GeoVista: Web-Augmented Agentic Visual Reasoning for Geolocalization

Yikun Wang[1,2], **Zuyan Liu**[3], **Ziyi Wang**[3], **Han Hu**[4], **Pengfei Liu**[2,5]*, **Yongming Rao**[4]*
[1]Fudan University [2]Shanghai Innovation Institute
[3]Tsinghua University [4]Tencent Hunyuan [5]Shanghai Jiao Tong University

Abstract

Current research on agentic visual reasoning enables deep multimodal understanding but primarily focuses on image manipulation tools, leaving a gap toward more general-purpose agentic models. In this work, we revisit the geolocalization task, which requires not only nuanced visual grounding but also web search to confirm or refine hypotheses during reasoning. Since existing geolocalization benchmarks fail to meet the need for high-resolution imagery and the localization challenge for deep agentic reasoning, we curate **GeoBench**, a benchmark that includes photos and panoramas from around the world, along with a subset of satellite images of different cities to rigorously evaluate the geolocalization ability of agentic models. We also propose **GeoVista**, an agentic model that seamlessly integrates tool invocation within the reasoning loop, including an image-zoom-in tool to magnify regions of interest and a web-search tool to retrieve related web information. We develop a complete training pipeline for it, including a cold-start supervised fine-tuning (SFT) stage to learn reasoning patterns and tool-use priors, followed by a reinforcement learning (RL) stage to further enhance reasoning ability. We adopt a hierarchical reward to leverage multi-level geographical information and improve overall geolocalization performance. Experimental results show that GeoVista surpasses other open-source agentic models on the geolocalization task greatly and achieves performance comparable to closed-source models such as Gemini-2.5-flash and GPT-5 on most metrics.

Webpage: https://geo-vista.github.io

1 Introduction

Recent advances in Vision-Language Models (VLMs) (Wang et al., 2024; Wu et al., 2024; Chen et al., 2025) enable deep reasoning over multimodal queries by invoking image-centric tools and utilizing long Chain-of-Thought approaches (Shao et al., 2024a; Hu et al., 2024), allowing these models to handle much more complex tasks. Some recent works (Su et al., 2025; Zheng et al., 2025) attempt to integrate seamless tool invocation into multi-turn interaction through reinforcement learning.

Among the latest multimodal reasoning milestones (Gao et al., 2025; Zhang et al., 2025c; Wang et al., 2025; Xie et al., 2025), the OpenAI o3 model (OpenAI, 2025b) enables a dynamic reasoning process with different tools integrated into it. This marks the transcendence of multimodal reasoning from one-turn queries to smooth "thinking with images" like humans, achieving a coordinated fashion of interleaving textual CoT (Wei et al., 2022) with image manipulation and other tool invocations. Some follow-up works (Lai et al., 2025; Zhang et al., 2025b) also explore combining image-centric tools with open-sourced models to achieve similar performance. However, these works only emphasize image manipulation during multimodal reasoning, thus making problem-solving rely solely on the model's inherent knowledge and lacking appropriate access to external information retrieval tools like web search.

*Corresponding author

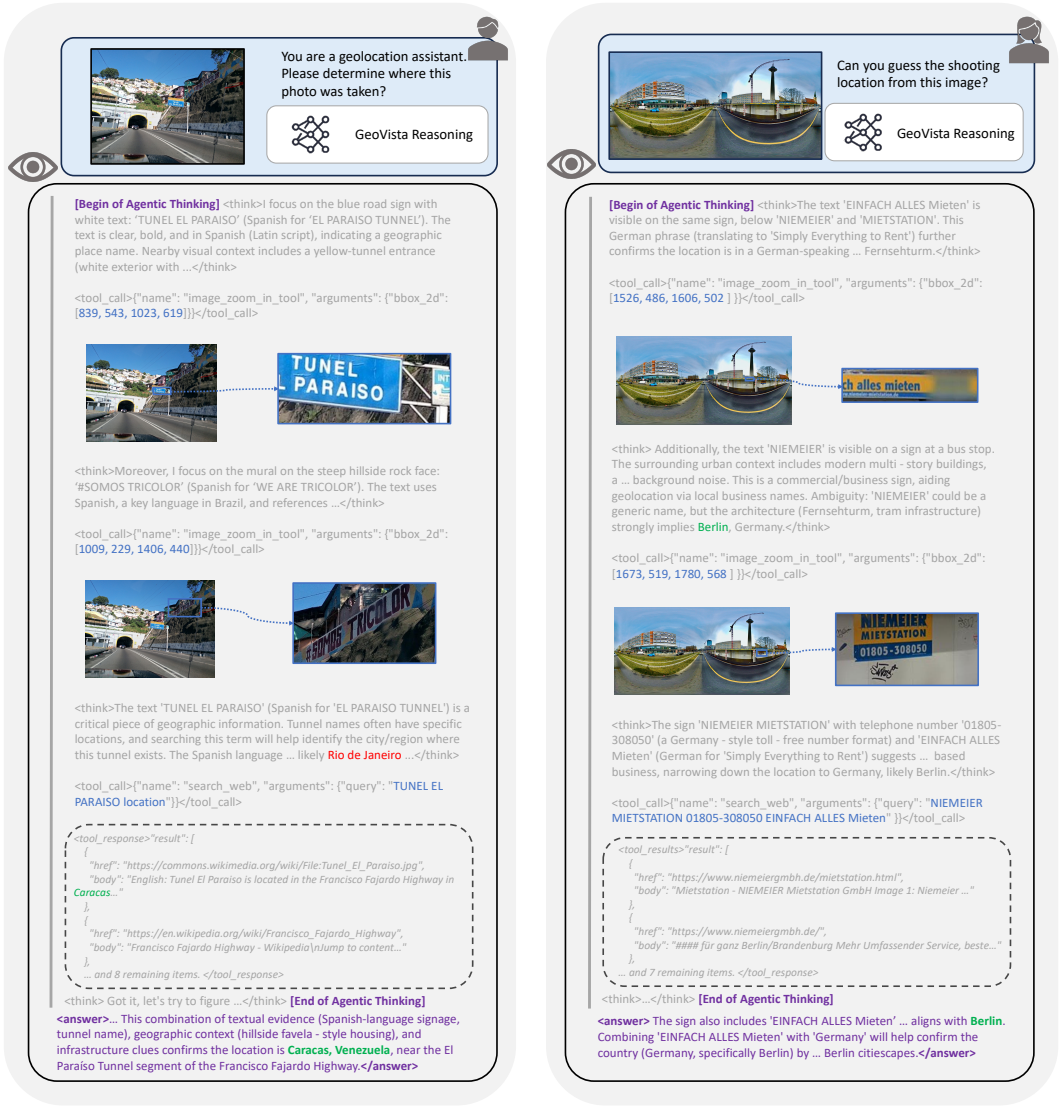

Figure 1: **Agentic thinking of GeoVista for Real-world geolocalization**. During the reasoning loop, our GeoVista seamlessly integrates the image–zoom-in tool to magnify regions of interest and the web-search tool to retrieve relevant information. This web-augmented visual reasoning process enables GeoVista validate or refine its geolocalization judgments.

To enable a new axis for agentic multimodal reasoning, we revisit a real-world scenario—geolocalization, in which models are required to extract visual clues in high-resolution images and rely on the web search to validate or refine their hypotheses (Li et al., 2025; Zhang et al., 2025a; FutureSearch et al., 2025). This makes the geolocalization scenario naturally combine visual tools and information retrieval tools. To rigorously evaluate the models, we propose **GeoBench**, which consists of high-resolution photos and panoramas of global coverage. To ensure localizability as well as challenge, we remove non-localizable ones and easily recognizable landmarks. To gain insights, GeoBench also supports level-wise evaluation and nuanced evaluation to fully assess models' geolocalization capability.

We also propose our **GeoVista**, an agentic multimodal model, which seamlessly integrates tool invocation like web-search and image-zoom-in within a dynamic reasoning loop for complex geolocalization queries. As illustrated in Fig.1, GeoVista actively decides when

and how to invoke tools, enabling a dynamic process of visual clue extraction and external information retrieval, reproducing reasoning behaviors similar to closed-source models like OpenAI o3. It not only utilizes visual operation and information retrieval tools to validate its hypotheses but also uses external information retrieval (Mühlbacher et al., 2024; Zhou et al., 2024; Pang et al., 2025) to justify its previous wrong hypotheses and reach the correct solution.

We also provide a complete pipeline for GeoVista training, including cold-start and reinforcement learning. First is the cold-start supervised finetuning (SFT) for learning tool-use and reasoning priors: We apply closed-source VLMs to generate tool invocation proposals with their rationales, execute the tool proposals to obtain the observations, and then serialize the rationales, tool invocations, and observations to generate multi-turn reasoning trajectories in order to conduct cold-start supervised finetuning. We control the number of interaction turns by limiting different tool invocation proposals.

Second is the reinforcement learning to further incentivize reasoning ability (DeepSeek-AI et al., 2025). We apply group relative policy optimization (GRPO) (Shao et al., 2024b) with geological labels to train the models. Geological information often contains hierarchical information; to fully utilize the multi-level information, we design a hierarchical reward based on multi-level labels. This simple yet effective strategy encourages the models to learn hierarchical geological contexts from the images and make more accurate judgments.

Our contributions are summarized as follows:

- We revisit the geolocalization task in the era of large reasoning models, which naturally requires visual clue extraction and external knowledge retrieval. We propose the **GeoBench** benchmark, which features high-resolution images with high localizability challenge, various data types of global coverage, and allows multi-level evaluation for insightful assessment.
- We propose **GeoVista**, which seamlessly integrates tool invocation within a dynamic reasoning loop for complex geolocalization queries. We also provide a complete training pipeline consisting of reasoning trajectory curation, cold-start SFT, and reinforcement learning. We further adopt a hierarchical reward during the RL stage for utilizing hierarchical information in geological labels.
- We also conduct extensive experiments to demonstrate the effectiveness of GeoVista on GeoBench and perform analysis experiments to gain insights into our approach.

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., 2025b) 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.

Figure 2: **Image examples from GeoBench and the training data, and the agentic pipeline of GeoVista.** Given a query and image, the policy model iteratively generates thoughts and actions; each action is parsed, executed, and yields a new observation, repeating this loop until it outputs a final geolocation prediction or reaches the maximum interaction turn limit.

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., 2020), 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, step-wise localization beyond raw appearance cues.

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 10 relevant information sources 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 geolocaliza-

tion metadata, including precise latitude and longitude. We state how we collect the raw data in Sup.A.

Comparison with Existing Geolocalization Benchmarks We compare our GeoBench with the existing benchmarks, we assess benchmarks along the following axes:

Table 1: **Comparison across geolocalization datasets**. GeoBench is the first benchmark designed to evaluate the general geolocation ability of agentic models. It features reasonable localizability, high-resolution imagery, 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., 2020)	2022				✓	
OSV-5M (Astruc et al., 2024)	2024	✓	✓			✓
GeoComp (Song et al., 2025)	2025	✓	✓			✓
GeoBench (*ours*)	2025	✓	✓	✓	✓	✓

- **Global Coverage.** 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.** Whether the benchmark filters out non-localizable images or easily localizable landmarks to maintain meaningful localization difficulty.
- **High Resolution.** Whether all images in the benchmark have at least 1 M pixels to support reliable visual clue extraction and grounding.
- **Data Variety.** Whether the benchmark includes two or more types of images to test the generalizability of reasoning models under varying data conditions.
- **Nuanced Evaluation.** Whether the benchmark includes geolocation coordinates to enable haversine distance computation for nuanced evaluation.

Localizability Filtering We also conduct localizability filtering to remove non-localizable images and easily localizable landmarks. As we believe that 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.

Figure 3: **Localizable vs Non-Localizable.** We remove the non-localizable (orange) and the landmarks (purple) from GeoBench, leaving only localizable images for rigorously evaluating models.

- **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.

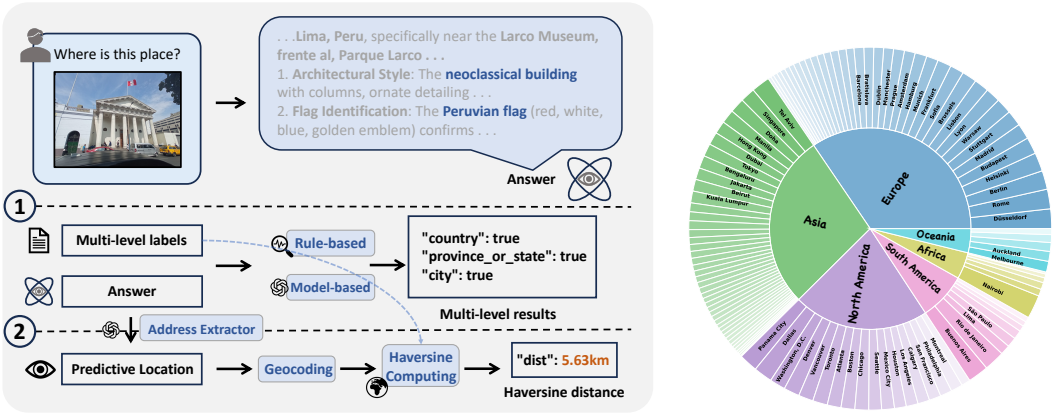

Figure 4: LEFT: **The evaluation pipeline of GeoBench dataset**. The evaluation system consists of (1) Level-wise evaluation, which employs both rule-based and model-based verifiers to determine correctness at different administrative levels, and (2) nuanced evaluation, which extracts the predicted address, applies geocoding to obtain the predicted geolocalization point, and computes the haversine distance to the ground-truth location. RIGHT: **Geological distribution of GeoBench.** GeoBench is a high-resolution, multi-source, globally annotated dataset to evaluate models' general geolocalization ability.

Level-wise Evaluation To support a fully automated, rule–based evaluation pipeline and to enable in–depth analysis of models' geolocalization capability, we develop multi–level labels that include each image's country, province or state, and city. With these multi–level geographical labels, we combine a rule–based verifier for matching specific terms with a model–based verifier (using *OpenAI gpt-4o-mini*) to validate the correctness of model responses at different administrative levels.

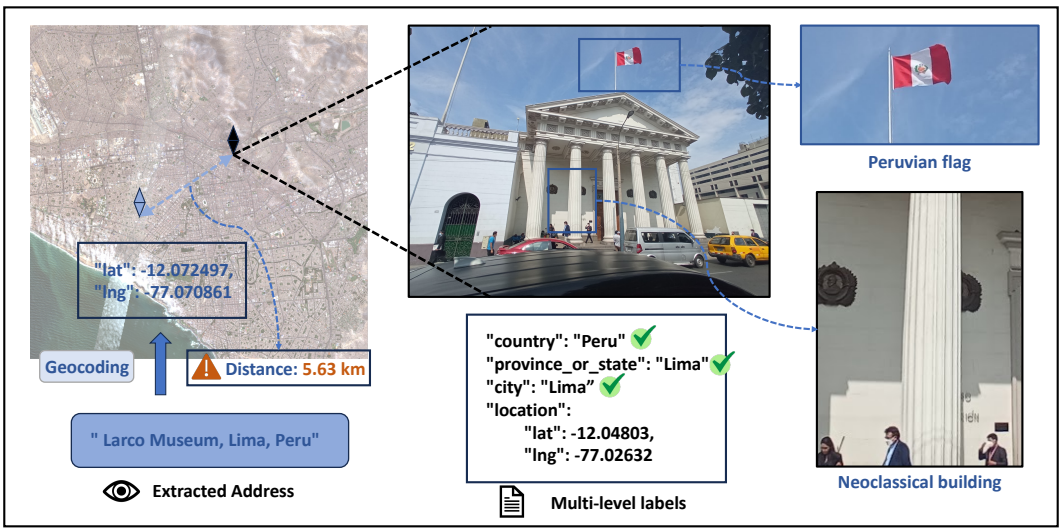

Figure 5: **Illustration of GeoBench dataset, along with level-wise and nuanced evaluation**.

Nuanced Evaluation and Haversine Distance For some images with richer geographic context, state–of–the–art (SOTA) models such as Gemini–2.5–Pro can recover much more

detailed addresses to street level, e.g., "Schöneberger Straße, 22149 Hamburg, Germany." Hence we posit that a more fine–grained evaluation beyond city–level is required. However, models often cannot predict the geolocalization point directly, which makes nuanced evaluation difficult.

To this end, as shown in Fig.4, for each response we extract the predicted textual location and convert it into geodetic coordinates (latitude and longitude) via geocoding services (e.g., *Google Geocoding API*), thereby allowing us to compute the estimated haversine distance (km) between the prediction point and the ground truth point (the geolocalization coordinates of the metadata) in an automated fasion:

$$
\begin{aligned}
d &= 2R_{\mathrm{e}} \arcsin\left(\sqrt{v}\right), \\
v &= \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)
\end{aligned}
\tag{1}
$$

where (ϕ_1, λ_1) and (ϕ_2, λ_2) are the latitude/longitude pairs of the prediction point and the ground truth point, and R_e is Earth's approximate radius.

Geological Distribution We aim to construct a dataset with diverse sources and broad geographic coverage to evaluate both closed–source and open–source models on general geolocalization ability. To this end, we sample 512 standard photos, 512 panoramas, and 108 satellite images from the raw data (see Sup.A) and conduct multi–level annotation for each image. The data are high–resolution to support fine–grained visual reasoning, and the images span 6 continents, 66 countries, and 108 cities worldwide, ranging from Xi'an to Dublin to Washington, D.C. (Fig. 4).

3.3 Cold Start and Thinking Trajectory 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.

Inspired by how humans identify a place during geolocalization—first selecting several candidate areas to inspect and then referencing external knowledge sources (e.g., Google Search) for further information—we inject this prior into the cold-start data. As shown in Fig.**??** 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.

3.4 Reinforcement Learning

We apply a vanilla GRPO (Shao et al., 2024b) setting: each question q is passed to the policy model, which generates a group of outputs $\{o_i\}_{i=1}^{G}$. Rewards r_i are computed based on response correctness (e.g., whether the model predicts the city where the photo is taken). In our implementation, we do not include KL or entropy regularization. Formally, the optimization objective is:

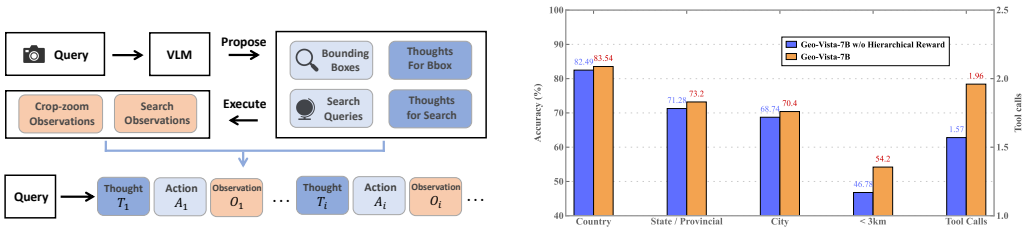

Figure 6: LEFT: **Thinking trajectory curation**. We mimic human geolocalization by using a VLM to propose tool calls and rationales, and assemble tool-call reasoning trajectories. RIGHT: **Comparison of GeoVista-7B and its counterpart w/o Hierarchical Reward**.

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \frac{1}{G}\sum_{i=1}^{G}\left[\min\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}A_i,\ \text{clip}\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\epsilon, 1+\epsilon\right)A_i\right)\right] \tag{2}$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \tag{3}$$

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.6-RIGHT). To address this, we adopt a **hierarchical reward** to fully leverage the multi-level structure:

$$r_i = \begin{cases} \beta^2, & \textit{if city-level correct}, \\ \beta, & \textit{if provincial/state-level correct}, \\ 1, & \textit{if country-level correct}, \\ 0, & \textit{else}. \end{cases} \tag{4}$$

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 Training Recipe

Supervised Finetuning During the SFT process, we use *Qwen2.5-VL-7B-Instruct* (Qwen et al., 2025) as the base model. In order to avoid out-of-memory error caused by overlong trajectories, we set a max context length of 32768. We train on approximately 2000 cold-start samples for 1 epochs. The learning rate is set to 1×10^{-5}, with the global batch size is 32.

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

5 Experiment

5.1 Settings

Models We compare **GeoVista** against a comprehensive suite of models. This suite is including closed-source systems—***Gemini–2.5–pro***, ***Gemini–2.5–flash*** (Team, 2025), ***GPT-5*** (OpenAI, 2025a), and ***Seed-1.6-vision*** (Seed, 2025)—which are supporting iterative tool calls within their reasoning process. We are also comparing open-source, vision-capable reasoning models such as ***Mini-o3-7B*** (Lai et al., 2025), ***DeepEyes-7B*** (Zheng et al., 2025), and ***Thyme-RL-7B*** (Zhang et al., 2025b), which are sharing the same 7B parameter size as our **GeoVista**. We also use the base model ***Qwen2.5-VL-Instruct*** (Qwen et al., 2025) for comparison. It is worth noting that the closed-source models, although not publicly specified, are likely having far larger parameter counts than 7B.

Tool Use Access We grant all open-source models identical access to the image–zoom-in tool for visual regional inspection and to a real-time 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 the closed-source models like GPT-5 (OpenAI, 2025a), which are already integrating comparable tools into their internal reasoning, we simply issue the query in a single turn.

Evaluation For a rigorous and insightful evaluation of geolocalization performance, we use **GeoBench** and conduct level-wise assessment at the **country level**, **provincial level**, and **city level**, reporting accuracy at each level. To analyze performance across different data types, we separately report city-level accuracy on panoramas, photos, and satellite images. To further assess each model's ability to produce fine-grained geolocalization results, we conduct the nuanced evaluation and report two metrics: the proportion of predictive locations with the **haversine distance less than 3 km** and the **median haversine distance**.

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 are setting the initial pixel budget to 2 M, meaning the original image is being downsampled to at most 2 M pixels before entering the visual encoder.

Table 2: **The Comparison on GeoBench.** The **bold** figures indicate the best performance among closed-source and open-source models, and the underlined figures indicate open-source results that surpass at least one of their closed-source counterparts.

Models	Country (%) ↑	Provincial / State (%) ↑	City (%) ↑	City (%) (Panorama) ↑	City (%) (Photo) ↑	City (%) (Satellite) ↑
Close-sourced Models						
Gemini-2.5-pro	**97.20**	**86.78**	**78.98**	**78.32**	**77.54**	**88.14**
GPT-5	94.09	77.69	67.11	69.47	67.92	53.39
Seed-VL-1.6	94.31	81.61	70.58	69.73	73.44	61.86
Gemini-2.5-flash	90.54	79.16	73.29	71.88	73.83	77.12
Open-sourced Models						
Qwen2.5-VL-7B	58.93	42.91	32.57	24.22	44.73	16.10
Mini-o3-7B	20.14	11.52	11.30	6.05	16.02	13.56
DeepEyes-7B	54.20	36.08	30.56	19.92	42.58	24.58
Thyme-RL-7B	69.61	44.31	30.21	26.17	35.94	22.88
Geo-Vista-7B (*ours*)	**92.64**	**79.60**	**72.68**	**79.49**	**72.27**	**44.92**

5.2 Main Results

Our experimental results demonstrate GeoVista's superior performance across metrics on GeoBench, as shown in Table 2. We report results at multiple geographical levels and additionally provide city-level accuracy on the GeoBench data types (i.e., panorama, photo, and satellite images). Across these metrics, GeoVista achieves state-of-the-art performance among open-source models. We also find that Gemini-2.5-pro achieves the best overall performance on GeoBench among its closed-source counterparts.

It is worth noting that, despite having far fewer parameters, GeoVista performs on par with closed-source models on most metrics. We attribute this performance to GeoVista's learned reasoning prior and its ability to use tool calls, especially the web-search tool. This demonstrates the effectiveness of GeoVista's reasoning capabilities, which extend beyond simple visual grounding.

Table 3: **Nuanced distance statistics of different models' performance on GeoBench.** The **bold** figures indicate the best performance among closed-source and open-source models.

Models	<3km (%) ↑	Median Distance (km) ↓
Closed-source Models		
Gemini-2.5-pro	**64.45**	**0.80**
GPT-5	55.12	1.86
Seed-VL-1.6	54.00	2.22
Gemini-2.5-flash	58.11	1.67
Open-source Models		
Qwen2.5-VL-7B	29.30	2209.82
Mini-o3-7B	9.57	13043.70
DeepEyes-7B	26.86	5174.93
Thyme-RL-7B	29.88	880.97
Geo-Vista-7B (*ours*)	**52.83**	**2.35**

We also conduct the **nuanced evaluation** of model predictions as shown in Tab.3. We find that GeoVista achieves high precision for real-world geolocalization. For the two nuanced metrics we report—the rate of haversine distance $< 3\,\text{km}$ and the median haversine distance (Tab.3)—GeoVista, while leaving a small gap to closed-source models, substantially outperforms other open-source models that think with images with the same tool access, highlighting its superior reasoning performance.

5.3 Analysis

5.3.1 RQ1: The Ablation Study

Table 4: **The Ablation Study.** Ablations on cold-start SFT, RL, and hierarchical rewards show SFT and RL are both indispensable, while hierarchical rewards further enhance multi-turn geolocalization accuracy on GeoBench.

Models	Median Distance (km) ↓	City (%) (Panorama) ↑	City (%) (Photo) ↑	City (%) (Satellite) ↑
Qwen-2.5-VL	2209.82	24.22	44.73	16.1
w/o Cold Start	55.32	48.52	43.63	27.46
w/o RL	11.17	54.88	57.23	29.66
w/o HR	4.11	75.0	68.95	40.68
Geo-Vista-7B	**2.35**	**79.49**	**72.27**	**44.92**

We present an ablation study to quantify the contribution of each component. The overall results appear in Table 4. Unless otherwise stated, we keep the same training hyperparameters and evaluation settings.

Cold Start (SFT) To assess the necessity of cold-start SFT, 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.

Hierarchical Reward (HR) We also evaluate the necessity of the hierarchical reward. We keep both the cold-start SFT and reinforcement learning, but disable the hierarchical reward during the RL stage, using only a city-level reward. The results confirm the importance of hierarchical reward.

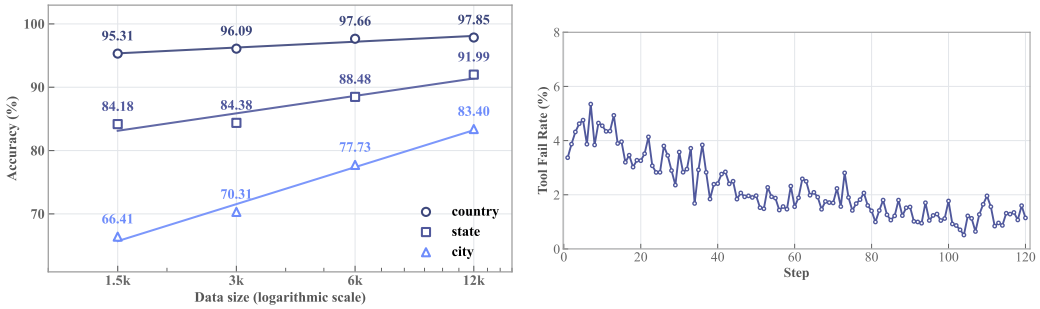

Figure 7: LEFT: **The performance on the panorama validation set during the RL stage.** We observe nearly log-linear performance gains on the 512-panorama validation set. RIGHT: **The tool fail rate during RL training.** The model's erroneous tool-call rate decreases during RL, suggesting it learns to avoid invalid or malformed calls, leading to improved performance.

5.3.2 RQ2: The Scaling Effect in 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 1,500, 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.7-LEFT), we observe a nearly perfect data-scaling effect.

5.3.3 RQ3: Failure Tool Calls during RL

To further analyze the model's behavior regarding tool calls during RL training, we record the error tool-call rate. Error tool calls typically arise from invalid crop-tool bounding-box parameters (e.g., `x_1` greater than `x_2` in `bbox_2d`) or incomplete `json` format tool-calls. An interesting observation is that, although we do not directly optimize tool-call behavior during RL, the model gradually produces fewer erroneous tool calls, showing a clear decreasing trend in error rate as training progresses (Fig.7-RIGHT). We hypothesize that erroneous tool calls reduce the model's likelihood of reaching the correct answer within limited turns, leading the model to implicitly learn to avoid such errors in its reasoning trajectories.

6 Conclusion

Our research focuses on a challenging task—real-world geolocalization—which requires searching for fine-grained visual clues and integrating external knowledge. We propose **GeoVista**, an agentic model capable of visual reasoning and tool use, including crop–zoom-in and web-search tools for deep, multi-step reasoning. To rigorously evaluate and obtain comprehensive metrics for real-world geolocalization, we introduce **GeoBench**, a benchmark containing 1,142 high-resolution images from diverse global locations and three distinct data types. We curate reasoning trajectories for both cold-start supervised fine-tuning and reinforcement learning to further enhance reasoning and tool-use capabilities. We also propose a hierarchical reward to provide nuanced supervision during reinforcement learning. Experimental results show that **GeoVista** outperforms other open-source baselines and achieves performance comparable to closed-source models such as **GPT-5** and **Gemini-2.5-flash** on most metrics. Furthermore, we conduct detailed analyses for deeper insights. We believe this work lays a solid foundation for future research on agentic visual reasoning and real-world geolocalization.

A Raw Data Collection

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

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:

- **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.

B Case Study

Figure 9: **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.

References

Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, and Loic Landrieu. 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.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

FutureSearch, :, Nikos I. Bosse, Jon Evans, Robert G. Gambee, Daniel Hnyk, Peter Mühlbacher, Lawrence Phillips, Dan Schwarz, and Jack Wildman. Deep research bench: Evaluating ai web research agents, 2025. URL https://arxiv.org/abs/2506.06287.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 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.

James Hays and Alexei A. Efros. 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.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (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-Conference.html.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. 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.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability, 2025. URL https://arxiv.org/abs/2504.21776.

Peter Mühlbacher, Nikos I. Bosse, and Lawrence Phillips. 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.

Xianghe Pang, Shuo Tang, Rui Ye, Yuwen Du, Yaxin Du, and Siheng Chen. Browsemaster: Towards scalable web browsing via tool-augmented programmatic agent pair, 2025. URL https://arxiv.org/abs/2508.09129.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 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.

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. *CoRR*, abs/2502.13759, 2025. doi: 10.48550/ARXIV.2502.13759. URL https://doi.org/10.48550/arXiv.2502.13759.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. 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.

Gemini Team. 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.

Nam N. Vo, Nathan Jacobs, and James Hays. 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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Ye Wang, Qianglong Chen, Zejun Li, Siyuan Wang, Shijie Guo, Zhirui Zhang, and Zhongyu Wei. Simple o3: Towards interleaved vision-language reasoning, 2025. URL https://arxiv.org/abs/2508.12109.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (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-Conference.html.

Tobias Weyand, Andre F. de Araújo, Bingyi Cao, and Jack Sim. 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.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Roy Xie, David Qiu, Deepak Gopinath, Dong Lin, Yanchao Sun, Chong Wang, Saloni Potdar, and Bhuwan Dhingra. Interleaved reasoning for large language models via reinforcement learning, 2025. URL https://arxiv.org/abs/2505.19640.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Yankai Chen, Chunkit Chan, Peilin Zhou, Xinyang Zhang, Chenwei Zhang, Jingbo Shang, Ming Zhang, Yangqiu Song, Irwin King, and Philip S. Yu. From web search towards agentic deep research: Incentivizing search with reasoning agents, 2025a. URL https://arxiv.org/abs/2506.18959.

Yifan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. *CoRR*, abs/2508.11630, 2025b. doi: 10.48550/ARXIV.2508.11630. URL https://doi.org/10.48550/arXiv.2508.11630.

Yizhen Zhang, Yang Ding, Shuoshuo Zhang, Xinchen Zhang, Haoling Li, Zhong-zhi Li, Peijie Wang, Jie Wu, Lei Ji, Yelong Shen, Yujiu Yang, and Yeyun Gong. Perl: Permutation-enhanced reinforcement learning for interleaved vision-language reasoning. *CoRR*, abs/2506.14907, 2025c. doi: 10.48550/ARXIV.2506.14907. URL https://doi.org/10.48550/arXiv.2506.14907.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 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.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents, 2024. URL https://arxiv.org/abs/2307.13854.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5316–5325, 2020. URL https://api.semanticscholar.org/CorpusID:227151840.