GeoBrowse: A Geolocation Benchmark for Agentic Tool Use with Expert-Annotated Reasoning Traces

Xinyu Geng*, Yanjing Xiao*, Yuyang Zhang*, Hanwen Wang, Xinyan Liu, Rui Min, Tianqing Fang, and Yi R. Fung†

MMSense Lab
The Hong Kong University of Science and Technology
xgengad@connect.ust.hk, yrfung@ust.hk

Abstract. Deep research agents integrate fragmented evidence through multi-step tool use. BrowseComp offers a text-only testbed for such agents, but existing multimodal benchmarks rarely require both weak visual cues composition and BrowseComp-style multi-hop verification. Geolocation is a natural testbed because answers depend on combining multiple ambiguous visual cues and validating them with open-web evidence. Thus, we introduce **GeoBrowse**, a geolocation benchmark that combines visual reasoning with knowledge-intensive multi-hop queries. **Level 1** tests extracting and composing fragmented visual cues, and **Level 2** increases query difficulty by injecting long-tail knowledge and obfuscating key entities. To support evaluation, we provide an agentic workflow **GATE** with five think-with-image tools and four knowledge-intensive tools, and release expert-annotated stepwise traces grounded in verifiable evidence for trajectory-level analysis. Experiments show that GATE outperforms direct inference and open-source agents, indicating that no-tool, search-only or image-only setups are insufficient. Gains come from coherent, level-specific tool-use plans rather than more tool calls, as they more reliably reach annotated key evidence steps and make fewer errors when integrating into the final decision. The GeoBrowse benchmark and codes are provided in https://github.com/ornamentt/GeoBrowse.

Keywords: Agentic tool use · Multimodal reasoning · Geolocation · Open-web information seeking

1 Introduction

Real-world tasks often require gathering scattered evidence from the open web and synthesizing it into a supported conclusion [18, 32]. Deep research agents tackle this by coupling Large Language Models (LLMs) with tools to plan actions, seek information, and verify evidence [14, 34]. To evaluate such agents,

* Equal contribution
† Corresponding author

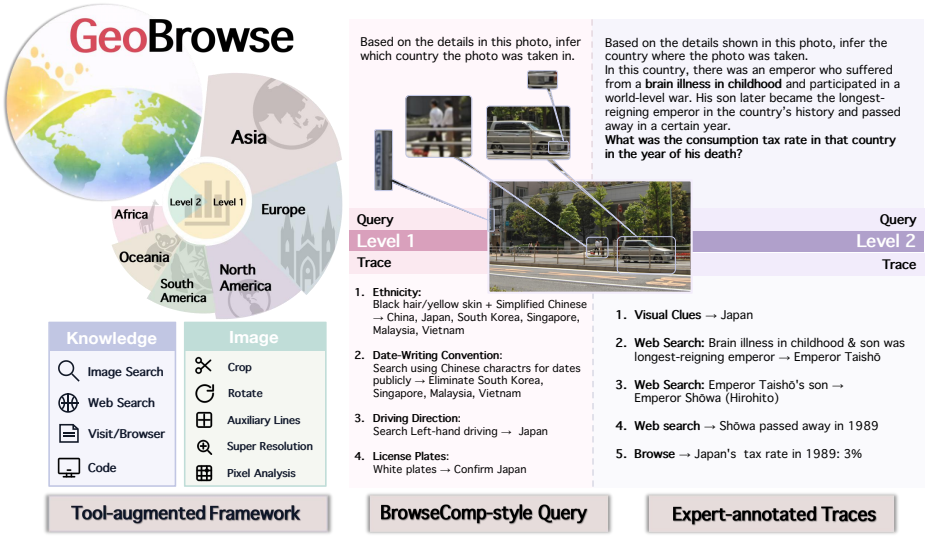

Fig. 1: GeoBrowse couples a tool-use framework with a geolocation benchmark: Level 1 emphasizes visual cue composition, while Level 2 contains BrowseComp-style queries, all paired with expert-annotated stepwise traces.

BrowseComp [49] provides a text-only benchmark where evidence is intentionally fragmented and obfuscated [12], so success depends on multi-hop browsing and verification [16, 23]. Yet, many real-world tasks are multimodal and require extracting weak visual cues before open-web evidence gathering [11, 21]. However, existing multimodal counterparts of BrowseComp either emphasize harder visual perception [46, 53] or treat the image as a single step within a largely text-driven reasoning chain [13, 25, 27]. Few benchmarks jointly stress **ambiguous visual cues** and **hard, evidence-grounded multi-hop query**, requiring agents to manipulate images to extract weak cues and then verify them through image-based retrieval, together with other tools to retrieve long-tail knowledge.

Geolocation is well suited for reasoning over ambiguous visual clues. Street-view images often contain multiple weak clues such as signs, road markings, and architectural style, where no single clue is decisive and candidates must be narrowed by combination [**?**]. Solving such cases also requires iterative reasoning and open-web verification beyond one-shot retrieval [29, 43]. Together, these properties align with BrowseComp, making geolocation a natural substrate for multimodal BrowseComp-style benchmarking.

However, existing geolocation benchmarks remain limited for evaluating agentic tool use and multi-step reasoning, as summarized in Tab. 1. First, many widely used datasets such as MP-16 [22], GLDv2 [50], VIGOR [58], Google-WS [8] and OSV [2], provide only coordinates or coarse labels without expert stepwise traces, which favors vision-only location prediction over tool-driven inference. Second, benchmarks that claim solvability such as GeoVista [48] and

Table 1: Comparison of existing geolocation datasets with GeoBrowse. "Local" denotes region-level coverage, whereas "Global" spans multiple continents. ✓ and ✗ indicate the presence and absence of a feature, respectively; ❍ indicates partial availability.

Dataset	Geographic Coverage	Human Annotation	Reasonable Localizability	CoT	Agentic Tool-use
MP-16 [22]	Local	✗	✗	✗	✗
GLDv2 [50]	Local	✗	✗	✗	✗
VIGOR [58]	Local	✗	✗	✗	✗
Google-WS [8]	Global	✗	✗	✗	✗
OSV [2]	Global	✗	✗	✗	✗
GeoComp [43]	Global	✓	❍	✗	✗
GeoVista [48]	Global	✗	❍	✗	❍
GeoBrowse (ours)	**Global**	✓	✓	✓	✓

GeoComp [43] often filter toward salient landmarks, which removes hard cases that rely on subtle environmental and architectural signals, thus reducing overall difficulty. Third, tool use is rarely evaluated comprehensively. Although recent GeoVista [48] includes zoom-in and web search, it leaves broader tools of image manipulation and browser largely untested.

We introduce **GeoBrowse**, a geolocation benchmark for evaluating agentic tool use that requires both think-with-image ability and open-web information seeking, as shown in Fig. 1. **Level 1** targets visual-first geolocation from weak and fragmented cues. Each instance includes expert stepwise annotations that list actionable visual fragments and required image manipulations. We also minimize reliance on salient landmarks to raise difficulty while remaining solvable. **Level 2** builds on the visual localization of Level 1 and further increases the difficulty of reasoning. We inject long-tail knowledge and rewrite questions into BrowseComp-style with key entities obfuscated, so solving requires multi-step retrieval and verification. Each instance is paired with an expert trace that specifies the intended evidence path. To support comprehensive evaluation of agentic capability, we provide a **G**eolocation **A**gentic-workflow with **T**ool **E**nhancement (**GATE**) with five think-with-image tools (*e.g.*, crop, rotate, pixel analysis, auxiliary lines, super-resolution) and four knowledge-oriented tools (*e.g.*, web text search, web image search, visit, code interpreter).

Experiments show that GATE delivers consistent gains, outperforming direct inference across 12 Multimodal Large Language Models (MLLMs) and 3 open source agents. We find that search-only or image-only tools are insufficient, especially on Level 2 where multi-step retrieval and verification are required. Analysis of tool-use traces and ablation studies suggest that improvements come from coherent, level-specific tool-use plans rather than more tool calls: Level 1 benefits mainly from image processing, while Level 2 relies more on evidence-driven web retrieval and verification. Trajectory analysis further shows that agentic planning improves accuracy by covering more annotated key evidence and reducing final integration errors.

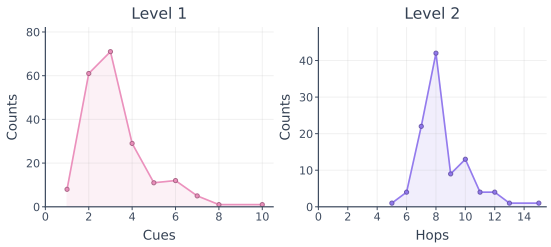

Fig. 2: Distribution of cues and hops on GeoBrowse. Cues count visual cues in Level 1 images, and hops count multi-hop steps in Level 2 queries, quantifying difficulty of visual and knowledge-intensive reasoning.

Table 2: Distribution of target administrative levels. The # and % signify count and percentage, respectively.

Target		Level 1	Level 2
Country	#	92	63
	%	46.2	62.4
State	#	55	19
	%	27.6	18.8
City	#	52	19
	%	26.2	18.8

Contributions. (1) **GeoBrowse dataset:** we present a two-level geolocation benchmark with 300 instances: Level 1 requires composing weak visual cues, and Level 2 introduces BrowseComp-style multi-hop queries with entity obfuscation. (2) **Expert-annotated traces:** we provide a verifiable evidence path with stepwise milestones for every instance, enabling trajectory-level analysis. (3) **Agentic workflow and tool suite:** we introduce **GATE**, an agentic workflow equipped with five think-with-image tools and four knowledge-intensive tools, supporting end-to-end cue extraction and open-web verification. (4) **Evaluation and analysis:** we benchmark 12 MLLMs and open-source agents, conduct single-tool ablations, and analyze milestone hit rates and an error taxonomy.

2 Related Work

Agentic Multimodal Tool Use. Recent advances in autonomous web agents have demonstrated the potential of agentic reasoning with external tools, particularly in open-domain information seeking and synthesis [26, 34, 42, 51, 55]. Extending such agents to multimodal settings further complicates reasoning, as models must integrate visual cues with textual knowledge and external verification. Prior work explores multimodal chain-of-thought prompting and structured visual reasoning [31, 56], as well as grounding reasoning in external knowledge via multimodal Retrieval-Augmented Generation (RAG) [5, 44, 45]. What remains most needed is an agentic tool-use setting that tightly couples image reasoning with web search and verification. Existing benchmarks rarely test this combination: BrowseComp-style multimodal suites primarily derive difficulty from text search, with vision playing a limited role [13, 46, 49], while think-with-image benchmarks emphasize visual manipulation without comparably challenging open-web reasoning [17, 19, 24, 41]. We bridge this gap by grounding high-difficulty, BrowseComp-like queries in geolocation, where both think-with-image operations and web search are essential for resolving fragmented visual evidence.

Geolocation Benchmarks. Early image-based geolocation benchmarks primarily provide coordinate supervision for recognition or retrieval, without explicitly

targeting multi-step reasoning or tool use [22, 47, 54]. Cross-view datasets focus on matching street-level images to aerial views within constrained regions [58], while more recent large-scale street-view corpora expand geographic coverage but still emphasize single-shot localization from images [2, 8]. Recent efforts begin to incorporate human signals or limited tool interaction, such as aggregating human gameplay data [43] or enabling zoom and web search for agentic models [48]. However, existing benchmarks rarely combine globally localizable imagery with high-difficulty, information-seeking queries and expert, stepwise multimodal tool-use annotations. This leaves limited support for evaluating agentic frameworks that require both think-with-image reasoning and web-based verification under realistic constraints.

3 GeoBrowse Benchmark

GeoBrowse is a geolocation benchmark designed to evaluate three agentic capabilities: *visual cue composition*, *open-web information seeking and verification*, and *tool-use planning*. It comprises 300 instances, with 199 in Level 1 and 101 in Level 2. As shown in Fig. 3, GeoBrowse spans broad geographic coverage across continents: Asia (47%), Africa (5%), Europe (22%), North America (13%), Oceania (4%), and South America (9%). Tab. 2 summarizes answer granularity by administrative level. Level 2 skews toward country-level targets due to multi-hop knowledge reasoning.

Level 1 targets image-centric reasoning from ambiguous cues. The number of visual cues is mostly around 2 to 4 with a mean of 3.21, whose distribution is shown in Fig. 2. Example instances and expert-annotated traces are provided in Appendix E.1. Level 2 extends Level 1 by adding multi-hop reasoning based on open-web information and cultural knowledge. Solving them perfectly requires 8.42 reasoning hops on average. Agents must first infer the location from the images, then use it as the start for iterative information seeking and verification chain. Each instance includes all Level 1 fields and an annotated evidence trace of solving multi-hop query shown in Appendix E.2.

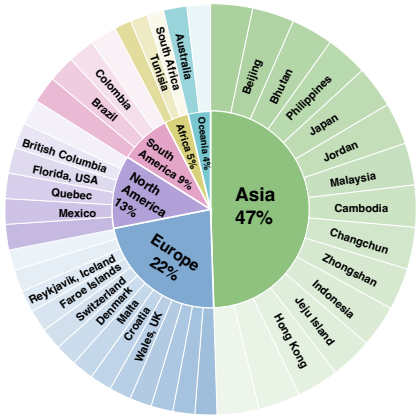

Fig. 3: Geographic coverage of GeoBrowse visual cues. The inner ring shows the percentage of instances by continent and the outer ring lists representative locations within to illustrate the diversity of covered places.

3.1 Data Collection

Level 1 We source raw candidates from publicly available geolocation videos created by community experts on platforms such as YouTube. These videos

include stepwise reasoning, providing a natural source of geolocation inference traces. We only use videos released under permissive licenses CC BY 4.0 and retain source attribution. From these videos, we sample representative frames for annotation. All instances are annotated by five experts with experience in geography research or geolocation. Annotators refine the candidate set by (i) identifying informative visual cues, (ii) collecting expert-level stepwise reasoning traces, and (iii) specifying appropriate tool-use strategies. These annotated candidates then undergo subsequent filtering and quality control in Sec. 3.2. To mitigate privacy concerns, all instances are filtered to avoid exposing personally identifiable information, which is detailed in Appendix A.

By sourcing candidates from geolocation videos and annotating them with traces from experts, our benchmark provides content-grounded, multi-hop reasoning chains that are largely missing from prior geolocation datasets, which typically offer only coordinates or final-answer supervision.

Level 2. Level 2 is constructed by transforming a Level 1 geolocation query $q^{(1)}$ into a BrowseComp-style, knowledge-intensive query $q^{(2)}$. Following [13, 23], let $a^{(1)}$ denote the gold answer of the Level 1 instance. We treat $a^{(1)}$ as the root entity and build a multi-hop chain on the Wikipedia hyperlink graph $G = (\mathcal{E}, \mathcal{R})$, where $\mathcal{E}$ is the set of Wikipedia entities (title of pages) and $(e, e') \in \mathcal{R}$ indicates a directed hyperlink from e to e'. For clarity, define the outgoing-neighbor operator

$$\mathrm{Out}(e) = \{e' \in \mathcal{E} \mid (e, e') \in \mathcal{R}\}. \tag{1}$$

Starting from $e_0 = a^{(1)}$, we sample a path π

$$\begin{aligned} &\pi = (e_0, e_1, \ldots, e_K), \\ \text{s.t.} \quad &e_{i+1} \in \mathrm{Out}(e_i), \\ &\forall i \in \{0, \ldots, K-1\}. \end{aligned} \tag{2}$$

where K controls the reasoning depth. We set the terminal entity as the Level 2 answer, $a^{(2)} = e_K$. Importantly, the hops in π are designed to be dependency-critical. Each step provides the minimal evidence needed to identify the next, so omitting or skipping a hop typically breaks downstream retrieval and verification, yielding an almost canonical solution path for following annotating process.

Given π, we generate the query $q^{(2)}$ through Gemini-3-Pro [10] under a strict constraint: all intermediate entities $\{e_i\}_{i=1}^{K-1}$ must be obfuscated with indirect descriptions and contextual clues rather than literal entity names, ensuring that the query cannot be solved via direct search without reasoning. The prompt of query generation is in Appendix D.1.

This construction requires an agent to first infer $a^{(1)}$ from the image to anchor the chain, and then perform multi-hop retrieval and verification to reach $a^{(2)}$. As a result, Level 2 jointly evaluates visual geolocation, knowledge-intensive reasoning, and tool-based planning.

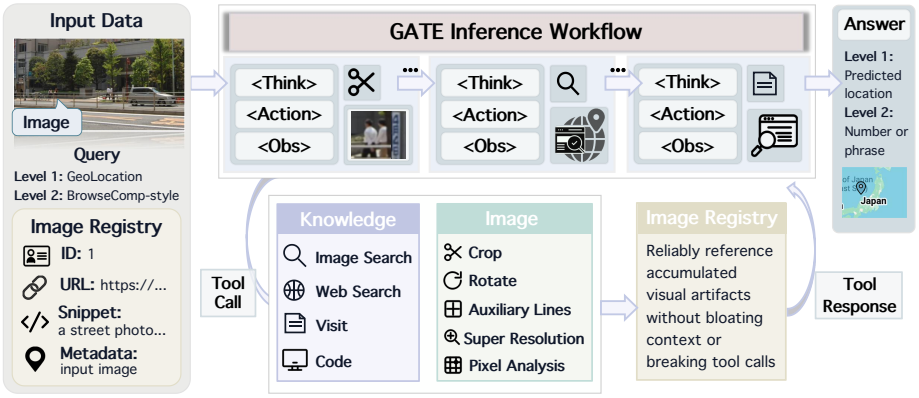

Fig. 4: The pipeline of GATE, our proposed Geolocation Agentic-workflow with Tool Enhancement approach. The input image is first registered into stable `img_id` references. GATE then follows a ReAct-style loop: `<Think>` summarizes the latest evidence and plans the next step, `<Action>` invokes an image or knowledge tool, and the tool response is returned as `<Obs>` to update the agent state. Any new images in `<Obs>` are re-registered, and the loop repeats until the final answer.

3.2 Quality Control and Reliability

In the beginning, we use LLM with tool access to judge quality of queries. Each instance is evaluated by OpenAI o4-mini [38] four times. We drop items solved in all runs as too easy and manually review items failed in all runs, discarding those with insufficient or incoherent evidence. This process filters both trivial and incomplete cases and improves the reliability of the benchmark. Importantly, o4-mini is not used in any subsequent evaluation, thus avoiding model leakage.

Then, all instances are validated by three experts with experience in geography-related research and geolocation problem solving. They deem an instance reliable only if the final answer is correct and supported by the verifiable evidence trace, which is internally consistent and reproducible by our tools. For Level 2, annotators additionally check that the query preserves ambiguity by avoiding explicit named-entity leakage while still admitting a valid multi-hop reasoning path. Detailed criteria and the validation protocol are provided in Appendix B.4.

Each instance is annotated by one expert and reviewed by a second expert. Disagreements are resolved by a third expert based on the completeness and verifiability of evidence. As a result, every instance receives at least two rounds of expert review. We report additional annotation details and agreement statistics in Appendix B.4.

4 Geolocation Agentic-workflow with Tool Enhancement

To address the limitations of pure LLM in grounded reasoning, such as missing fine-grained visual cues and unreliable factual verification, we augment the model

with a structured tool suite and introduce our **G**eolocation **A**gentic-workflow with **T**ool **E**nhancement (**GATE**). These tools provide external perceptual and knowledge support, enabling more reliable multi-modal reasoning.

4.1 Tools

As shown in Fig. 4, we group our tool suite into two categories: *think-with-image* tools for targeted visual cue extraction, and *knowledge* tools for open-web retrieval and verification. The formal tool definitions, interfaces, and implementation details are provided in Appendix D.5.

Think-with-image tools. These tools operate directly on the input image to expose weak and localized clues that are hard to access with a single forward pass. We support five primitive tools:

1) **Crop** for isolating salient subregions and reducing background clutter,
2) **Rotate** for correcting camera orientation and aligning text or structural cues,
3) **Auxiliary Lines** for overlaying reference guides such as vanishing lines or heading aids,
4) **Local Super-Resolution** for selective enhancement of small regions,
5) **Pixel Analysis** to figure out color statistics, edge density and texture cues.

Think-with-image tools are implemented in a Python sandbox [57]. When image manipulation is needed, the agent writes executable code and invokes the corresponding tool. The sandbox returns either an edited image or a brief analyzing report with statistics. For reliability and flexibility, the system prompt includes runnable reference snippets for each primitive tool. The sandbox also ships with standard libraries such as Pillow [7] and OpenCV [4].

Knowledge tools. Knowledge tools enable evidence gathering and disambiguation, which is particularly important for BrowseComp-style Level 2 instances where multiple weak cues must be validated against external sources. We include:

1) **Web Image Search** via Google SerpApi [15] for retrieving relevant images with captions and URLs,
2) **Web Text Search** for open-domain information seeking,
3) **Visit** via Jina [20] for navigating specific URLs and goal-conditioned page reading,
4) **Code Interpreter**, following the implementation in [6], for symbolic computation and numerical reasoning.

4.2 GATE the Agentic Workflow

Action space and trajectories. We model query answering as an interactive ReAct-style [52] trajectory τ. As shown in Fig. 4, at each step, the agent

conditions on the accumulated history and produces a structured turn consisting of (1) a **Thought**, summarizing experience in the last step and choosing which tool to call, wrapped in `<think>...</think>`; (2) an **Action**, either a tool invocation in `<tool_call>...</tool_call>` or a terminal response in `<answer>...</answer>`; and (3) an **Observation**, i.e., the return of the environment, wrapped in `<tool_response>...</tool_response>`.

Let $\mathcal{A}$ denote the discrete tool-action set, and let **Finish** terminate the episode by emitting the final answer. The trajectory τ of length L is

$$\tau = \{(a_0, o_0), (a_1, o_1), \ldots, (a_L, o_L)\}, \tag{3}$$

where $a_i \in \mathcal{A} \cup \{\textbf{Finish}\}$ and o_i is the observation returned after executing a_i.

In-trajectory image registry. GATE needs to preserve images produced in a trajectory, since both image tools and web image search can generate new visual artifacts. However, directly appending image paths, URLs or embeddings to trajectory is impractical under bounded context window and I/O budget, and long histories can also increase cost and reduce reliability. More importantly, once images accumulate, the agent must repeatedly restate long pointers to select a specific image, where small formatting errors can make the image inaccessible and break downstream tool calls. We maintain an in-trajectory image registry as a persistent visual state.

As shown in Tab. 3, each registry stores an `img_id`, a resolvable pointer (*i.e.*, URLs or Paths), a short snippet, and provenance metadata including producing tool and parent image. When a tool produces a new image, the environment registers it and exposes the updated registry to agent at the next step. The agent references images by `img_id`, and the environment resolves an identifier to underlying pointer for subsequent inspection or tool calls. For example, Tab. 3 shows a simple lineage where `img_id=0` is cropped into `img_id=1` and then super-resolved into `img_id=2`, with `from` explicitly tracking provenance. This avoids copying long paths, reduces hallucinated references, and improves the robustness of multimodal tool use.

Table 3: Example of image registration table.

Field	Value
img_id	0
pointer	`input.png`
snippet	input image
metadata	input
img_id	1
pointer	`crop_1.png`
snippet	sign region crop
metadata	`Crop(img_id=0)`
img_id	2
pointer	`sr_1.png`
snippet	2× SR on sign
metadata	`SuperRes(img_id=1)`

5 Experiments

5.1 Experimental Setup

We compare three classes of baselines on GeoBrowse:
(1) Direct Inference. Models give answers in a single pass without tool calls. We evaluate GPT-4o [33], GPT-4.1 [37], GPT-5 [35], GPT-5.2 [36], Gemini-2.5-Pro/Flash [9], Gemini-3-Pro, Claude-4.5-Opus [1], Llama-3.2-90B [30], Qwen-2.5-VL family (7B/32B) [3] and Qwen-3-VL-32B [40].

Table 4: Main results on GeoBrowse. All accuracy scores are reported as percentages. Each entry is averaged over 3 runs. *Avg.* is the macro-average over city/state/country.

Backbone	Level 1: Visual Cues				Level 2: BrowseComp-style			
	City	State	Country	Avg.	City	State	Country	Avg.
Direct Inference								
GPT-4o	5.8	16.4	35.7	23.1	10.5	15.8	11.1	11.9
GPT-4.1	5.8	14.5	30.4	19.6	5.3	5.3	15.9	11.9
GPT-5	11.5	21.8	38.0	27.6	15.8	10.8	20.6	17.8
GPT-5.2	7.7	9.1	30.4	18.6	21.1	21.1	20.6	20.8
Gemini-2.5-Pro	11.5	21.8	41.3	28.1	10.8	10.8	17.5	14.8
Gemini-2.5-Flash	7.7	18.2	40.2	22.6	5.3	10.8	9.9	12.9
Gemini-3-Pro	13.5	27.3	53.3	35.7	21.1	21.1	23.8	22.8
Claude-4.5-Opus	13.5	23.6	43.4	30.2	21.1	26.3	27.0	25.7
Llama-3.2-90B	3.8	10.9	19.6	13.1	5.3	5.3	9.5	7.9
Qwen-3-VL-32B	7.7	16.4	16.3	14.1	0.0	5.3	9.5	6.9
Qwen-2.5-VL-7B	1.9	6.3	11.1	7.5	0.0	0.0	3.2	1.9
Qwen-2.5-VL-32B	3.8	9.1	17.3	11.1	0.0	0.0	6.3	4.0
Open Source Agents								
OmniSearch	9.6	18.2	37.0	24.6	15.8	21.1	19.0	18.8
WebWatcher	7.7	18.2	38.0	24.6	15.8	26.3	20.6	20.8
Pyvision	15.4	20.0	43.5	29.6	10.5	15.8	14.3	13.9
GATE								
GPT-4o	7.7	20.0	52.2	31.8	15.8	21.1	22.2	21.8
GPT-4.1	7.7	20.0	45.7	28.6	15.8	21.1	23.8	22.8
GPT-5	11.5	23.6	65.2	39.7	21.1	31.6	31.8	30.7
GPT-5.2	11.5	23.6	54.3	34.7	26.3	31.6	33.3	31.7
Gemini-2.5-Pro	13.5	23.6	52.2	35.7	15.8	36.8	30.2	29.7
Gemini-2.5-Flash	9.6	20.0	50.0	30.7	10.5	21.1	23.8	20.8
Gemini-3-Pro	24.0	42.6	65.2	**48.2**	26.3	36.8	34.9	**34.7**
Claude-4.5-Opus	15.4	32.7	63.0	42.2	21.1	36.8	33.3	32.7
Llama-3.2-90B	3.8	10.9	19.6	13.1	5.3	10.5	12.7	11.9
Qwen-3-VL-32B	7.7	12.7	30.4	19.6	5.3	5.3	9.5	8.9
Qwen-2.5-VL-7B	3.8	9.1	13.0	9.5	0.0	5.3	6.4	4.9
Qwen-2.5-VL-32B	5.8	10.9	21.7	14.1	0.0	5.3	7.9	5.9

(2) Open-source Agents. We include three representative open-source agents that span complementary tool capabilities. OmniSearch [27] is search-centric. WebWatcher [13] supports end-to-end web search, browsing, and lightweight computation for multi-step verification. PyVision [57] provides a code-driven vision sandbox for flexible image manipulation. Together, they cover search-only, browse-and-verify, and image-tool use baselines. For fair comparison and consistency with common practice in these work, we use the default tool interfaces and GPT-4o as the backbone.

(3) GATE. We evaluate GATE across backbones with standardized tool use via a fixed protocol, tool suite, and call budget. All models use the same interfaces, return formats, and image registry. The complete system prompt is in Appendix D.3.

Metrics and Judging. We report accuracy as the primary metric. For a test set of n instances, we compute

$$\text{pass@1} = \frac{1}{n}\sum_{i=1}^{n} p_i, \tag{4}$$

where $p_i \in \{0, 1\}$ indicates whether the i-th prediction is correct. Because both gold answers and model outputs are short and concrete, the correctness can be determined using *LLM-as-judge* [28], whose prompt follows Humanity's Last Exam [39] in Appendix D.4. To ensure reliability, we compare human decisions against the LLM-as-judge and report agreement statistics in Appendix B.3.

5.2 Main Results

Tab. 4 shows that **GATE** consistently improves over direct backbone inference and also surpasses representative open-source agents. While these agents benefit from external tools, their gains are more limited as they cover a narrower subset of tool functionalities. On the same backbone, GATE improves GPT-4o from 23.1% to 31.8% on Level 1 and from 11.9% to 21.8% on Level 2, exceeding search-centric OmniSearch at 24.6% and 18.8% and vision-centric PyVision at 29.6% and 13.9%. This suggests search-only or image-only tools are insufficient, especially on Level 2 where multi-step retrieval and verification are required.

Breaking results down by target granularity, fine-grained localization is substantially harder: city-level accuracy of Level 1 is generally lower than state- and country-level accuracy. Besides, Level 2 city and state splits each contain 19 questions, thus a 5.2% change corresponds to only one additional solved instance. This indicates that for harder fine-grained locations, GATE yields modest absolute improvements for most models on Level 2. This pattern is consistent with our Level 2 design: each instance uses geolocation as the root, and failure of localization prevents the agent from reliably initiating the subsequent multi-hop reasoning chain. The obfuscated, BrowseComp-style clues make it difficult to back-infer the correct location.

Overall, GATE with the **Gemini-3-Pro** backbone achieves the best performance, reaching 48.2% on Level 1 and 34.7% on Level 2, followed by Claude-4.5-Opus and GPT-5. GATE provides large absolute gains for strong models, improving Gemini-3-Pro from 35.7% to 48.2% on Level 1 and from 22.8% to 34.7% on Level 2. Moreover, gains are larger on Level 1, consistent with the importance of image-side processing, while Level 2 improvements are smaller and more model-dependent, reflecting the added difficulty of long-tail knowledge retrieval and evidence integration. We therefore analyze tool traces and single-tool ablations to localize where the improvements originate.

5.3 Analysis

Tool Usage Profile Fig. 5 characterizes how agents allocate tool calls. Percentages denote the proportion of total tool invocations within each level. Level 1

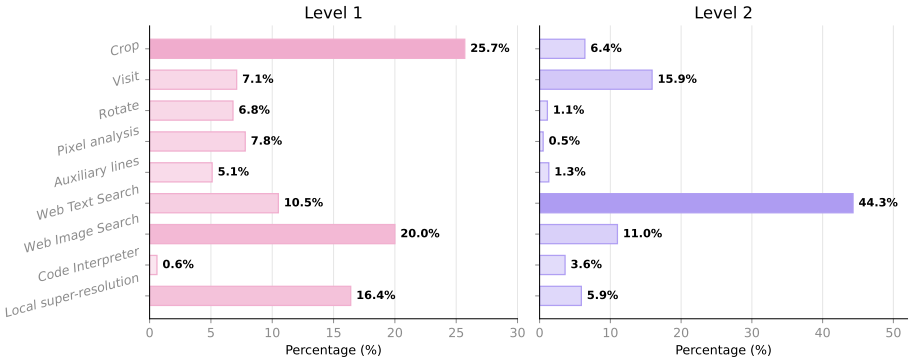

Fig. 5: Tool-use distribution on GeoBrowse. Statistics are aggregated over all tool calls produced by GATE with the **Gemini-3-Pro** backbone, across Level 1 (geolocation tasks) and Level 2 (multi-step reasoning tasks requiring external knowledge).

concentrates on visual preprocessing and image-based retrieval, with *Crop* accounting for 25.7%, *Web Image Search* 20.0%, and *Local Super-resolution* 16.4%. Level 2 shifts toward web evidence gathering, where *Web Text Search* accounts for 44.3% of all tool calls and *Visit* increases to 15.9%.

Call frequency, however, does not directly measure utility. We therefore run single-tool ablations and quantify coordination with the **synergy gap**, defined as the all-tools accuracy minus the best single-tool accuracy. Since *Visit* depends on retrieved URLs, we pair it with the corresponding search tool. As shown in Tab. 5, on Level 1, *Image Processor* reaches 38.6%, only 1.1% below the all-tools score 39.7%, indicating strong visual clue extraction captures most of the attainable gain. On Level 2, the best single-tool setting is *Web Text Search + Visit* at 27.0%. Level 2 leaves a larger 3.7% gap to the all-tools score 30.7%, suggesting a stronger dependence on tool coordination. This pattern matches GeoBrowse's design, where Level 1 emphasizes image-side cue extraction and Level 2 adds knowledge-intensive multi-hop retrieval and verification. Finally, *Code* alone underperforms Level 1 baseline, indicating that computation without external evidence is not effective for this task.

Table 5: Single-tool ablation results. Accuracy (%) generated by GPT-5. Purple marks the best single-tool setting per level, and Pink indicates degradation.

Single-tool setting	Level 1 (%)	Level 2 (%)
Baseline	26.6	17.8
All tools enabled	39.7	30.7
Web Text Search + Visit	28.6	27.0
Web Image Search + Visit	35.7	22.9
Image Processor	38.6	21.2
Code Interpreter	23.1	18.6
Synergy gap	1.1	3.7

Fixed Policy vs. Agentic Planning We compare the agentic planning with a fixed tool-use policy, which is constructed by fixing a representative tool chain observed in trajectories, using the same tool suite but removing all adaptive decisions and retries. Details are in Appendix B.1. Tab. 6 leads to three findings:

(1) Fixed policy captures some of the tool gain, but remains below agentic planning. Fixed tool use cannot recover from early errors and is brittle to top-1 retrieval noise and ambiguous entities, lacking query refinement and tool reordering that would waste budget on plausible but irrelevant candidates.

(2) The gap between fixed and agentic widens on Level 2. For example, agentic planning of GPT-4o adds 6.2% accuracy over the fixed policy on Level 2, compared to 3.4% on Level 1. Level 2 requires more multi-hop retrieval and explicit verification, where adaptive planning and evidence switching provide larger returns.

(3) Benefits depend on model strength, and weak models may degrade. While strong models consistently benefit from fixed and agentic tool use, Qwen-2.5-VL-7B drops under both regimes on Level 1 (7.5% to 6.1% with fixed; 7.5% to 6.5% with agentic), suggesting strong models filter noisy tool outputs and integrate evidence across steps, while weaker models can be misled by noisy observation.

Table 6: Accuracy comparison of three tool regimes. *B*: Baseline without tools. *F*: Fixed tool-use policy. *A*: Agentic tool-use. Purple highlights strong performance, and Pink indicates performance degradation.

	Level 1 (%)			Level 2 (%)		
Model	B	F	A	B	F	A
GPT-4o	23.1	28.4	31.8	11.9	15.6	21.8
GPT-5	26.6	34.9	39.7	17.8	23.8	30.7
Gemini-3-Pro	28.1	40.8	48.2	22.8	29.2	34.7
Claude-4.5-Opus	30.2	37.6	42.2	25.7	28.0	32.7
Qwen-3-VL-32B	14.1	16.8	19.6	6.9	6.9	8.9
Qwen-2.5-VL-7B	7.5	6.1	6.5	3.0	2.7	4.9

Trajectory-level Analysis To test whether these gains reflect better planning rather than trial-and-error, we analyze agentic trajectories on Level 2 using human-annotated milestones. We extract milestones from the expert traces, each specifying a key visual cue or a retrievable entity, and count a milestone hit if it appears in tool responses. Each instance has 10 milestones on average. Tab. 7 reveals two findings:

(1) Incorrect runs call more tools but do not convert them into evidence. For GPT-5 and Gemini-3-Pro, tool calls increase from 12.2 to 15.7 and from 12.8 to 16.1 in incorrect cases, but milestone hit rates remain far lower than in correct cases, indicating inefficient recovery after early noise. Weaker models show smaller call increases with limited milestone coverage. Overall, the gains from agentic tool use are better explained by more reliable milestone coverage and effective recovery.

Table 7: Trajectory-level analysis on Level 2 of GATE. *Acc.* is the accuracy. *Split* denotes correct (T) and incorrect (F) instances. *MS Hit* is the milestone hit rate.

Model	Acc. (%)	Split	MS Hit (%)	Tool Calls
GPT-5	30.7	T	78.0	12.2
		F	57.0	15.7
Gemini-3-Pro	34.7	T	80.0	12.8
		F	59.0	16.1
Qwen-3-VL-32B	8.9	T	61.0	10.7
		F	29.0	12.6
Qwen-2.5-VL-7B	3.0	T	42.0	8.6
		F	15.0	9.6

(2) Strong and weak models fail at different stages. For GPT-5 and Gemini-3-Pro, correct runs reach 78–80% milestone hit rate, while incorrect runs still

retain 57–59%, indicating failures after substantial evidence is retrieved, typically during synthesis, candidate selection, or verification. For weaker models, milestone coverage drops sharply when incorrect (29% for Qwen-3-VL-32B and 15% for Qwen-2.5-VL-7B), suggesting earlier breakdowns in extracting or grounding retrievable entities.

Failure Diagnosis To diagnose failures and inform system improvements, we label incorrect Level 2 trajectories with six coarse error types, whose detailed definitions are in Appendix B.2.

E1 Perception and grounding failure, E2 Retrieval strategy and querying failure, E3 Noisy or ambiguous evidence selection, E4 Missing or insufficient verification, E5 Ordering and budgeting failure, and E6 Synthesis and final decision failure.

Table 8: Error-type distribution (%) on Level 2 under GATE, computed over incorrect cases only. For each model, we highlight the most frequent error.

E	GPT-5	Gemini-3-Pro	Qwen3-VL-32B	Qwen2.5-VL-7B
E1	7.0	6.1	28.4	41.4
E2	16.9	15.0	30.0	32.5
E3	16.2	16.8	14.3	9.7
E4	13.3	14.9	7.9	5.2
E5	8.4	7.5	10.3	7.6
E6	38.2	39.7	9.1	3.6

Tab. 8 shows clear stage differences by model strength. For GPT-5 and Gemini-3-Pro, errors are dominated by **E6** at 38.2–39.7%, indicating that failures often arise during evidence synthesis or final decision-making rather than retrieval, aligning with Finding 1 in Sec. 5.3. For weaker models, most errors fall into **E1**+**E2** at 58.4% for Qwen-3-VL-32B and 73.9% for Qwen-2.5-VL-7B, reflecting difficulty in extracting retrievable entities and forming effective queries.

Text-only Probing for Visual Grounding. To investigate whether Level 2 questions can be answered without the image, we evaluated all Level 2 instances under two text-only protocols: *Text-only Query*, which removes the image and requires the model to answer the Level 2 question from text alone, and *Text-only Root*, which tests whether the geographic anchor can be inferred solely from the text. Tab. 9 shows that shortcuts exist but are not universal:

visual information remains important, although strong models can sometimes compensate for visual reasoning through textual back-inference. These results further demonstrate that GeoBrowse can diagnose multimodal shortcut behavior by checking whether agents ground the visual root or substitute it with downstream textual priors, and they motivate a deeper analysis of why stronger models may bypass visual grounding when text-based reasoning is available.

Table 9: Text-only audit of Level 2 instances, comparing original direct answering with image against two text-only protocols.

Model	Original	Text-only	
	Direct	Query	Root
GPT-5.4	30.4	14.8	20.5
GPT-4.1	11.9	1.9	4.0
Gemini-3-Pro	22.8	9.0	15.8
Claude-4.5-Opus	25.7	11.9	15.8
Qwen-3-VL-32B	6.9	0.0	0.0
Qwen-2.5-VL-7B	1.9	0.0	0.0

Table 10: Accuracy with 95% confidence intervals for representative models on Level 1, Level 2, and diagnostic Level 2 geographic subgroups.

Setting	Model	L1	L2	City	State	Country
Direct	GPT-5	19.6 [17.1, 21.1]	11.9 [8.9, 12.9]	15.8 [5.3, 21.1]	10.5 [0, 15.8]	20.6 [15.8, 23.8]
Direct	Gemini-3-Pro	35.7 [32.2, 37.2]	22.8 [17.8, 25.7]	21.1 [15.7, 26.3]	21.1 [10.5, 26.3]	23.8 [20.6, 28.9]
Direct	Claude-4.5-Opus	30.2 [28.6, 32.7]	25.7 [22.8, 26.7]	21.1 [15.7, 26.3]	26.3 [15.7, 31.6]	27.0 [23.8, 31.7]
GATE	GPT-5	28.6 [25.1, 35.2]	22.8 [15.8, 30.7]	21.1 [15.8, 36.8]	31.6 [21.1, 42.1]	31.8 [25.4, 39.7]
GATE	Gemini-3-Pro	48.2 [45.2, 52.8]	34.7 [27.7, 43.6]	26.3 [15.8, 42.1]	36.8 [26.3, 47.3]	34.9 [27.0, 42.9]
GATE	Claude-4.5-Opus	42.2 [39.6, 45.7]	32.7 [29.7, 42.3]	21.1 [15.7, 26.3]	36.8 [26.3, 47.3]	33.3 [27.0, 45.6]

Confidence Intervals We analyze 95% confidence intervals for representative models in Tab. 10. The aggregate Level 1/Level 2 intervals show clear separation or only minimal overlap for the main trends, while the small city/state slices have wide intervals as expected and are used for diagnostic analysis rather than strong subgroup-level claims. We will continue expanding GeoBrowse to improve the robustness of subgroup evaluation.

6 Conclusion

We introduced GeoBrowse, a tool-augmented geolocation benchmark designed for multi-step reasoning with fragmented visual evidence and open-web verification. GeoBrowse includes two levels: Level 1 focuses on visual-first cue composition, and Level 2 adds BrowseComp-style multi-hop knowledge reasoning. Each instance is paired with expert stepwise annotations. We also introduce an agentic workflow GATE equipped with a unified tool suite. Our experiments show improvements come from coherent, level-specific tool-use plans rather than increased tool usage. Level 1 primarily benefits from visual processing, whereas Level 2 relies more on evidence-driven retrieval. Agentic planning further improves accuracy by hitting the annotated milestones and reducing errors in final integration.

As an expert-curated benchmark, GeoBrowse prioritizes trace quality and verifiable evidence, which currently bounds its scale and introduces some distributional skew, such as a predominance of country-level instances and uneven geographic coverage. Beyond dataset expansion, future work will focus on core agentic reasoning improvements, including adaptive tool-use planning, milestone-aware hypothesis tracking, and stronger evidence aggregation across visual cues and open-web sources. These directions aim to move agents from reactive tool use toward iterative formation, verification, and revision of geographically grounded hypotheses.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2025YFE0200500) and the HKUST-Tsinghua Joint Research Initiative Fund (Grant No. TSINGHUA26EG05).

References

1. Anthropic: Claude opus 4.5 (2025), https://www.anthropic.com/claude/opus
2. Astruc, G., Dufour, N., Siglidis, I., Aronssohn, C., Bouia, N., Fu, S., Loiseau, R., Nguyen, V.N., Raude, C., Vincent, E., et al.: Openstreetview-5m: The many roads to global visual geolocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21967–21977 (2024)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bradski, G.: The opencv library. Dr. Dobb's Journal: Software Tools for the Professional Programmer **25**(11), 120–123 (2000)
5. Chen, Z., Wang, X., Jiang, Y., Zhang, Z., Geng, X., Xie, P., Huang, F., Tu, K.: Detecting knowledge boundary of vision large language models by sampling-based inference. arXiv preprint arXiv:2502.18023 (2025)
6. Cheng, Y., Chen, J., Chen, J., Chen, L., Chen, L., Chen, W., Chen, Z., Geng, S., Li, A., Li, B., et al.: Fullstack bench: Evaluating llms as full stack coders. arXiv preprint arXiv:2412.00535 (2024)
7. Clark, A., et al.: Pillow (pil fork) documentation. readthedocs (2015)
8. Clark, B., Kerrigan, A., Kulkarni, P.P., Cepeda, V.V., Shah, M.: Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23182–23190 (2023)
9. DeepMind, G.: Gemini 2.5 (2025), https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/
10. DeepMind, G.: A new era of intelligence with gemini 3 (2025), https://blog.google/products/gemini/gemini-3/
11. Dong, Y., Liu, Z., Sun, H.L., Yang, J., Hu, W., Rao, Y., Liu, Z.: Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9062–9072 (2025)
12. Du, M., Xu, B., Zhu, C., Wang, X., Mao, Z.: Deepresearch bench: A comprehensive benchmark for deep research agents. arXiv preprint (2025)
13. Geng, X., Xia, P., Zhang, Z., Wang, X., Wang, Q., Ding, R., Wang, C., Wu, J., Zhao, Y., Li, K., et al.: Webwatcher: Breaking new frontier of vision-language deep research agent. arXiv preprint arXiv:2508.05748 (2025)
14. Google: Try deep research and our new experimental model in gemini, your ai assistant (2024), https://blog.google/products/gemini/google-gemini-deep-research/
15. Google: Serpapi (2025), https://serpapi.com/
16. Gu, J., Xian, Z., Xie, Y., Liu, Y., Liu, E., Zhong, R., Gao, M., Tan, Y., Hu, B., Li, Z.: Toward structured knowledge reasoning: Contrastive retrieval-augmented generation on experience. arXiv preprint arXiv:2506.00842 (2025)
17. Guo, X., Tyagi, U., Gosai, A., Vergara, P., Park, J., Montoya, E.G.H., Zhang, C.B.C., Hu, B., He, Y., Liu, B., et al.: Beyond seeing: Evaluating multimodal llms on tool-enabled image perception, transformation, and reasoning. arXiv preprint arXiv:2510.12712 (2025)
18. Hu, M., Zhou, Y., Fan, W., Nie, Y., Xia, B., Sun, T., Ye, Z., Jin, Z., Li, Y., Chen, Q., et al.: Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation. arXiv preprint arXiv:2505.23885 (2025)

19. Jiang, G., Su, Z., Qu, X., et al.: Xskill: Continual learning from experience and skills in multimodal agents. arXiv preprint arXiv:2603.12056 (2026)
20. Jina.ai: Jina (2025), https://jina.ai/
21. Koh, J.Y., Lo, R., Jang, L., Duvvur, V., Lim, M., Huang, P.Y., Neubig, G., Zhou, S., Salakhutdinov, R., Fried, D.: Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 881–905 (2024)
22. Larson, M., Soleymani, M., Gravier, G., Ionescu, B., Jones, G.J.: The benchmarking initiative for multimedia evaluation: Mediaeval 2016. IEEE MultiMedia **24**(1), 93–96 (2017)
23. Li, K., Zhang, Z., Yin, H., Zhang, L., Ou, L., Wu, J., Yin, W., Li, B., Tao, Z., Wang, X., et al.: Websailor: Navigating super-human reasoning for web agent. arXiv preprint arXiv:2507.02592 (2025)
24. Li, M., Zhong, J., Zhao, S., Zhang, H., Lin, S., Lai, Y., Wei, C., Psounis, K., Zhang, K.: Tir-bench: A comprehensive benchmark for agentic thinking-with-images reasoning. arXiv preprint arXiv:2511.01833 (2025)
25. Li, S., Bu, X., Wang, W., Liu, J., Dong, J., He, H., Lu, H., Zhang, H., Jing, C., Li, Z., Li, C., Tian, J., Zhang, C., Peng, T., He, Y., Gu, J., Zhang, Y., Yang, J., Zhang, G., Huang, W., Zhou, W., Zhang, Z., Ding, R., Wen, S.: Mm-browsecomp: A comprehensive benchmark for multimodal browsing agents (2025), https://arxiv.org/abs/2508.13186
26. Li, X., Jin, J., Dong, G., Qian, H., Zhu, Y., Wu, Y., Wen, J.R., Dou, Z.: Webthinker: Empowering large reasoning models with deep research capability. arXiv preprint arXiv:2504.21776 (2025)
27. Li, Y., Li, Y., Wang, X., Jiang, Y., Zhang, Z., Zheng, X., Wang, H., Zheng, H.T., Yu, P.S., Huang, F., Zhou, J.: Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent (2025), https://arxiv.org/abs/2411.02937
28. Liu, Y., Yang, T., Huang, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., Zhang, Q.: Calibrating llm-based evaluator. In: Calzolari, N., Kan, M., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy. pp. 2638–2656. ELRA and ICCL (2024), https://aclanthology.org/2024.lrec-main.237
29. Luo, W., Lu, T., Zhang, Q., Liu, X., Hu, B., Zhao, Y., Zhao, J., Gao, S., McDaniel, P., Xiang, Z., et al.: Doxing via the lens: Revealing location-related privacy leakage on multi-modal large reasoning models. arXiv preprint arXiv:2504.19373 (2025)
30. Meta: Llama 3.2 (2024), https://huggingface.co/meta-llama/Llama-3.2-90B-Vision
31. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14420–14431 (2024)
32. Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al.: Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332 (2021)
33. OpenAI: Hello GPT-4o (2024), https://openai.com/index/hello-gpt-4o/
34. OpenAI: Deep research system card (2025), https://cdn.openai.com/deep-research-system-card.pdf
35. OpenAI: Gpt-5 large language model (2025), https://openai.com/gpt-5/

36. OpenAI: Introducing gpt-5.2 (2025), https://openai.com/index/introducing-gpt-5-2/
37. OpenAI: Introducing openai gpt-4.1 (2025), https://openai.com/index/gpt-4-1/
38. OpenAI: Openai o3 and o4-mini system card. System card, OpenAI (Apr 2025), https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf
39. Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C.B.C., Shaaban, M., Ling, J., Shi, S., et al.: Humanity's last exam. arXiv preprint arXiv:2501.14249 (2025)
40. Qwen, T.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
41. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems **37**, 8612–8642 (2024)
42. Song, H., Jiang, J., Min, Y., Chen, J., Chen, Z., Zhao, W.X., Fang, L., Wen, J.R.: R1-searcher: Incentivizing the search capability in llms via reinforcement learning. arXiv preprint arXiv:2503.05592 (2025)
43. Song, Z., Yang, J., Huang, Y., Tonglet, J., Zhang, Z., Cheng, T., Fang, M., Gurevych, I., Chen, X.: Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework. arXiv preprint arXiv:2502.13759 (2025)
44. Su, Z., Gao, J., Guo, H., Liu, Z., Zhang, L., Geng, X., Huang, S., Xia, P., Jiang, G., Wang, C., et al.: Agentvista: Evaluating multimodal agents in ultra-challenging realistic visual scenarios. arXiv preprint arXiv:2602.23166 (2026)
45. Su, Z., Xia, P., Guo, H., Liu, Z., Ma, Y., Qu, X., Liu, J., Li, Y., Zeng, K., Yang, Z., Li, L., Cheng, Y., Ji, H., He, J., Fung, Y.R.: Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers (2025), https://arxiv.org/abs/2506.23918
46. Tao, X., Teng, Y., Su, X., Fu, X., Wu, J., Tao, C., Liu, Z., Bai, H., Liu, R., Kong, L.: Mmsearch-plus: A simple yet challenging benchmark for multimodal browsing agents. arXiv preprint arXiv:2508.21475 (2025)
47. Vo, N., Jacobs, N., Hays, J.: Revisiting im2gps in the deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 2621–2630 (2017)
48. Wang, Y., Liu, Z., Wang, Z., Liu, P., Hu, H., Rao, Y.: Geovista: Web-augmented agentic visual reasoning for geolocalization. arXiv preprint arXiv:2511.15705 (2025)
49. Wei, J., Sun, Z., Papay, S., McKinney, S., Han, J., Fulford, I., Chung, H.W., Passos, A.T., Fedus, W., Glaese, A.: Browsecomp: A simple yet challenging benchmark for browsing agents (2025), https://arxiv.org/abs/2504.12516
50. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2575–2584 (2020)
51. Wu, J., Li, B., Fang, R., Yin, W., Zhang, L., Tao, Z., Zhang, D., Xi, Z., Fu, G., Jiang, Y., et al.: Webdancer: Towards autonomous information seeking agency. arXiv preprint arXiv:2505.22648 (2025)
52. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR) (2023)
53. Yuan, H., Sun, Y., Li, Y., Zhang, T., Deng, X., Ding, H., Qi, L., Wang, A., Li, X., Yang, M.H.: Visual reasoning tracer: Object-level grounded reasoning benchmark (2025), https://arxiv.org/abs/2512.05091

54. Zamir, A.R., Shah, M.: Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs. IEEE transactions on pattern analysis and machine intelligence **36**(8), 1546–1558 (2014)
55. Zhang, D., Zhao, Y., Wu, J., Li, B., Yin, W., Zhang, L., Jiang, Y., Li, Y., Tu, K., Xie, P., Huang, F.: Evolvesearch: An iterative self-evolving search agent (2025), https://arxiv.org/abs/2505.22501
56. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
57. Zhao, S., Zhang, H., Lin, S., Li, M., Wu, Q., Zhang, K., Wei, C.: Pyvision: Agentic vision with dynamic tooling. arXiv preprint arXiv:2507.07998 (2025)
58. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021)