

GeoVista: Web-Augmented Agentic Visual Reasoning for Geolocalization

Anonymous CVPR submission

Paper ID 9358

Abstract

001 *Current research on agentic visual reasoning enables*
002 *deep multimodal understanding but primarily focuses on*
003 *image manipulation tools, leaving a gap toward more*
004 *general-purpose agentic models. In this work, we revisit*
005 *the geolocalization task, which requires not only nuanced*
006 *visual grounding but also web search to confirm or refine*
007 *hypotheses during reasoning. Since existing geolocaliza-*
008 *tion benchmarks fail to meet the need for high-resolution*
009 *imagery and the localization challenge for deep agentic rea-*
010 *soning, we curate **GeoBench**, a benchmark that includes*
011 *photos and panoramas from around the world, along with*
012 *a subset of satellite images of different cities to rigorously*
013 *evaluate the geolocalization ability of agentic models. We*
014 *also propose **GeoVista**, an agentic model that seamlessly*
015 *integrates tool invocation within the reasoning loop, includ-*
016 *ing an image-zoom-in tool to magnify regions of interest and*
017 *a web-search tool to retrieve related web information. We*
018 *develop a complete training pipeline for it, including a cold-*
019 *start supervised fine-tuning (SFT) stage to learn reasoning*
020 *patterns and tool-use priors, followed by a reinforcement*
021 *learning (RL) stage to further enhance reasoning ability.*
022 *We adopt a hierarchical reward to leverage multi-level geo-*
023 *graphical information and improve overall geolocalization*
024 *performance. Experimental results show that GeoVista sur-*
025 *passes other open-source agentic models on the geolocali-*
026 *zation task greatly and achieves performance comparable*
027 *to closed-source models such as Gemini-2.5-flash and GPT-*
028 *5 on most metrics.*

24, 28, 31], the OpenAI o3 model [13] 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 [25] with image manipulation and other tool invocations. Some follow-up works [9, 32] 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.

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 [5, 10, 30]. 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 [11, 14, 34] 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

1. Introduction

030 Recent advances in Vision-Language Models (VLMs) [2,
031 23, 27] enable deep reasoning over multimodal queries by
032 invoking image-centric tools and utilizing long Chain-of-
033 Thought approaches [8, 17], allowing these models to han-
034 dle much more complex tasks. Some recent works [20, 33]
035 attempt to integrate seamless tool invocation into multi-turn
036 interaction through reinforcement learning.
037 Among the latest multimodal reasoning milestones [6,

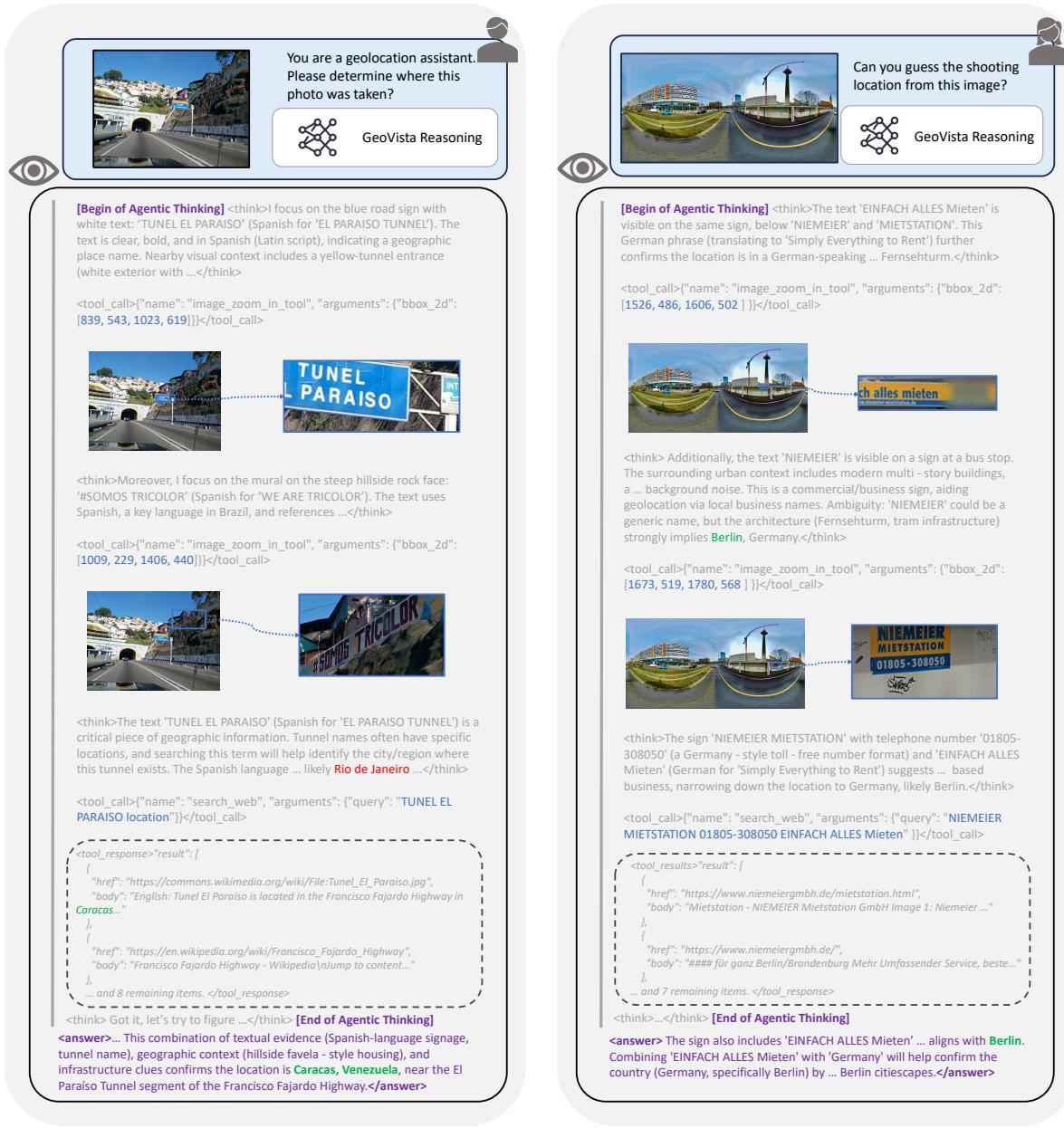


Figure 1. **Agentic thinking of GeoVista for Real-world geolocation.** During the reasoning loop, our GeoVista seamlessly integrates the image-zoom-in tool to magnify regions of interest and the web-search tool to retrieve relevant information. This web-augmented visual reasoning process enables GeoVista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

079
080
081
082
083
084
085
086

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 invoca-

cation proposals.

087
088
089
090
091
092
093
094

Second is the reinforcement learning to further incentivize reasoning ability [4]. We apply group relative policy optimization (GRPO) [18] 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 [17] introduced localized intermediate steps (e.g., boxes/regions) to guide attention; Visual Sketchpad [8] 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 [13] marked a watershed by productizing tool-mediated visual reasoning inside the chain (zoom, crop, rotate), triggering open replications.

After the emergence of OpenAI o3 [13], Thyme [32] extends this paradigm with a code-executing visual sandbox that emits and runs image operators; mini-o3 [9] trains an agent to alternate “think–act” cycles with iterative region selection and overturn masking, scaling to deep multi-turn search; OpenThinkIMG [20] unifies detectors, OCR, and drawing tools under a standardized controller with RL-learned tool policies; and DeepEyes [33] 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 [7] and curated YFCC100M subsets [22], emphasizing retrieval and metric learning. Landmark-centric recognition leveraged Google Landmarks v2 [26], improving precision where distinc-

tive structures exist. Cross-view methods advanced with VIGOR [35], stressing generalization across cities for ground-to-aerial matching. Scaling to worldwide street scenes, OpenStreetView-5M [1] enabled training and fair evaluation at unprecedented diversity and size. Complementing purely visual supervision, GeoComp [19] 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 geolocalization metadata, including precise latitude and longitude. We state how we collect the raw data in Sup.A and compare our GeoBench with other existing geolocalization benchmarks in Sup.C. We also conduct localizability filtering to remove non-localizable images and easily localizable landmarks (see Sup.B).

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.

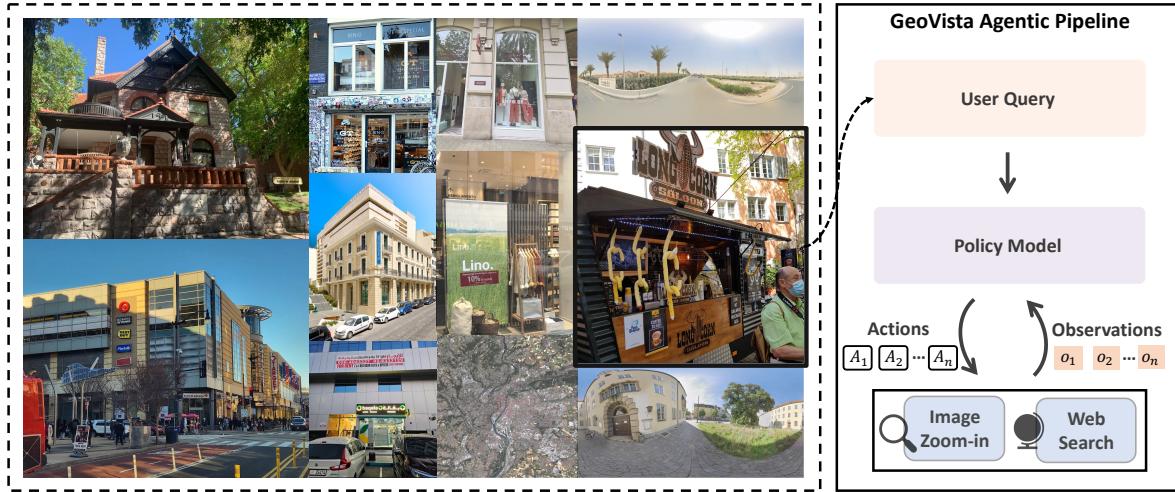


Figure 2. **Image examples from GeoBench and the training data, and the agentic pipeline of GeoVista.** Given a query and image, the policy model iteratively generates thoughts and actions; each action is parsed, executed, and yields a new observation, repeating this loop until it outputs a final geolocation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

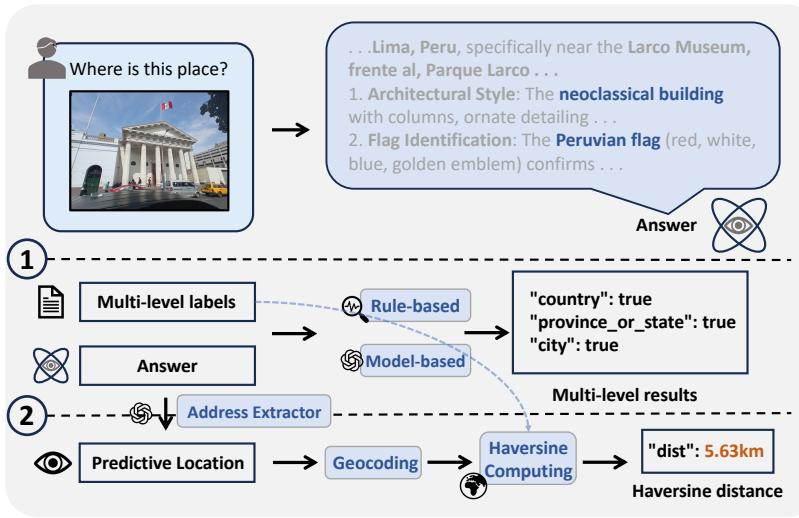


Figure 3. LEFT: **The evaluation pipeline of GeoBench dataset.** The evaluation system consists of (1) Level-wise evaluation, which employs both rule-based and model-based verifiers to determine correctness at different administrative levels, and (2) nuanced evaluation, which extracts the predicted address, 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.

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.3, for each response we extract the predicted textual location and convert it into geographic 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 fashion:

202
203
204
205
206
207
208

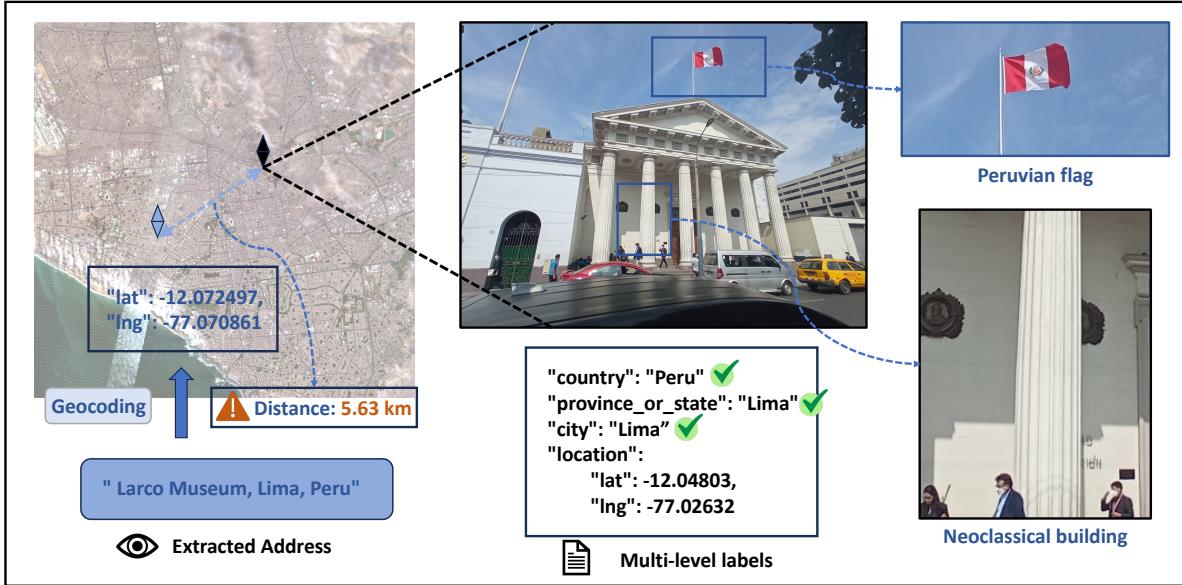


Figure 4. Illustration of GeoBench dataset, along with level-wise and nuanced evaluation.

$$d = 2R_e \arcsin(\sqrt{v}),$$

$$v = \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1) \cos(\phi_2) \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right) \quad (1)$$

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 geolocation ability. To this end, we sample 512 standard photos, 512 panoramas, and 108 satellite images from the raw data 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. 3).

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.

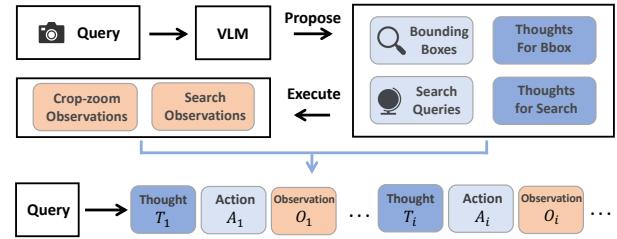


Figure 5. Thinking trajectory curation. We mimic human geolocation by using a VLM to propose tool calls and rationales, and assemble tool-call reasoning trajectories.

Inspired by how humans identify a place during geolocation—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. 5 we use a VLM (*Seed-1.6-vision* [16]) 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.

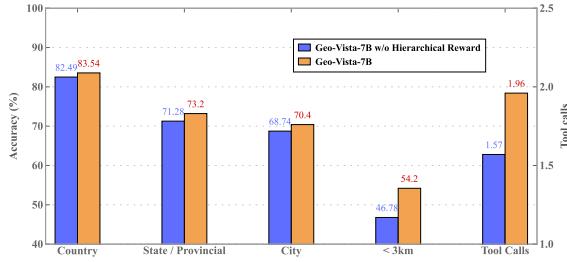


Figure 6. The comparison of GeoVista-7B and its counterpart w/o Hierarchical Reward.

3.4. Reinforcement Learning

We apply a vanilla GRPO [18] 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:

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

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

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

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

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* [15] 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 *ver1* for GRPO [18] 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 [3]. 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** [21], **GPT-5** [12], and **Seed-1.6-vision** [16]—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** [9], **DeepEyes-7B** [33], and **Thyme-RL-7B** [32], which are sharing the same 7B parameter size as our **GeoVista**. We also use the base model *Qwen2.5-VL-Instruct* [15] 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 [29] pattern of tool calls in multi-turn interactions. For the closed-source models like GPT-5 [12], 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 con-

Table 1. **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

Table 2. **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) ↓
Close-sourced Models		
Gemini-2.5-pro	64.45	0.80
GPT-5	55.12	1.86
Seed-VL-1.6	54.0	2.22
Gemini-2.5-flash	58.11	1.67
Open-sourced Models		
Qwen2.5-VL-7B	29.30	2209.82
Mini-o3-7B	9.57	13043.7
DeepEyes-7B	26.86	5174.93
Thyme-RL-7B	29.88	880.97
Geo-Vista-7B (ours)	52.83	2.35

duct 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 geolocation 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 [9] 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.

5.2. Main Results

Our experimental results demonstrate GeoVista’s superior performance across metrics on GeoBench, as shown in Ta-

ble 1. 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.

We also conduct the nuanced evaluation of model predictions as shown in Tab.2. We find that GeoVista achieves high precision for real-world geolocation. For the two nuanced metrics we report—the rate of haversine distance < 3 km and the median haversine distance (Tab.2)—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.

5.3. Analysis

5.3.1. RQ1: The Ablation Study

Table 3. **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 geolocation 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 3.

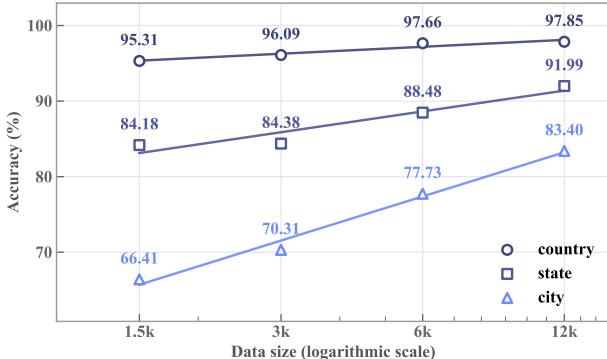


Figure 7. **The Performance on the panorama validation set during RL stage.** We scale RL data from 1.5k to 12k samples and consistently observe nearly log-linear performance gains on the 512-panorama validation set.

364 Unless otherwise stated, we keep the same training hyper-
365 parameters and evaluation settings.

366 **Cold Start (SFT)** To assess the necessity of cold-start
367 SFT, we remove the cold-start stage and conduct reinforce-
368 ment learning directly. The results show that cold-start
369 SFT is essential for multi-turn tool use, as performance on
370 **GeoBench** collapses without it.

371 **Reinforcement Learning (RL)** To examine the ne-
372 cessity of reinforcement learning, we remove the RL and
373 only conduct the cold-start SFT. The results show that SFT
374 alone is not sufficient: although the model learns a reason-
375 ing prior, it requires reinforcement learning to incentivize
376 and strengthen its reasoning capability.

377 **Hierarchical Reward (HR)** We also evaluate the ne-
378 cessity of the hierarchical reward. We keep both the cold-
379 start SFT and reinforcement learning, but disable the hier-
380 archical reward during the RL stage, using only a city-level
381 reward. The results confirm the importance of hierarchical
382 reward.

383 5.3.2. RQ2: The Scaling Effect in RL Stage

384 We hypothesize that model performance increases as the
385 data size grows. Since RL data do not require reasoning-
386 trajectory annotations, we can easily scale the RL dataset
387 to 12k samples. We apply different RL data sizes, in-
388 cluding 1,500, 3k, 6k, and 12k, using the same cold-start
389 SFT checkpoint. We report performance on a validation set
390 consisting of 512 panoramas. The results show that per-
391 formance consistently improves as the data size increases.
392 When plotting data size on a logarithmic scale against per-
393 formance (Fig.7), we observe a nearly perfect data-scaling
394 effect.

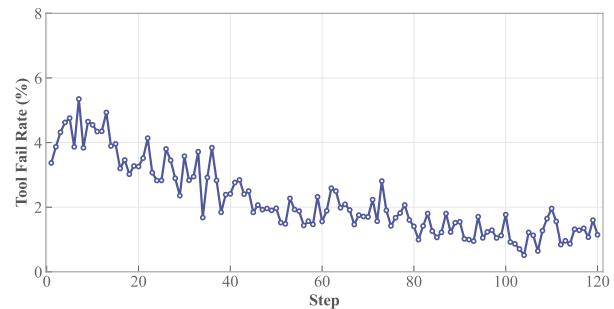


Figure 8. **The tool fail rate during RL training.** During RL, the model’s erroneous tool-call rate decreases, suggesting it learns to avoid invalid or malformed calls and improve performance.

395 5.3.3. RQ3: Failure Tool Calls during RL

396 To further analyze the model’s behavior regarding tool calls
397 during RL training, we record the error tool-call rate. Error
398 tool calls typically arise from invalid crop-tool bounding-
399 box parameters (e.g., x_1 greater than x_2 in $bbox_2d$) or
400 incomplete json format tool-calls. An interesting observa-
401 tion is that, although we do not directly optimize tool-call
402 behavior during RL, the model gradually produces fewer er-
403 roneous tool calls, showing a clear decreasing trend in error
404 rate as training progresses (Fig.8). We hypothesize that er-
405 roneous tool calls reduce the model’s likelihood of reaching
406 the correct answer within limited turns, leading the model
407 to implicitly learn to avoid such errors in its reasoning tra-
408 jectories.

409 6. Conclusion

410 Our research focuses on a challenging task—real-world
411 geolocation—which requires searching for fine-grained
412 visual clues and integrating external knowledge. We pro-
413 pose **GeoVista**, an agentic model capable of visual reason-
414 ing and tool use, including crop-zoom-in and web-search
415 tools for deep, multi-step reasoning. To rigorously eval-
416 uate and obtain comprehensive metrics for real-world geolo-
417 calization, we introduce **GeoBench**, a benchmark contain-
418 ing 1,142 high-resolution images from diverse global loca-
419 tions and three distinct data types. We curate reasoning tra-
420 jectories for both cold-start supervised fine-tuning and rein-
421 forcement learning to further enhance reasoning and tool-use ca-
422 pabilities. We also propose a hierarchical reward to provide
423 nuanced supervision during reinforcement learning. Experi-
424 mental results show that **GeoVista** outperforms other open-
425 source baselines and achieves performance comparable to
426 closed-source models such as **GPT-5** and **Gemini-2.5-flash**
427 on most metrics. Furthermore, we conduct detailed analy-
428 ses for deeper insights. We believe this work lays a solid
429 foundation for future research on agentic visual reasoning
430 and real-world geolocation.

GeoVista: Web-Augmented Agentic Visual Reasoning for Geolocalization

Supplementary Material

431

A. Raw Data Collection

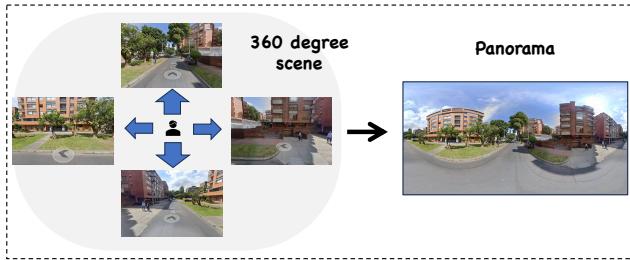


Figure 9. The panorama pipeline in GeoBench and GeoVista training data.

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

- 436 • **Normal Photos.** To obtain high-quality photographs of
 437 diverse scenarios (e.g., libraries, supermarkets, suburban
 438 areas), we collect photos from the internet. These photos
 439 typically have least a resolution of 1600×1200 .
- 440 • **Panoramas.** The source data are 360° street-view scenes
 441 from cities across the globe. To make them compatible
 442 with multimodal LLM input, we convert them into planar
 443 panoramas by stitching tiles retrieved via the Mapillary
 444 API and assembling them locally. To balance detail with
 445 storage, each panorama is fixed at a resolution of $4096 \times$
 446 2048 .
- 447 • **Satellite Images.** The typical size of our satellite images
 448 is 2000×2000 . We retrieve recent Sentinel-2 Level-2A
 449 imagery for cities worldwide from the Microsoft Planetary
 450 Computer, mosaic several low-cloud scenes within
 451 each city's bounding box, and save multiple images to-
 452 gether with their metadata from different viewport vari-
 453 ants.

B. Localizability Filtering

455 We believe that images collected from the Internet exhibit
 456 varying levels of localizability [1], especially when the data
 457 types and sources differ. Therefore, we remove two cate-
 458 gories of data via model-based filtering:

- 459 • **Non-localizable images.** These images usually lack
 460 identifiable geographical clues and contain generic ob-
 461 jects or scenes, such as close-up food photos, indoor
 462 rooms, plain natural landscapes, or single animals. Such
 463 content provides almost no regional or cultural context,
 464 making localization infeasible.



Figure 10. Localizable vs Non-Localizable. We remove the non-localizable (orange) and the landmarks (purple) from GeoBench, leaving only localizable images for rigorously evaluating models.

- 465 • **Easily localizable landmarks.** These images contain
 466 strong geographic priors, typically featuring iconic land-
 467 marks or globally recognizable sites. Since VLMs have
 468 likely encountered such images multiple times during
 469 pretraining, including them would make geolocation triv-
 470 ial and fail to reflect genuine reasoning ability.

C. Benchmark Comparison

Table 4. Comparison across geolocation datasets.

Benchmark	Year	GC	RC	HR	DV	NE
Im2GPS [7]	2008	✓				
YFCC4k [22]	2017	✓				
Google Landmarks v2 [26]	2020	✓				
VIGOR [35]	2022					✓
OSV-5M [1]	2024	✓	✓			✓
GeoComp [19]	2025	✓	✓			✓
GeoBench (ours)	2025	✓	✓	✓	✓	✓

We assess benchmarks along the following axes:

- 472 • **Global Coverage.** Whether the benchmark contains im-
 473 ages from across the globe, ensuring that the model does
 474 not overfit or bias its performance toward specific regions.
- 475 • **Reasonable Localizability.** Whether the benchmark fil-
 476 ters out non-localizable images or easily localizable land-
 477 marks to maintain meaningful localization difficulty.
- 478 • **High Resolution.** Whether all images in the benchmark
 479 have at least 1 M pixels to support reliable visual clue ex-
 480 traction and grounding.
- 481 • **Data Variety.** Whether the benchmark includes two or
 482 more types of images to test the generalizability of rea-
 483 soning models under varying data conditions.
- 484 • **Nuanced Evaluation.** Whether the benchmark includes
 485 geolocation coordinates to enable haversine distance
 486 computation for nuanced evaluation.

D. Case Study



Please analyze where is the place.

 GeoVista Reasoning

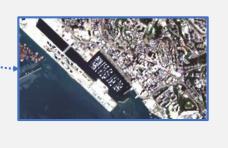
[Begin of Agentic Thinking] <think>I focus on the upper right portion of the image, which showcases a prominent city with an extensive urban layout and a dense network of roads. This urban area is situated on the coast, adjacent to a large body of water that appears to be a bay. The coastline is ... </think>

```
<tool_call>{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [645, 97, 1385, 471]}}</tool_call>
```



<think>The isolated white patch in this zoomed satellite image appears to be a large man-made structure, likely a public or private park, sports facility, or industrial complex. Its distinct shape contrasts sharply with the surrounding densely packed urban grid and waterways. The location is a major coastal city, consistent with the original image's context (e.g., Hong Kong). This feature could be Victoria Park (a historic park) or a modern waterfront development. The clarity and prominence of the white area suggest it holds regional significance. Analyzing this specific feature can reveal details about the urban planning, natural-human interaction, or cultural infrastructure of the city, providing deeper insight into its geography, economy, and social fabric.</think>

```
<tool_call>{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [1066, 235, 1297, 353]}}</tool_call>
```



<think>Got it, let's try to figure out the location. First, the original image shows a large bay with a city on the coast, surrounded by mountains. This is characteristic of Hong Kong. The zoomed images show a city with an intricate road network, a major harbor with shipping activity, and a large white feature that looks like a harbor or port area. The streetscape, the bay, and the mountainous terrain all point to Hong Kong. The white feature could be a harbor, like Victoria Harbour. So putting it together, the original image is likely Hong Kong, specifically the Kowloon Peninsula area, with the HK International Airport (visible in the top left) and the main urban core of **Hong Kong** Island.</think> **[End of Agentic Thinking]**

<answer>The original image depicts **Hong Kong, China**. Key evidence:
- Urban Layout ... align with Hong Kong's urban planning and geography.

This matches Hong Kong's iconic coastal layout, with Kowloon Peninsula stretching into Victoria Harbour.</answer>



Please analyze where is the place.

 GeoVista Reasoning

[Begin of Agentic Thinking] <think>I focus on the black signboard above the entrance, which displays Chinese characters: 黃南皓珺博物館. Transcription: 'Yunnan Hua Gu Museum' (transliteration: 'Yūnnán Huàgǔ Bówùguǎn'). The text is yellow on black, traditional Chinese script (traditional for Yunnan). This signage likely identifies a museum's name. The building has ... suggesting a cultural....</think>

```
<tool_call>{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [253, 69, 1074, 353]}}</tool_call>
```



<think>The museum's name '云南皓珺博物馆' (Yunnan Hua Gu Museum) suggests a museum in Yunnan Province, China. The architectural style (gray stone facade, red doors) aligns with Yunnan's historic districts, likely Kunming. Searching this museum name will reveal its location, confirming if it's in Kunming (e.g., part of the city's heritage sites) and provide details about its exhibits or significance, helping determine the geographic location from the visual cues like stone walls, red doors, and urban setting.</think>

```
<tool_call>{"name": "search_web", "arguments": {"query": "云南皓珺博物馆 位置"}}</tool_call>
```

```
<tool_response>"result": [
  {
    "href": "https://kunming.city8.com/zhuanyejigou/60034244_l1hc",
    "body": "【云南皓珺博物馆】地址:电话:公交指南:位置示意图:地图位置:周边酒店:昆明专业机构黄页:昆明地图...\n"
  },
  {
    "href": "https://map.baidu.com/place/394a88d86219bcc53062128d",
    "body": "云南皓珺博物馆:地址:云南省昆明市西山区敬德巷附5号...\n"
  },
  {
    "href": "https://m.city8.com/kunming/scope/7ufcyd730qpvbdd973_address",
    "body": "【昆明市西山区敬德巷附5号云南皓珺博物馆】详细位置:位置示意图:地图位置:公交指南:附近酒店:昆明地址黄页:昆明地图\n...\n"
  }
], and 7 remaining items. </tool_response>
```

<think>...The sign says "云南皓珺博物馆" (Yunnan Haojun Museum) and the architectural style (gray stone, red doors) matches Yunnan's historic districts. From the search results, the museum is in **Kunming, Yunnan**, at 5 Jinding Lane, Xishan District...</think> **[End of Agentic Thinking]**

<answer>The original image is located at 云南皓珺博物馆 (Yunnan Haojun Museum) in **Kunming, Yunnan Province, China**. Key evidence: 1. Museum Signage: The black sign ...</answer>

Figure 11. **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.

489

References

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

- [1] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronsohn, 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)*, pages 21967–21977, 2024. 3, 1
- [2] 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 Wenhui Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 1
- [3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. 6
- [4] 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, Shanghai 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, Wan-jia 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, Zhehan 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. 2
- [5] 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. 1
- [6] 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*, pages 19520–19529. Computer Vision Foundation / IEEE, 2025. 1
- [7] 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. 3, 1
- [8] 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 *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. 1, 3
- [9] 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. 1, 3, 6, 7
- [10] 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. 1
- [11] Peter Mühlbacher, Nikos I. Bosse, and Lawrence Phillips. Towards a realistic long-term benchmark for open-web research agents, 2024. 1
- [12] OpenAI. Introducing gpt-5. OpenAI Blog, 2025. Accessed: 2025-11-13. 6
- [13] OpenAI. Introducing openai o3 and o4-mini, 2025. Accessed: 2025-11-13. 1, 3
- [14] 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. 1
- [15] 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, Yuan

- 603 Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang
604 Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang,
605 Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan
606 Qiu. Qwen2.5 technical report, 2025. 6
- 607 [16] ByteDance Seed. Seed1.6 vision. Official Site, 2025. Ac-
608 cessed: 2025-11-13. 5, 6
- 609 [17] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan
610 Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot:
611 Advancing multi-modal language models with a comprehen-
612 sive dataset and benchmark for chain-of-thought reasoning,
613 2024. 1, 3
- 614 [18] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao
615 Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo.
616 Deepseekmath: Pushing the limits of mathematical reason-
617 ing in open language models. *CoRR*, abs/2402.03300, 2024.
618 2, 6
- 619 [19] Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet,
620 Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and
621 Xiuying Chen. Geolocation with real human gameplay data:
622 A large-scale dataset and human-like reasoning framework.
623 *CoRR*, abs/2502.13759, 2025. 3, 1
- 624 [20] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao,
625 Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Jun-
626 tao Li, Xiaoye Qu, and Yu Cheng. Openthinkimg: Learning
627 to think with images via visual tool reinforcement learning.
628 *CoRR*, abs/2505.08617, 2025. 1, 3
- 629 [21] Gemini Team. Gemini 2.5: Pushing the frontier with ad-
630 vanced reasoning, multimodality, long context, and next gen-
631 eration agentic capabilities. *CoRR*, abs/2507.06261, 2025. 6
- 632 [22] Nam N. Vo, Nathan Jacobs, and James Hays. Revisiting
633 im2gps in the deep learning era. 2017 IEEE International
634 Conference on Computer Vision (ICCV), pages 2640–2649,
635 2017. 3, 1
- 636 [23] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan,
637 Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin
638 Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui
639 Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Jun-
640 yang Lin. Qwen2-vl: Enhancing vision-language model’s
641 perception of the world at any resolution, 2024. 1
- 642 [24] Ye Wang, Qianglong Chen, Zejun Li, Siyuan Wang, Shijie
643 Guo, Zhirui Zhang, and Zhongyu Wei. Simple o3: Towards
644 interleaved vision-language reasoning, 2025. 1
- 645 [25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten
646 Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and
647 Denny Zhou. Chain-of-thought prompting elicits reasoning
648 in large language models. In *Advances in Neural Infor-
649 mation Processing Systems 35: Annual Conference on Neural
650 Information Processing Systems 2022, NeurIPS 2022, New
651 Orleans, LA, USA, November 28 - December 9, 2022*. 2022.
652 1
- 653 [26] Tobias Weyand, Andre F. de Araújo, Bingyi Cao, and Jack
654 Sim. Google landmarks dataset v2 – a large-scale benchmark
655 for instance-level recognition and retrieval. 2020 IEEE/CVF
656 Conference on Computer Vision and Pattern Recognition
657 (CVPR), pages 2572–2581, 2020. 3, 1
- 658 [27] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu,
659 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. 1
660 661 662 663 664 665
[28] Roy Xie, David Qiu, Deepak Gopinath, Dong Lin, Yanchao
Sun, Chong Wang, Saloni Potdar, and Bhuwan Dhingra. In-
terleaved reasoning for large language models via reinfor-
celearning, 2025. 1
666 667 668 669
[29] 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 In-
ternational Conference on Learning Representations, ICLR
2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net,
2023. 6
670 671 672 673 674 675
[30] 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: Incentiviz-
ing search with reasoning agents, 2025. 1
676 677 678 679 680 681 682 683
[31] 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 inter-
leaved vision-language reasoning. *CoRR*, abs/2506.14907,
2025. 1
684 685 686 687 688 689
[32] 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 im-
ages. *CoRR*, abs/2508.11630, 2025. 1, 3, 6
690 691 692 693 694 695
[33] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao,
Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deep-
eyes: Incentivizing “thinking with images” via reinforce-
ment learning. *CoRR*, abs/2505.14362, 2025. 1, 3, 6
696 697 698 699
[34] 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. We-
barena: A realistic web environment for building au-
tonomous agents, 2024. 1
700 701 702 703 704
[35] 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 Pat-
tern Recognition (CVPR), pages 5316–5325, 2020. 3, 1
705 706 707 708