT. Collectively, these systems push beyond perception toward interactive, auditable, tool-centric visual reasoning.

2.2. Real-World Geolocalization

Prior work on real-world geolocalization spans single-image, landmark, and cross-view settings. Early global photo localization built on Im2GPS (Hays & Efros, 2008) and curated YFCC100M subsets (Vo et al., 2017), emphasizing retrieval and metric learning. Landmark-centric recognition leveraged Google Landmarks v2 (Weyand et al., 2020), improving precision where distinctive structures exist. Cross-view methods advanced with VIGOR (Zhu et al., 2021), stressing generalization across cities for ground-to-aerial matching. Scaling to worldwide street scenes, OpenStreetView-5M (Astruc et al., 2024) enabled training and fair evaluation at unprecedented diversity and size. Complementing purely visual supervision, GeoComp (Song et al., 2025) introduced human gameplay traces and reasoning sequences, catalyzing explainable localization beyond appearance cues.

GeoCLIP (Cepeda et al., 2023) and GeoRanker (Jia et al., 2025) embed images and GPS coordinates into a shared space and introduce distance-aware ranking for worldwide image-to-GPS retrieval. Building on large multimodal models, G3 (Jia et al., 2024) and Img2Loc (Zhou et al., 2024b) couple foundation models with retrieval-augmented generation for adaptive worldwide geolocalization, while (Li et al., 2025a; Wang et al., 2025a) fine-tune VLMs on street-view reasoning traces for step-by-step localization. Beyond these supervision-centric efforts, a growing line of work investigates agentic tool use for geo-localization: EmbodiedWebAgents (Hong et al., 2025) and DoxingBench (Luo et al., 2025) approach it from physical-digital embodiment and privacy-oriented stress testing, and Li et al. (2025b) further apply RL with geo-localization rewards to boost image-level prediction.

3. Approach

3.1. Agentic Pipeline

Given a user query and an input image for geolocalization, the policy model iteratively produces a thought T_i and an action A_i (Fig. 2). The action is parsed and executed to interact with the environment, which yields a new observation O_i. This observation is then appended to the interaction history and fed back into the policy model. The thought–action–observation loop terminates when the model decides to present its final answer or reaches the limit of interaction turns. The tools available to the model are of two types:

- **Crop-and-Zoom.** The policy model outputs a bounding box parameterized with `bbox_2d`, which contains pixel coordinates used to crop and magnify regions of interest. The observation is the magnified cropped subfigure.
- **Web-Search.** The policy model initializes a web search `query` to retrieve up to at most ten relevant information webpages from the internet. The web search service is provided by a third-party provider, and the observation consists of textual documents with web URLs.

3.2. GeoBench Benchmark

To ensure distributional diversity, we curate **GeoBench** and training data of GeoVista from the cities worldwide. For automated labeling, each sample is accompanied by geolocalization metadata, including precise latitude and longitude. We state how we collect the raw data in Sup.C.

We compare GeoBench with existing benchmarks and describe localizability filtering in Sup. H and Sup. I, respectively. GeoBench spans 6 continents, 66 countries, and 108 cities worldwide (see Fig. 7b in Sup. B for details).

Figure 2. **Examplars from GeoBench and GeoVista's agentic pipeline.** Given a query and image, the model iteratively generates thoughts and actions; each action is executed and yields a new observation. The loop repeats until the model outputs a final prediction or reaches the turn limit.

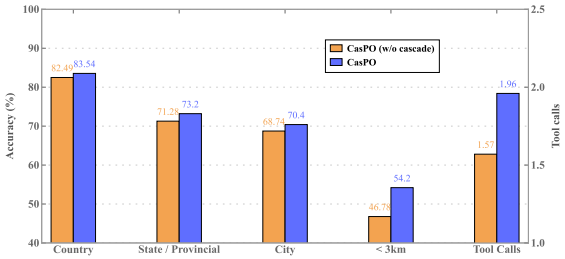


Figure 3. **Effect of the CasPO cascade.** GeoVista-7B with CasPO makes more tool calls and achieves higher accuracy.

Evaluation Pipeline. As illustrated in Fig. 7a (Sup. B), we develop multi–level labels at country, province/state, and city levels. We combine a rule–based verifier $\mathcal{V}_{\text{rule}}$ with a model–based verifier $\mathcal{V}_{\text{model}}$ (GPT-4o-mini) to validate responses at each level. Given a model answer a_i, the per-level correctness indicator is:

$$\mathbb{1}[\ell] = \mathbb{1}\left[\mathcal{V}^{\ell}_{\text{rule}}(a_i)\right] \vee \mathbb{1}\left[\mathcal{V}^{\ell}_{\text{model}}(a_i)\right], \quad \ell \in \{\text{country, province, city}\}, \tag{1}$$

where $\vee$ is logical OR. These indicators are later reused by CasPO in Eq. 3. For fine-grained evaluation (or nuanced evaluation) beyond city-level, we additionally compute haversine distance between geocoded predictions and ground-truth coordinates (see Sup. A).

3.3. Cold Start and Reinforcement Learning with CasPO

Cold-Start SFT. Training solely with RL from *Qwen-2.5-VL-Instruct* produced overly concise responses that rarely invoked tools (Lai et al., 2025), motivating a supervised warm-up. Inspired by how humans identify a place—first inspecting candidate regions, then consulting external knowledge—we prompt a VLM (*Seed-1.6-vision* (Seed, 2025)) to propose bounding boxes and web-search queries with rationales, and assemble them into coherent multi-turn trajectories. We curate 2,000 cold-start examples without answer-based filtering, aiming only to instill a reasoning prior (see Sup. D).

RL with CasPO. Our RL backbone is GRPO (Shao et al., 2024b; Schulman et al., 2017): each question q is passed to the policy model, which generates a group of outputs $\{o_i\}_{i=1}^{G}$. Rewards r_i are then computed based on rubrics (e.g., whether the model predicts the right place). In our implementation, we do not include KL or entropy regularization. Formally, the optimization objective is:

$$\mathcal{J}_{\text{RL}}(\theta) = \mathbb{E}_{q,\{o_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_i(\theta)\,A_i, \text{clip}(\rho_i(\theta),\,1-\epsilon,\,1+\epsilon)\,A_i\Big)\Bigg], \tag{2}$$

where $\rho_i(\theta) = \pi_\theta(o_i \mid q)/\pi_{\theta_{\text{old}}}(o_i \mid q)$ is the importance ratio, and $A_i = (r_i - \text{mean}(\mathbf{r}))/\text{std}(\mathbf{r})$ is the group-normalized advantage with $\mathbf{r} = \{r_j\}_{j=1}^{G}$.

However, because the data have multi-level labels, a reward that only grants credit when the model predicts the correct city does not fully utilize the hierarchical information. Under this simple reward, the model underperforms on GeoBench and makes fewer tool calls (Fig. 3). To address this, we propose **Cascaded Policy Optimization (CasPO)**, which retains the GRPO objective but replaces its binary outcome reward with a *cascaded reward* that leverages the multi-level label structure:

Table 1. **Results on GeoBench. Bold** = best within each category (closed-/open-source) and underline = open-source result surpassing at least one closed-source counterpart model.

Models	Geographic Level (%) ↑		City-Level Accuracy (%) ↑				Nuanced	
	Country	Prov./State	Overall	Panorama	Photo	Satellite	<3km (%) ↑	Med. Dist. (km) ↓
Close Source Model								
Gemini-2.5-pro	**97.20**	**86.78**	**78.98**	**78.32**	**77.54**	**88.14**	**64.45**	**0.80**
GPT-5	94.09	77.69	67.11	69.47	67.92	53.39	55.12	1.86
Seed-VL-1.6	94.31	81.61	70.58	69.73	73.44	61.86	54.00	2.22
Gemini-2.5-flash	90.54	79.16	73.29	71.88	73.83	77.12	58.11	1.67
Open Source Geolocalization Model								
GeoReasoner-7B	76.43	63.98	52.77	53.47	57.34	27.78	37.18	16.62
GLOBE-7B	84.65	73.54	63.74	68.32	66.93	26.95	48.54	3.89
Open Source Agentic Model								
Mini-o3-7B	20.14	11.52	11.30	6.05	16.02	13.56	9.57	13043.70
DeepEyes-7B	54.20	36.08	30.56	19.92	42.58	24.58	26.86	5174.93
Thyme-RL-7B	69.61	44.31	30.21	26.17	35.94	22.88	29.88	880.97
Geo-Vista-7B (*ours*)	**92.64**	**79.60**	**72.68**	**79.49**	**72.27**	**44.92**	**52.83**	**2.35**
Δ (vs Qwen2.5-VL)	+33.71	+36.69	+40.11	+55.27	+27.54	+28.82	+23.53	-2207.47

$$r_i = \max\big(\beta^2\,\mathbb{1}[\text{city}],\ \beta\,\mathbb{1}[\text{province}],\ \mathbb{1}[\text{country}],\ 0\big)\,, \tag{3}$$

where $\mathbb{1}[\cdot]$ indicates whether the prediction matches the ground truth at the corresponding administrative level. Since a correct city implies a correct province and country, the $\max$ keeps only the finest-grained reward, recovering the four-way case definition.

We set $\beta > 1$ so that correct answers at smaller administrative divisions receive larger rewards. For example, for a photo taken in Los Angeles, we give a higher reward to the answer *"Los Angeles"* than to *"San Francisco,"* because the former is correct at the city level, although both are correct at the state level. To prevent β from being so large that reward gaps become excessive, or so small that rewards collapse, empirically we choose a compromise value of $\beta = 2$ in later experiments. As reinforcement learning incurs substantial cost, particularly due to search API usage and the computational overhead of response-group rollouts, we do not experiment with additional β values.

4. Experiment

4.1. Settings

We evaluate on GeoBench and IM2GPS3k (Vo et al., 2017), comparing GeoVista against closed-source models (Gemini-2.5-pro/flash (Team, 2025), GPT-5 (OpenAI, 2025a), Seed-1.6-vision (Seed, 2025)), open-source geolocalization models (GeoReasoner-7B (Li et al., 2025a), GLOBE-7B (Li et al., 2025b)), and open-source 7B agentic models (Mini-o3 (Lai et al., 2025), DeepEyes (Zheng et al., 2025), Thyme-RL (Zhang et al., 2025c)); the base Qwen2.5-VL-Instruct (Qwen et al., 2025) is also included. All open-source agentic models share identical crop-and-zoom and web-search tools under a ReAct-style (Yao et al., 2023) loop, with inputs downsampled to $\leq 2\,\mathrm{M}$ pixels following Mini-o3. On GeoBench we report country/province/city accuracy (city broken down by data type), the <3 km rate, and median haversine distance; on IM2GPS3k we report accuracy at 1/25/200/750 km thresholds, evaluating both the full set and a high-information subset filtered by a Gemini-2.5-pro scene-informativeness judge. Full baseline, tool-use, evaluation, and inference details, along with the training recipe and ablation, are deferred to Appendix E, F, and J.

4.2. Main Results

Our experimental results on GeoBench are shown in Tab. 1. We report geographic-level accuracy (country, provincial/state), city-level accuracy broken down by data type (panorama, photo, satellite) with a weighted average, and two nuanced distance metrics (<3 km rate and median haversine distance). Across these metrics, GeoVista achieves state-of-the-art performance among open-source models, while Gemini-2.5-pro leads among closed-source counterparts.

Despite having far fewer parameters, GeoVista performs on par with closed-source models on most metrics. We attribute this to GeoVista's learned reasoning prior and its ability to use tool calls, especially the web-search tool. For the nuanced metrics, GeoVista achieves a median distance of 2.35 km and a <3 km rate of 52.83%, leaving a small gap to closed-source models but substantially outperform-

Table 2. **Results on IM2GPS3k.** We report accuracy (%) at four distance thresholds. IM2GPS3k denotes the full test set, and IM2GPS3k* denotes the high-information subset filtered by scene informativeness. **Bold** indicates best open-source result while underline indicates second best.

Models	IM2GPS3k				IM2GPS3k* (*high-info*)			
	Street 1 km	City 25 km	Region 200 km	Country 750 km	Street 1 km	City 25 km	Region 200 km	Country 750 km
Close Source Model								
Gemini-2.5-pro	20.56	54.32	67.94	78.76	29.45	63.89	77.37	90.73
GPT-5	15.23	50.86	63.21	76.59	24.38	60.25	73.58	88.46
Gemini-2.5-flash	17.49	51.95	61.19	74.32	26.85	61.37	69.38	83.62
Open Source Model								
GeoCLIP	14.11	34.47	50.65	69.67	14.47	33.57	49.38	69.34
GeoRanker	18.79	45.05	61.49	**76.31**	17.38	44.59	60.98	75.74
GeoReasoner-7B	7.67	26.94	36.63	52.27	9.76	28.47	38.68	54.96
GLOBE-7B	9.84	40.18	56.19	71.45	11.43	41.85	58.58	72.39
Geo-Vista-7B (*w/o web search*)	**19.76**	**50.84**	**62.39**	76.15	21.39	52.69	63.93	78.53
Geo-Vista-7B (*ours*)	16.48	47.43	58.34	73.57	**25.37**	**55.38**	**67.49**	**84.75**

Table 3. **Out-of-domain generalization on general visual search and perception benchmarks.** We evaluate three training stages: the base model (Qwen2.5-VL), after cold-start SFT, and after SFT + RL (GeoVista). Δ measures the gain of RL. **Bold** indicates the best result per column.

Model	V*Bench			HRBench-8K	MMBench
	Attribute	Spatial	Overall		
Qwen2.5-VL	78.24	73.63	76.42	65.34	83.53
w/ SFT	76.17 -2.07	72.41 -1.22	74.96 -1.46	63.07 -2.27	81.28 -2.25
w/ SFT + RL	**79.48** +1.24	**74.58** +0.95	**77.53** +1.11	**65.61** +0.27	**85.97** +2.44
Δ (w/ RL)	+3.31	+2.17	+2.57	+2.54	+4.69

ing all other open-source models, highlighting its superior reasoning performance.

4.3. Analysis

4.3.1. RQ1: Evaluation on Existing Geolocalization Benchmarks

We further evaluate on IM2GPS3k (Vo et al., 2017), the most widely used geolocalization benchmark. Since most of its images depict low-informativeness scenes (e.g., deserts, rural highways, featureless suburbs), we additionally use Gemini-2.5-pro as a judge to score each image's *scene informativeness* and select the top 15% as a high-information subset.

Tab. 2 reveals a intriguing pattern: On the full set, GeoVista *without* web search outperforms the full pipeline: low-informativeness scenes often yield vague queries (e.g., "desert with sparse grass") whose results carry few information gain or even burden subsequent reasoning. On the high-information subset, the picture reverses—scenes with text, distinctive architecture, or landmarks let all VLM-based methods (e.g., Gemini-2.5-pro, GeoReasoner and GLOBE-7B) improve substantially, and GeoVista with web search achieves the strongest open-source results across all four thresholds (e.g., 84.75% at country level). This confirms that agentic search is most valuable when the visual scene provides sufficient anchoring for effective queries.

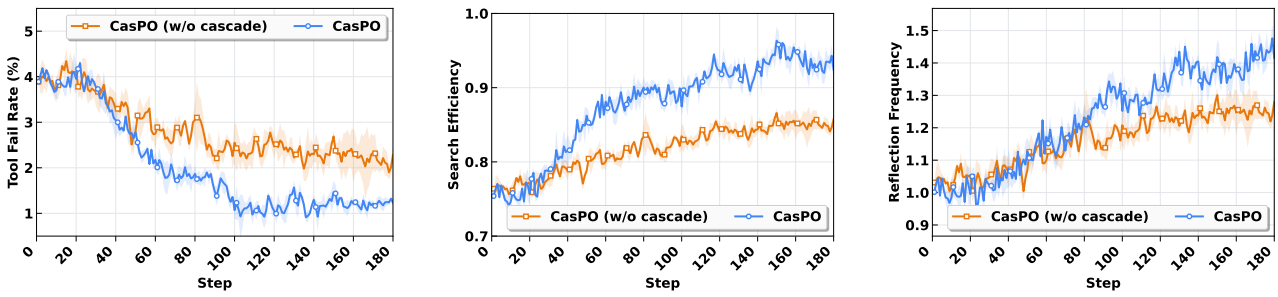
4.3.2. RQ2: Out-of-Domain Generalization

A key concern with domain-specific RL is whether it narrows the model's general capabilities. To investigate this, we evaluate GeoVista at each training stage—base model (Qwen2.5-VL), after cold-start SFT, and after SFT + RL—on three out-of-domain benchmarks that test visual search and perception abilities: V*Bench (Wu & Xie, 2023), HRBench (Wang et al., 2024b) for high-resolution understanding, and MMBench (Liu et al., 2024) for generic multimodal evaluation.

As shown in Tab. 3, cold-start SFT causes a consistent drop across all benchmarks (e.g., -2.25 on MMBench), as expected when adapting to a narrow domain format. However, the subsequent RL stage not only recovers but *surpasses* the base model on every benchmark, with gains of $+3.31$ on V*Bench Attribute, $+2.54$ on HRBench-8K, and $+4.69$ on MMBench. The Δ row highlights that RL consistently reverses the SFT degradation and yields net positive transfer. This suggests that the reasoning and tool-use skills acquired during geo-specific RL—systematic visual inspection, evidence-based search, and self-reflection—generalize to broader visual understanding tasks.

4.3.3. RQ3: Behavior Analysis of RL Training

To understand *how* RL shapes reasoning behavior beyond final accuracy, we track behavioral metrics throughout training for both CasPO and CasPO w/o cascade. Notably, the



(a) Tool Error Rate *(b)* Search Efficiency *(c)* Reflection Frequency

Figure 4. **Emergent behavioral changes during RL training.** Both CasPO and CasPO w/o cascade show the same qualitative trends: RL reduces tool errors (a), improves search targeting (b), and induces self-reflection (c), despite optimizing only for location accuracy.

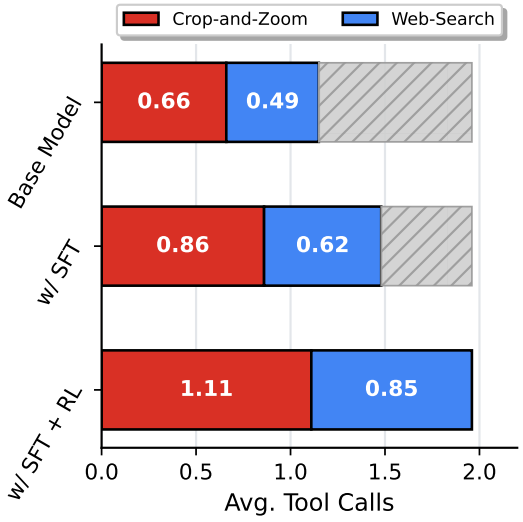


Figure 5. **Avg. tool calls per sample on GeoBench** across training stages. RL substantially increases both crop-and-zoom and web-search usage.

RL reward is purely accuracy-based, with no explicit supervision on tool usage or reasoning structure. Yet both variants exhibit the same qualitative trends (Fig. 4), indicating that RL itself drives these emergent shifts; the cascade amplifies the effect but does not create it. Meanwhile, as shown in Fig. 5, the average number of tool calls per sample increases substantially from the base model through SFT to RL, confirming that RL encourages more active tool engagement throughout multi-turn reasoning trajectories.

We highlight three emergent behavioral trends, none of which is explicitly rewarded: (1) **Tool error rate** drops steadily (Fig. 4a)—the model learns that malformed JSON calls waste reasoning budget and yield no useful information, so accuracy pressure alone is sufficient to suppress them. (2) **Search efficiency** improves over training (Fig. 4b): the model's queries shift from generic (e.g., *"tropical city with palm trees"*) to evidence-grounded (e.g., *"Carrefour supermarket Abidjan"*), showing that RL teaches the model *when* a scene contains searchable cues and *what* to search for. (3) **Reflection frequency** rises steadily for both variants (Fig. 4c): self-correction markers (e.g., *"wait"*, *"let me re-examine"*) become increasingly frequent, as re-examining hypotheses against newly gathered evidence yields more accurate final predictions.

5. Conclusion

Teaching small VLMs to autonomously reason with tools remains an open challenge. Real-world geolocalization is an ideal testbed, naturally demanding both visual clue extraction and external knowledge retrieval. We present **GeoVista**, a 7B agentic model that interleaves crop–zoom-in and web-search within multi-step reasoning, approaching closed-source models such as GPT-5 and Gemini-2.5-flash. At its core, **CasPO** cascades credit along the geographic hierarchy atop cold-start SFT, turning a single ground truth into dense multi-scale supervision. Beyond accuracy gains, our analysis reveals that RL reshapes model behavior: reducing tool errors, inducing more targeted search, and eliciting emergent self-reflection absent in SFT-only training. We further introduce **GeoBench**, a benchmark of 1,142 high-resolution images with global coverage and multi-level metrics. These results suggest that compact models can acquire sophisticated agentic reasoning with the right training recipe, and that geolocalization is a fertile testbed for studying such emergent behaviors. We hope GeoVista and GeoBench spur broader progress on tool-augmented visual reasoning and agentic multimodal models at practical scale.

Impact Statement

This work advances Machine Learning with broad societal implications that need no specific highlighting. Although geolocalization may raise privacy concerns for personal imagery, our study focuses on public-scene imagery.

LLM/Agent Usage Statement

Large language models and coding agents were used *only* for minor text polishing (grammar, phrasing, typo fixes). All research ideas, experimental design, implementation, analysis, and conclusions are the authors' own. No LLM was used to generate scientific content, results, or references.

References

Astruc, G., Dufour, N., Siglidis, I., Aronssohn, C., Bouia, N., Fu, S., Loiseau, R., Nguyen, V. N., Raude, C., Vincent, E., Xu, L., Zhou, H., and Landrieu, L. Openstreetview-5m: The many roads to global visual geolocation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21967–21977, 2024. URL https://api.semanticscholar.org/CorpusID:269448726.

Cepeda, V. V., Nayak, G. K., and Shah, M. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization, 2023. URL https://arxiv.org/abs/2309.16020.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Zhang, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., and Wang, W. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley, 2 edition, 2006. ISBN 978-0-471-24195-9.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

FutureSearch, :, Bosse, N. I., Evans, J., Gambee, R. G., Hnyk, D., Mühlbacher, P., Phillips, L., Schwarz, D., and Wildman, J. Deep research bench: Evaluating ai web research agents, 2025. URL https://arxiv.org/abs/2506.06287.

Gao, J., Li, Y., Cao, Z., and Li, W. Interleaved-modal chain-of-thought. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 19520–19529. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01818. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gao_Interleaved-Modal_Chain-of-Thought_CVPR_2025_paper.html.

Hays, J. and Efros, A. A. IM2GPS: estimating geographic information from a single image. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008. doi: 10.1109/CVPR.2008.4587784. URL https://doi.org/10.1109/CVPR.2008.4587784.

Hong, Y., Sun, R., Li, B., Yao, X., Wu, M., Chien, A., Yin, D., Wu, Y. N., Wang, Z. J., and Chang, K.-W. Embodied web agents: Bridging physical-digital realms for integrated agent intelligence, 2025. URL https://arxiv.org/abs/2506.15677.

Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N. A., and Krishna, R. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/fb82011040977c7712409fbdb5456647-Abstract-Confere html.

Jia, P., Liu, Y., Li, X., Wang, Y., Du, Y., Han, X., Wei, X., Wang, S., Yin, D., and Zhao, X. G3: An effective and adaptive framework for worldwide geolocalization using large multi-modality models, 2024. URL https://arxiv.org/abs/2405.14702.

Jia, P., Park, S., Gao, S., Zhao, X., and Li, S. Georanker: Distance-aware ranking for worldwide image geolocalization, 2025. URL https://arxiv.org/abs/2505.13731.

Lai, X., Li, J., Li, W., Liu, T., Li, T., and Zhao, H. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *CoRR*, abs/2509.07969, 2025. doi: 10.48550/ARXIV.2509.07969. URL https://doi.org/10.48550/arXiv.2509.07969.

Li, L., Ye, Y., Zhou, Y., Jiang, B., and Zeng, W. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model, 2025a. URL https://arxiv.org/abs/2406.18572.

Li, L., Zhou, Y., Liang, Y., Tsung, F., and Wei, J. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models, 2025b. URL https://arxiv.org/abs/2506.14674.

Li, X., Jin, J., Dong, G., Qian, H., Wu, Y., Wen, J.-R., Zhu, Y., and Dou, Z. Webthinker: Empowering large reasoning models with deep research capability, 2025c. URL https://arxiv.org/abs/2504.21776.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., and Lin, D. Mmbench: Is your multi-modal model an all-around player?, 2024. URL https://arxiv.org/abs/2307.06281.

Luo, W., Lu, T., Zhang, Q., Liu, X., Hu, B., Zhao, Y., Zhao, J., Gao, S., McDaniel, P., Xiang, Z., and Xiao, C. Doxing via the lens: Revealing location-related privacy leakage on multi-modal large reasoning models, 2025. URL https://arxiv.org/abs/2504.19373.

Mühlbacher, P., Bosse, N. I., and Phillips, L. Towards a realistic long-term benchmark for open-web research agents, 2024. URL https://arxiv.org/abs/2409.14913.

OpenAI. Introducing gpt-5. OpenAI Blog, August 2025a. URL https://openai.com/index/introducing-gpt-5/. Accessed: 2025-11-13.

OpenAI. Introducing openai o3 and o4-mini, April 2025b. URL https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-11-13.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong,

H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Pang, X., Tang, S., Ye, R., Du, Y., Du, Y., and Chen, S. Browsemaster: Towards scalable web browsing via tool-augmented programmatic agent pair, 2025. URL https://arxiv.org/abs/2508.09129.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Seed, B. Seed1.6 vision. Official Site, June 2025. URL https://seed.bytedance.com/en/seed1_6. Accessed: 2025-11-13.

Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., and Li, H. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning, 2024a. URL https://arxiv.org/abs/2403.16999.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024b. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Song, Z., Yang, J., Huang, Y., Tonglet, J., Zhang, Z., Cheng, T., Fang, M., Gurevych, I., and Chen, X. Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework. *CoRR*, abs/2502.13759, 2025. doi: 10.48550/ARXIV.2502.13759. URL https://doi.org/10.48550/arXiv.2502.13759.

Su, Z., Li, L., Song, M., Hao, Y., Yang, Z., Zhang, J., Chen, G., Gu, J., Li, J., Qu, X., and Cheng, Y. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *CoRR*, abs/2505.08617, 2025. doi: 10.48550/ARXIV.2505.08617. URL https://doi.org/10.48550/arXiv.2505.08617.

Team, G. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Vo, N. N., Jacobs, N., and Hays, J. Revisiting im2gps in the deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2640–2649, 2017. URL https://api.semanticscholar.org/CorpusID:7449120.

Wang, C., Ye, X., Pan, X., Pan, Z., Wang, H., and Song, Y. Gre suite: Geo-localization inference via fine-tuned vision-language models and enhanced reasoning chains, 2025a. URL https://arxiv.org/abs/2505.18700.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., and Tao, D. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models, 2024b. URL https://arxiv.org/abs/2408.15556.

Wang, Y., Chen, Q., Li, Z., Wang, S., Guo, S., Zhang, Z., and Wei, Z. Simple o3: Towards interleaved vision-language reasoning, 2025b. URL https://arxiv.org/abs/2508.12109.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Confere
html.

Weyand, T., de Araújo, A. F., Cao, B., and Sim, J. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2572–2581, 2020. URL https://api.semanticscholar.org/CorpusID:214802288.

Wu, P. and Xie, S. V*: Guided visual search as a core mechanism in multimodal llms, 2023. URL https://arxiv.org/abs/2312.14135.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., Xie, Z., Wu, Y., Hu, K., Wang, J., Sun, Y., Li, Y., Piao, Y., Guan, K., Liu, A., Xie, X., You, Y., Dong, K., Yu, X., Zhang, H., Zhao, L., Wang, Y., and Ruan, C. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Xie, R., Qiu, D., Gopinath, D., Lin, D., Sun, Y., Wang, C., Potdar, S., and Dhingra, B. Interleaved reasoning for large language models via reinforcement learning, 2025. URL https://arxiv.org/abs/2505.19640.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Zhang, W., Li, Y., Bei, Y., Luo, J., Wan, G., Yang, L., Xie, C., Yang, Y., Huang, W.-C., Miao, C., Zou, H. P., Luo, X., Zhao, Y., Chen, Y., Chan, C., Zhou, P., Zhang, X., Zhang, C., Shang, J., Zhang, M., Song, Y., King, I., and Yu, P. S. From web search towards agentic deep research: Incentivizing search with reasoning agents, 2025a. URL https://arxiv.org/abs/2506.18959.

Zhang, Y., Ding, Y., Zhang, S., Zhang, X., Li, H., Li, Z., Wang, P., Wu, J., Ji, L., Shen, Y., Yang, Y., and Gong, Y. Perl: Permutation-enhanced reinforcement learning for interleaved vision-language reasoning. *CoRR*, abs/2506.14907, 2025b. doi: 10.48550/ARXIV.2506.14907. URL https://doi.org/10.48550/arXiv.2506.14907.

Zhang, Y., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., Fan, H., Chen, K., Chen, J., Ding, H., Tang, K., Zhang, Z., Wang, L., Yang, F., Gao, T., and Zhou, G. Thyme: Think beyond images. *CoRR*, abs/2508.11630, 2025c. doi: 10.48550/ARXIV.2508.11630. URL https://doi.org/10.48550/arXiv.2508.11630.

Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., and Yu, X. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *CoRR*, abs/2505.14362, 2025. doi: 10.48550/ARXIV.2505.14362. URL https://doi.org/10.48550/arXiv.2505.14362.

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., and Neubig, G. Webarena: A realistic web environment for building autonomous agents, 2024a. URL https://arxiv.org/abs/2307.13854.

Zhou, Z., Zhang, J., Guan, Z., Hu, M., Lao, N., Mu, L., Li, S., and Mai, G. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, pp. 2749–2754. ACM, July 2024b. doi: 10.1145/3626772.3657673. URL http://dx.doi.org/10.1145/3626772.3657673.

Zhu, S., Yang, T., and Chen, C. Vigor: Cross-view image geo-localization beyond one-to-one retrieval, 2021. URL https://arxiv.org/abs/2011.12172.

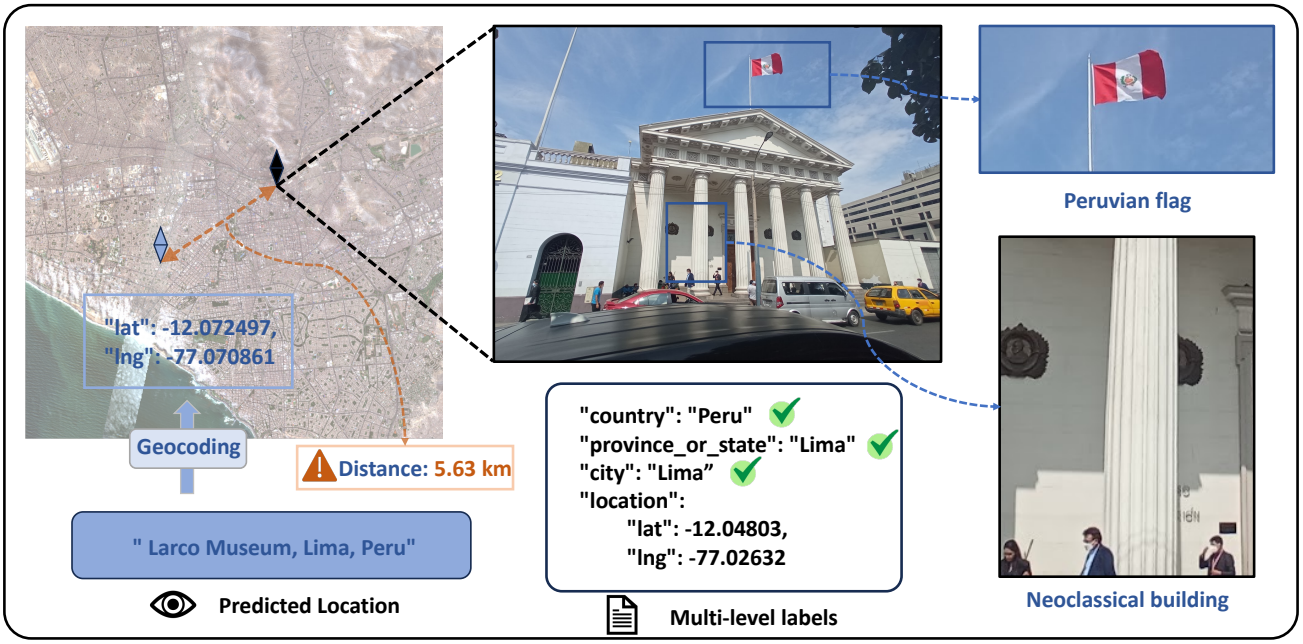


Figure 6. **GeoBench overview: sample images with level-wise and nuanced evaluation.**

A. Nuanced Evaluation and Haversine Distance

For images with richer geographic context, SOTA models can recover addresses down to street level (e.g., "Schöneberger Straße, 22149 Hamburg, Germany"). To support fine-grained evaluation beyond city-level, we extract each predicted textual location and convert it into geodetic coordinates via geocoding services (e.g., Google Geocoding API), then compute the haversine distance (km) to the ground-truth coordinates:

$$
\begin{aligned}
d &= 2R_{\mathrm{e}} \arcsin\left(\sqrt{v}\right), \\
v &= \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)
\end{aligned}
\tag{4}
$$

where (ϕ_1, λ_1) and (ϕ_2, λ_2) are the latitude/longitude pairs of the predicted and ground-truth points, and R_e is Earth's approximate radius. We report the proportion of predictions within 3 km and the median haversine distance as our nuanced metrics.

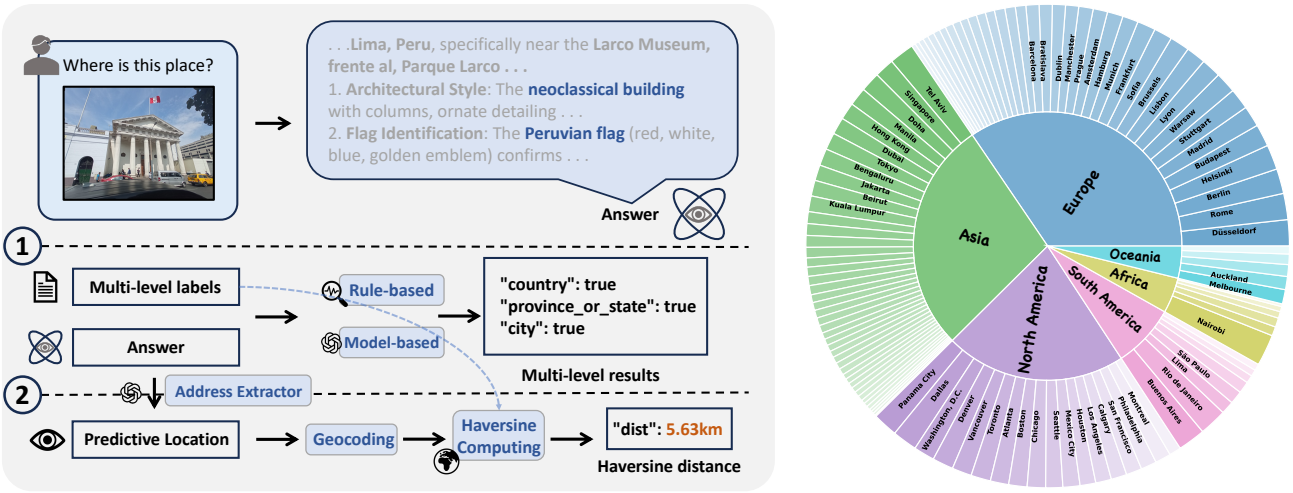
B. GeoBench Evaluation Pipeline and Geographic Distribution

Evaluation pipeline. Fig. 7a illustrates the two-pronged evaluation used throughout the paper: a level-wise verifier cascade at the country, province/state, and city levels, and a nuanced geocoding-based step that compares predicted and ground-truth coordinates via haversine distance.

Geographic distribution. We sample 512 standard photos, 512 panoramas, and 108 satellite images and conduct multi-level annotation for each. All images are high-resolution to support fine-grained visual reasoning. As shown in Fig. 7b, the dataset spans 6 continents, 66 countries, and 108 cities worldwide, ranging from Xi'an to Dublin to Washington, D.C.

C. Raw Data Collection

To improve the generalizability of our model rather than fitting it to a single data type, we query multiple types of raw data for GeoBench curation and training. The data types include:



(a) **GeoBench evaluation pipeline.** Rule-based and model-based verifiers at country/province/city levels; nuanced evaluation geocodes predictions and computes haversine distance (see Sup. A).

(b) **Geographic distribution of GeoBench**, spanning 6 continents, 66 countries, and 108 cities worldwide.

Figure 7. GeoBench evaluation system (left) and geographic coverage (right).

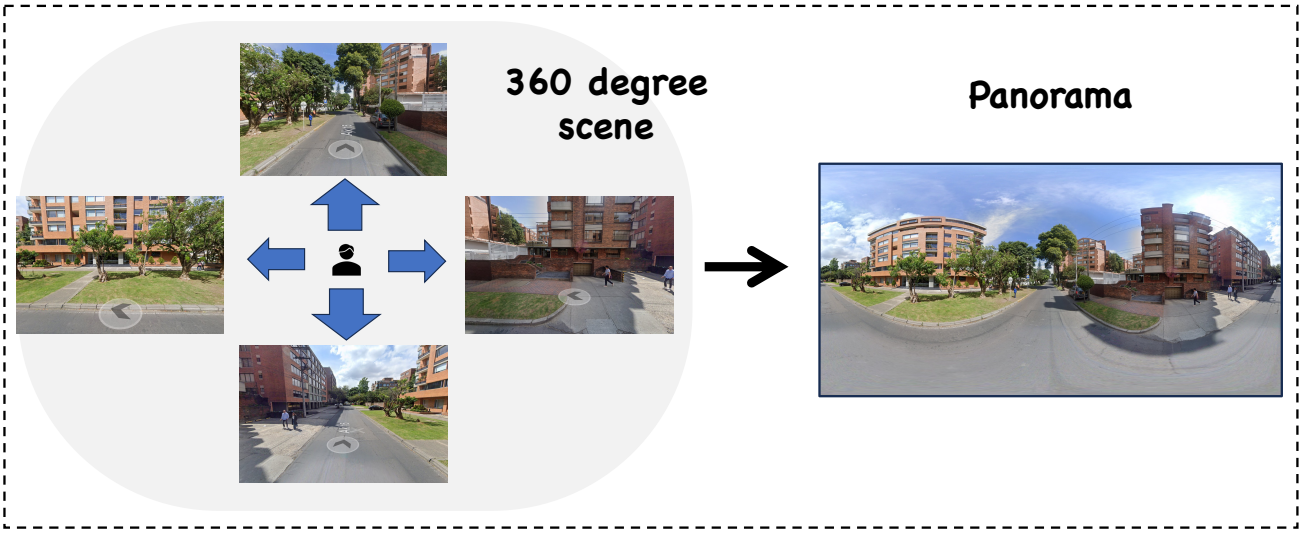


Figure 8. **The panorama stitching pipeline in GeoBench and GeoVista training data.**

- **Normal Photos.** To obtain high-quality photographs of diverse scenarios (e.g., libraries, supermarkets, suburban areas), we collect photos from the internet. These photos typically have least a resolution of 1600×1200.

- **Panoramas.** The source data are 360° street-view scenes from cities across the globe. To make them compatible with multimodal LLM input, we convert them into planar panoramas by stitching tiles retrieved via the Mapillary API and assembling them locally. To balance detail with storage, each panorama is fixed at a resolution of 4096×2048.

- **Satellite Images.** The typical size of our satellite images is 2000×2000. We retrieve recent Sentinel–2 Level–2A imagery for cities worldwide from the Microsoft Planetary Computer, mosaic several low-cloud scenes within each city's bounding box, and save multiple images together with their metadata from different viewport variants.

D. Cold-Start Data Curation

We initially attempted to train the model (i.e., *Qwen-2.5-VL-Instruct*) using reinforcement learning only, removing the need for cold-start supervised fine-tuning. However, the model tended to produce overly concise responses and hesitated to make tool calls, leading to unsatisfactory performance. This observation motivates the inclusion of explicit thinking trajectories for supervised fine-tuning, thereby incentivizing multi-turn tool-use capabilities (Lai et al., 2025).

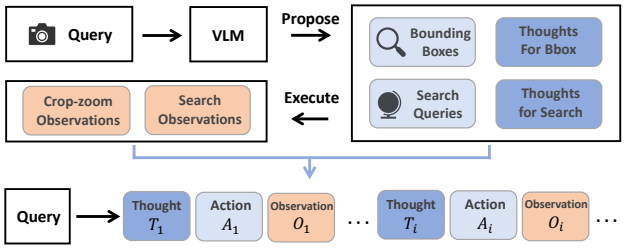


Figure 9. **Thinking trajectory curation.** A VLM proposes bounding boxes, search queries, and rationales that are assembled into multi-turn reasoning trajectories.

Inspired by how humans identify a place during geolocalization—first selecting several candidate areas to inspect and then referencing external knowledge sources (e.g., Web Search) for further information—we inject this prior into the cold-start data. As shown in Fig. 9, we use a VLM (*Seed-1.6-vision* (Seed, 2025)) to propose multiple regions (bounding boxes) along with intermediate reasoning. After perceiving salient geographic cues, the VLM is prompted to generate several web-search queries together with the accompanying rationale, then we ask it to generate the reasoning for the final judgement.

Finally, we assemble the reasoning steps, bounding boxes, and web-search queries into a coherent thinking trajectory with tool calls. As we only intend to provide the model with a reasoning pattern prior, we did not apply answer-based filtering to the reasoning trajectories. In this way, we curate 2,000 cold-start reasoning trajectory examples for geolocalization.

E. Experimental Settings Details

Models. We compare GeoVista against three categories of models. (1) *Closed-source models*: Gemini-2.5-pro, Gemini-2.5-flash (Team, 2025), GPT-5 (OpenAI, 2025a), and Seed-1.6-vision (Seed, 2025), which support iterative tool calls within their reasoning process and likely have far larger parameter counts than 7B. (2) *Open-source geolocalization models*: GeoReasoner-7B (Li et al., 2025a), which combines chain-of-thought reasoning with visual grounding for geolocalization, and GLOBE-7B (Li et al., 2025b), which leverages recognition-through-reasoning to perform geospatial prediction. (3) *Open-source agentic models*: Mini-o3-7B (Lai et al., 2025), DeepEyes-7B (Zheng et al., 2025), and Thyme-RL-7B (Zhang et al., 2025c), which share the same 7B parameter size and are capable of multi-turn tool-augmented reasoning. We also include the base model Qwen2.5-VL-Instruct (Qwen et al., 2025) for comparison.

Tool Use Access. All open-source agentic models are granted identical access to the image zoom-in tool for visual crop-zoom-in and the web-search tool for external information retrieval. We adopt a thought–action–observation, ReAct-style (Yao et al., 2023) pattern of tool calls in multi-turn interactions. For closed-source models such as GPT-5 (OpenAI, 2025a), which already integrate comparable tools into their internal reasoning, we simply issue the query in a single turn.

Evaluation. On **GeoBench**, we report level-wise accuracy at country, provincial/state, and city levels, with city-level results broken down by data type (panorama, photo, satellite). For fine-grained evaluation, we additionally report the median haversine distance and the proportion of predictions within 3 km. To assess out-of-domain generalization, we also evaluate on IM2GPS3k (Vo et al., 2017), a standard geolocalization benchmark, reporting accuracy at four distance thresholds (1, 25, 200, 750 km). Since IM2GPS3k contains many low-informativeness scenes, we additionally sample a high-information subset to better reveal each model's reasoning capacity.

Inference. Following the Mini-o3 (Lai et al., 2025) setting, to prevent the models from being overwhelmed by the context of the original high-resolution image, we set the initial pixel budget to 2 M, meaning the original image is downsampled to at most 2 M pixels before entering the visual encoder.

F. Training Recipe

Supervised Finetuning During the SFT stage, we use *Qwen2.5-VL-7B-Instruct* (Qwen et al., 2025) as the base model. To avoid out-of-memory errors from overlong trajectories, we set the maximum context length to 32768 tokens. We train on approximately 2,000 cold-start samples for 1 epoch, using a learning rate of 1×10^{-5} and a global batch size of 32.

Reinforcement Learning During reinforcement learning, we use `verl` to implement CasPO on top of GRPO (Shao et al., 2024b) with 12 k training samples. The global batch size is 64, with a mini-batch size of 32, and we use a constant learning rate of 1×10^{-6}. We remove KL regularization (Cover & Thomas, 2006), cap the maximum number of turns at 6, and set the maximum context length to 32K tokens for training efficiency. We also implement concurrent workers for tool interactions during rollout to accelerate training.

G. Scaling Effect of RL Stage

We hypothesize that model performance increases as the data size grows. Since RL data do not require reasoning-trajectory annotations, we can easily scale the RL dataset to 12 k samples. We apply different RL data sizes, including 1500, 3 k, 6 k, and 12 k, using the same cold-start SFT checkpoint. We report performance on a validation set consisting of 512 panoramas. The results show that performance consistently improves as the data size increases. When plotting data size on a logarithmic scale against performance (Fig. 10), we observe a nearly perfect data-scaling effect.

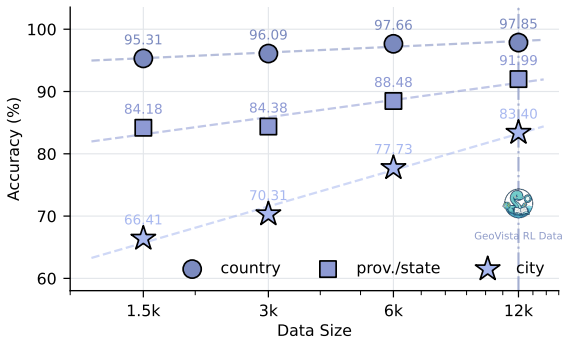


Figure 10. **RL data scaling effect.** Log-linear gains on the 512-panorama validation set as RL data size grows to 12k.

H. Benchmark Comparison

We compare GeoBench with existing geolocalization benchmarks along the following axes:

Table 4. **Comparison across geolocalization datasets.** GeoBench is the first benchmark targeting agentic geolocalization models, combining localizability filtering, high-resolution imagery, diverse data types, and hierarchical evaluation.

Benchmark	Year	GC	RC	HR	DV	NE
Im2GPS(Hays & Efros, 2008)	2008	✓				
YFCC4k (Vo et al., 2017)	2017	✓				
Google Landmarks v2(Weyand et al., 2020)	2020	✓				
VIGOR (Zhu et al., 2021)	2022				✓	
OSV-5M (Astruc et al., 2024)	2024	✓	✓			✓
GeoComp (Song et al., 2025)	2025	✓	✓			✓
GeoBench (*ours*)	2025	✓	✓	✓	✓	✓

- **Global Coverage (GC).** Whether the benchmark contains images from across the globe, ensuring that the model does not overfit or bias its performance toward specific regions.
- **Reasonable Localizability (RC).** Whether the benchmark filters out non-localizable images or easily localizable landmarks to maintain meaningful localization difficulty.
- **High Resolution (HR).** Whether all images in the benchmark have at least 1 M pixels to support reliable visual clue extraction and grounding.
- **Data Variety (DV).** Whether the benchmark includes two or more types of images to test the generalizability of reasoning models under varying data conditions.
- **Nuanced Evaluation (NE).** Whether the benchmark includes geolocation coordinates to enable haversine distance computation for nuanced evaluation.

I. Localizability Filtering

We conduct localizability filtering to remove non-localizable images and easily localizable landmarks. Images collected from the Internet exhibit varying levels of localizability (Astruc et al., 2024), especially when the data types and sources differ. Therefore, we remove two categories of data via model-based filtering:

- **Non-localizable images.** These images usually lack identifiable geographical clues and contain generic objects or scenes, such as close-up food photos, indoor rooms, plain natural landscapes, or single animals. Such content provides almost no regional or cultural context, making localization infeasible.
- **Easily localizable landmarks.** These images contain strong geographic priors, typically featuring iconic landmarks or globally recognizable sites. Since VLMs have likely encountered such images multiple times during pretraining, including them would make geolocation trivial and fail to reflect genuine reasoning ability.

Figure 11. **Localizable vs. non-localizable examples.** Non-localizable images (generic scenes, close-ups) and iconic landmarks are filtered out, retaining only images that require genuine reasoning to localize.

J. Ablation Study

Table 5. **Ablation study on GeoBench.** We ablate along two dimensions: *training* (top) removes a training stage or component; *inference tools* (bottom) disables a tool at test time. Removing cold-start SFT or RL collapses performance; disabling either tool degrades accuracy across all levels.

Variant	Median Dist. (km) ↓	City-Level Accuracy (%) ↑			
		Overall	Panorama	Photo	Satellite
Qwen-2.5-VL	2209.82	32.57	24.22	44.73	16.10
Training Ablation					
w/o Cold Start	55.32	44.30	48.52	43.63	27.46
w/o RL	11.17	53.54	54.88	57.23	29.66
w/o cascade	4.11	68.99	75.00	68.95	40.68
Inf. Tool Ablation					
w/o crop-and-zoom	2.94	71.20	77.89	70.61	42.28
w/o web-search	3.98	68.89	75.34	68.38	40.72
Geo-Vista-7B	**2.35**	**72.68**	**79.49**	**72.27**	**44.92**

We present an ablation study to quantify the contribution of each component. The overall results appear in Tab. 5. Unless otherwise stated, we keep the same training hyperparameters and evaluation settings.

Cold Start (SFT) To assess the necessity of cold-start supervised finetuning, we remove the cold-start stage and conduct reinforcement learning directly. The results show that cold-start SFT is essential for multi-turn tool use, as performance on GeoBench collapses without it.

Reinforcement Learning (RL) To examine the necessity of reinforcement learning, we remove the RL and only conduct the cold-start SFT. The results show that SFT alone is not sufficient: although the model learns a reasoning prior, it requires reinforcement learning to incentivize and strengthen its reasoning capability.

CasPO Cascade We evaluate the necessity of the cascaded reward in CasPO. We keep both the cold-start SFT and reinforcement learning, but disable the cascade during the RL stage (*CasPO w/o cascade*), using only a city-level reward. The results confirm the importance of the cascade.

We further ablate the two inference-time tools. (1) **Crop-and-Zoom** Removing the image crop-and-zoom tool causes a moderate drop (overall city accuracy 72.68 → 71.20), indicating that regional visual inspection contributes to fine-grained

reasoning. (2) **Web Search** Removing the web-search tool leads to a larger degradation (72.68 → 68.89; median distance 2.35 → 3.98 km), confirming that external knowledge retrieval is the more critical tool for geolocalization, particularly for scenes containing searchable cues such as text, signage, or distinctive buildings.

K. Case Study

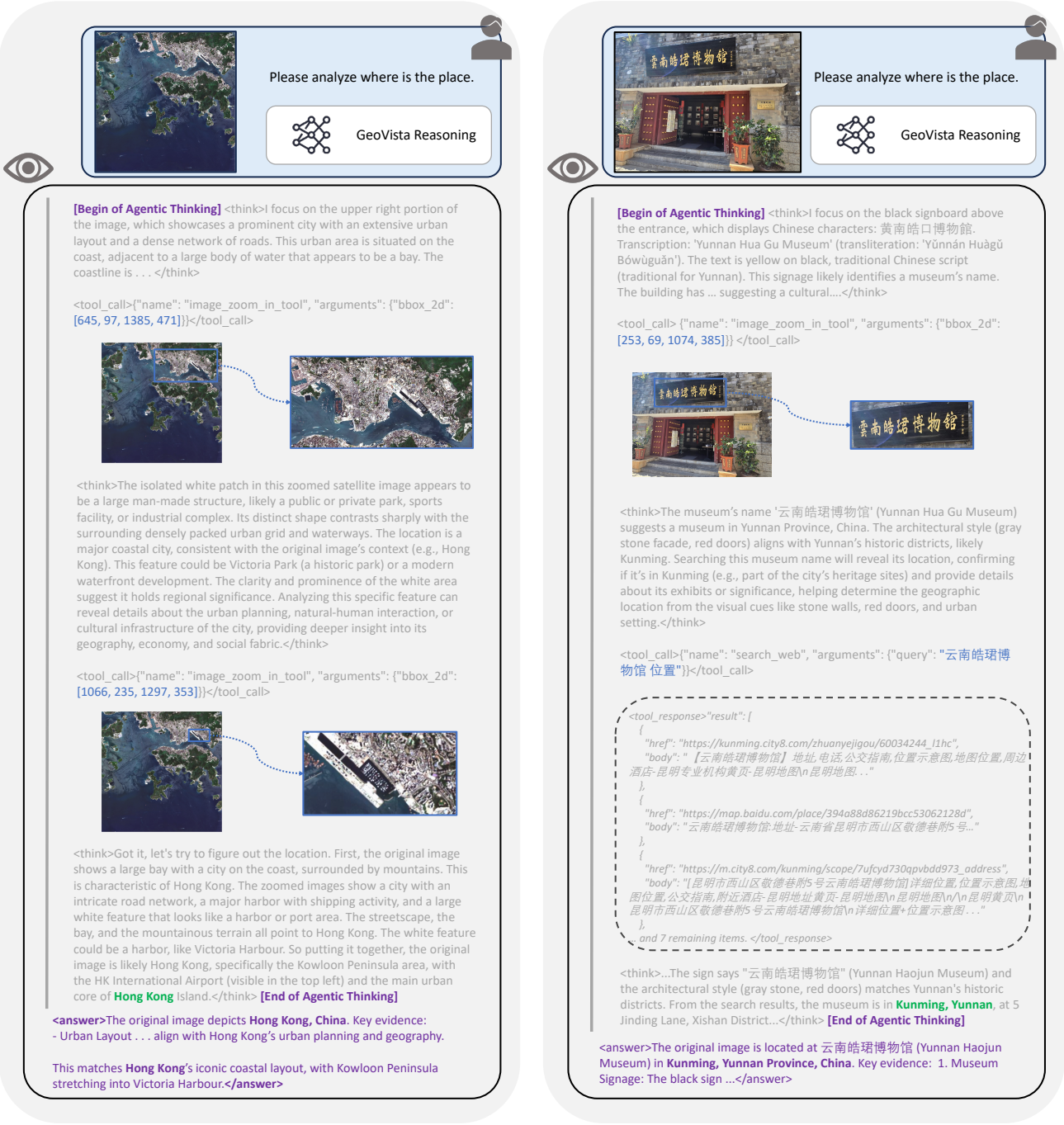


Figure 12. **The Reasoning Trajectory of GeoVista.** We provide additional cases to facilitate the analysis of GeoVista's reasoning trajectories and behavioral patterns, including one satellite-image example and one photo example from GeoBench.

L. System Prompt

SYS PROMPT

```
You are a helpful assistant.
Answer the user's question based on the image provided.
# Tools

You may call one or more functions to assist with the user query.

You are provided with function signatures within <tools></tools> XML tags:

# How to call a tool
Return a json object with function name and arguments within <tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>

# Tool definition

## Image crop and zoom in tool
<tools>
{
"name": "image_zoom_in_tool",
"description": "Zoom in on a specific region of an image by cropping it based on a bounding box (bbox).",
"parameters": {
    "properties": {
        "bbox_2d": {
            "type": "array",
            "description": "The bounding box as [x1, y1, x2, y2] on the original image."
        }
    },
    "required": ["bbox_2d"]
}
}
</tools>

## Search web tool
<tools>
{
  "type": "function",
  "function": {
    "name": "search_web",
    "description": "Execute a web search and return normalized results containing titles, snippets, and
    ↪  URLs.",
    "parameters": {
      "type": "object",
      "properties": {
        "query": {
          "type": "string",
          "description": "The search query string."
        }
      },
      "required": ["query"]
    }
  }
}
</tools>


**Example**:
<tool_call>
{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [10, 20, 100, 200]}}
</tool_call>
<tool_call>
{"name": "search_web", "arguments": {"query": "The palace museum"}}
</tool_call>

When you call a tool, think first. Place your internal reasoning inside <think>...</think>, followed by the
↪  <tool_call>...</tool_call>. And when you are ready to answer, also think first. Place your internal
↪  reasoning inside <think>...</think>, followed by the <answer>...</answer>.
```

M. Scene Informativeness Scoring Prompt

To construct the high-information subset IM2GPS3k*, we prompt Gemini-2.5-pro to score each image's scene informativeness on a 0–20 scale. The full prompt is shown below.

SCENE INFORMATIVENESS PROMPT

```
You are an image geolocation informativeness scorer. Evaluate how much useful geographic information is
↪  present in the image and assign a score from 0 to 20.

## Scoring Guidelines

- 0-3: No useful clues. Featureless landscape, generic road, or unreadable content.
- 4-7: Only coarse-grained clues (e.g., language, driving side, country-level vegetation/architecture style).
↪  Not enough to narrow below country/region level.
- 8-12: Moderate clues. Some readable text, partial signage, or recognizable regional features that could
↪  narrow to a province/state level but not a specific city.
- 13-16: Strong clues. Readable shop names, street signs, route numbers, advertisements, or district names
↪  that can plausibly identify a specific city.
- 17-20: Definitive clues. Iconic landmarks with strong geographic priors (e.g., Eiffel Tower, Oriental Pearl
↪  Tower, Tiananmen), or multiple corroborating city-level identifiers (e.g., city name on signage + local
↪  transit branding).

## What to Focus On

- Shop/store names and signage
- Road signs, street names, traffic signs
- Lane markings and road infrastructure
- Roadside advertisements and billboards
- Text on clothing, vehicles, or banners
- Any other clearly visible text or graphic symbols
- Recognizable landmark buildings or structures

## What to Ignore

- Google Maps watermarks or any map-provider watermarks -- these are NOT valid evidence.
- Do NOT fabricate or guess information not visible in the image.

## Output Format

Return a strict JSON object with no extra text:

{
  "score": <integer 0-20>,
  "has_city_level_geo_cues": <bool>,
  "has_strong_prior_landmark": <bool>,
  "landmark_names": [<string>, ...],
  "detected_texts": [<key readable texts found in the image>],
  "evidence": [<brief descriptions, e.g. "sign: XX Avenue", "storefront: XX Kebab">],
  "city_candidates": [<candidate cities inferred from evidence, may be empty>],
  "reasoning": "<20-60 word summary of scoring rationale>"
}
```