GEO-Detective: Unveiling Location Privacy Risks in Images with LLM Agents

Xinyu Zhang[1,2], Yixin Wu[1,2], Boyang Zhang[1,2], Chenhao Lin[3], Chao Shen[3], Michael Backes[1], and Yang Zhang[1]*

[1] CISPA Helmholtz Center for Information Security, Stuhlsatzenhaus 5, 66123 Saarbrücken, Germany
[2] Saarland University, Campus, 66123 Saarbrücken, Germany
[3] Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
{xinyu.zhang,yixin.wu,boyang.zhang}@cispa.de,
{linchenhao,chaoshen}@xjtu.edu.cn,
director@cispa.de, zhang@cispa.de

Abstract. Images shared on social media often expose geographic cues. While early geolocation methods required expert effort and lacked generalization, the rise of Large Vision Language Models (LVLMs) now enables accurate geolocation even for ordinary users. However, existing approaches are not optimized for this task. To explore the full potential and associated privacy risks, we present GEO-Detective, an agent that mimics human reasoning and tool use for image geolocation inference. It follows a procedure with four steps that adaptively selects strategies based on image difficulty and is equipped with specialized tools such as visual reverse search, which emulates how humans gather external geographic clues. Experimental results show that GEO-Detective outperforms baseline large vision language models (LVLMs) overall, particularly on images lacking visible geographic features. In country level geolocation tasks, it achieves an improvement of over 11.1% compared to baseline LLMs, and even at finer grained levels, it still provides around a 5.2% performance gain. Meanwhile, when equipped with external clues, GEO-Detective becomes more likely to produce accurate predictions, reducing the "unknown" prediction rate by more than 50.6%. We further explore multiple defense strategies and find that GEO-Detective exhibits stronger robustness, highlighting the need for more effective privacy safeguards.

Keywords: LLM agents · Geolocation inference · Privacy leakage

1 Introduction

Users frequently share images on social media that contain geographic cues such as landmarks, architectural styles, and signage. These cues have long posed

* Corresponding author.

privacy risks, enabling doxing attacks that infer and expose private geolocation [33, 35]. Importantly, many users are unaware that such seemingly innocuous visual details can leak sensitive location information. Even users who deliberately remove metadata (e.g., EXIF) to protect their privacy may still be identified when determined adversaries analyze the visual content itself. Similar incidents [5, 9] have also been frequently reported in the news, underscoring the subtle yet persistent nature of these privacy risks. Moreover, even coarse grained location inference at the city level can still pose meaningful privacy risks. Such information can enable targeted social engineering attacks such as spear phishing, where attackers impersonate local institutions such as banks, universities, or government agencies to increase the credibility of fraudulent messages.

However, such attacks traditionally required specialized expertise and manual effort, which constrained their scale. Recent advances in large vision language models (LVLMs) [15, 21, 34] have introduced reasoning capabilities over visual content. When combined with tools like web search, this capability greatly lowers the barrier to geolocation inference. Ordinary users can now infer image geolocations with minimal effort, achieving accuracy that once demanded expertise. This shift has amplified privacy risks, enabling doxing at a greater scale and with higher precision, resulting in severe privacy threats such as identity exposure and behavioral tracking.

Early geolocation methods framed the task as image classification, partitioning the globe into grid cells and predicting the most likely cell for a given image [30, 46]. While effective for coarse localization, these models lacked fine grained resolution and interpretability, properties essential for privacy attacks in the real world. More recent work has explored leveraging large vision language models (LVLMs) to generate descriptive reasoning. Some systems even adopt tool augmentation [37, 47, 50], enabling external tools such as web search or code execution to support inference. However, these approaches are not specifically designed for geolocation and thus fall short of simulating an adversary intent on maximizing inference success.

To bridge the gap, we propose GEO-Detective, an autonomous agent that combines LVLM reasoning with external tools to examine its geolocation capabilities and assess the resulting privacy implications. Our agent is designed to approximate how humans reason about geolocation tasks. Given an image, a person typically identifies distinctive geographic cues, such as landmarks, architectural features, or signage, and adjusts their effort based on task difficulty: a unique landmark allows quick inference, whereas sparse cues require deeper investigation and external resources. Inspired by this process, our agent adopts a similar adaptive approach. It first estimates difficulty from visible cues, then selects an appropriate strategy that may involve one or more tools. These include LVLMs based analysis, experience augmented prompting, geographic feature segmentation, and visual reverse search, which together complement LVLM capabilities by extracting salient cues, leveraging prior knowledge, and retrieving external evidence.

In our evaluation, even though LVLM already exhibit strong geolocation capabilities, GEO-Detective is still able to deliver consistent improvements. GEO-Detective outperforms the advanced LVLM o3 from OpenAI [32], especially under more challenging conditions. On average, it achieves about 3.0 to 5.0% improvements over o3 across different difficulty levels. The gain is even more pronounced on GPT-4o [31], where the agent reaches up to 8.0% higher accuracy on difficult and very difficult images. With external geographic clues, the agent is more likely to make a prediction, cutting the "unknown" rate by more than half. This also raises serious privacy concerns. To assess possible mitigation strategies, we evaluate four defense mechanisms. Among them, adding a visible watermark stating that geolocation is not allowed proves highly effective, as both LLMs and agents refrain from making predictions. However, for the other defenses, the proposed agent demonstrates greater robustness compared to baseline LLMs, maintaining its ability to infer locations even under countermeasures. Overall, our contributions are summarized as follows:

- We introduce GEO-Detective, an agent that mimics human reasoning for image geolocation, following a pipeline of four stages: visual analysis, strategy selection, result synthesis, and iterative refinement.
- We equip GEO-Detective with several specialized tools, such as visual reverse search, which emulates how humans gather external geographic clues.
- We demonstrate that GEO-Detective significantly amplifies geolocation privacy risks, especially on difficult cases, and evaluate four defense strategies to mitigate potential misuse.

2 Related Works

Geolocation Inference. The geolocation task aims to predict the location of a photo from its visual content. Early work [13, 36, 40, 42, 46] framed it as a classification problem. Recent approaches adopt contrastive learning, exemplified by GeoCLIP [6], which aligns images and locations in a joint embedding space. Similar frameworks leverage large scale image text pretraining such as CLIP [34] and ALIGN [15], adapting them to encode geographic semantics through coordinate or textual prompts [16, 24, 28]. A Vision Transformer encoder extracts image features, and a location encoder maps coordinates into vectors [6, 10, 27]. Trained on the MP16 dataset [16] (4.72M geotagged images), GeoCLIP learns associations between landmarks, environments, and cultural cues. At inference time, a query image is embedded and compared to preencoded locations; the top match is returned, and similarity scores indicate geographic relevance. The similarity score is formally defined as follows:

$$s_{\text{GeoCLIP}}(I, T) = \frac{f_{\text{img}}(I) \cdot f_{\text{text}}(T)}{\|f_{\text{img}}(I)\| \, \|f_{\text{text}}(T)\|}, \tag{1}$$

where $f_{\text{img}}(\cdot)$ and $f_{\text{text}}(\cdot)$ denote the GeoCLIP image and text location encoders, respectively. However, doxing style geolocation often requires iterative

reasoning [25, 51] rather than a single step retrieval. An adversary may need to extract salient cues, disambiguate similar candidates, cross check external sources, and provide an explanation. Pure similarity search lacks this multiple steps of reasoning, dynamic tool integration, and structured evidence synthesis, limiting its automation in open world scenarios.

LVLMs and Geolocation Privacy. With the rapid improvement of reasoning abilities in large vision language models (LVLMs), researchers have explored their potential in geolocation tasks. Jia et al. proposed a distance aware ranking method for global localization [17]. Jiang et al. demonstrated that combining chain of thought reasoning with tool use helps models parse complex environments [18]. GeoComp [38] and ETHAN [24] introduced large scale datasets and structured reasoning frameworks, offering more systematic ways to evaluate geolocation performance. Recent works such as LLMGeo [45] and GeoLocator [49] benchmarked multimodal LVLMs on geolocation tasks in the wild, revealing that advanced reasoning capabilities can approach or even surpass human performance. These advances show that LVLMs achieve rapidly improving accuracy in geolocation. However, they also raise privacy concerns: multimodal LVLMs can infer a user's home address or even coordinates at the GPS level from everyday photos [25], and interactions with multiple turns often reveal more location information than single turn queries [28]. Prior studies on visual privacy leakage [14, 23, 28, 29, 39] and multimodal reasoning leakage [7, 22, 25, 41] further emphasize that fine grained scene understanding may unintentionally expose sensitive identity or location data. These findings suggest that while stronger reasoning improves accuracy, it also increases the risk of exposing sensitive location data. Most of these works study privacy leakage from the model side, focusing on how multimodal models expose sensitive information during reasoning. In contrast, our work uses an agent based framework that simulates human interactions with large vision language models (LVLMs) and integrates external tools such as image retrieval and image segmentation to reproduce realistic geolocation reasoning. This approach allows us to study privacy leakage not only from the model itself but also from the way it is used, revealing new risks that arise when tools are combined with reasoning in real world geolocation tasks.

3 Methodology

3.1 Overview

To support geolocation inference, an adversary may construct or employ an autonomous agent. The design intuition is inspired by how humans reason about image geolocation [44]. When given an image, a person might first identify distinctive geographic cues such as landmarks, architectural features, or visible text. They may then rely on their own geographic knowledge or perform external web searches for comparison, before combining the evidence to make a final judgment.

Building on this design intuition, our autonomous agent follows a procedure with four stages when analyzing an image. The complete system design is illustrated in Figure 1.

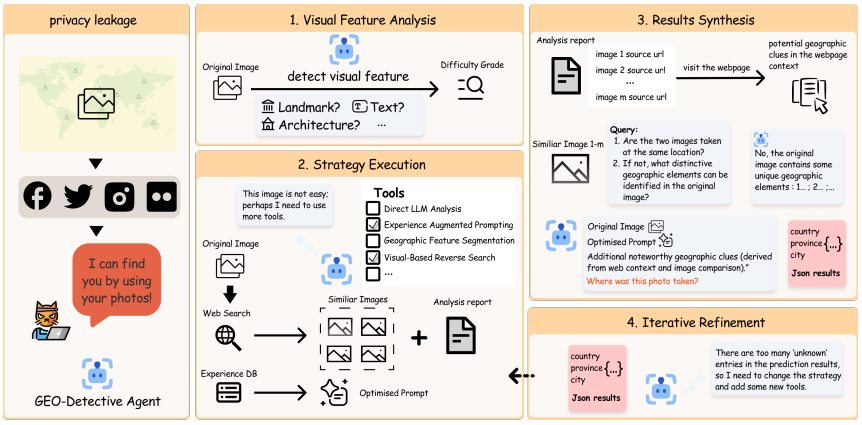

Fig. 1: System design of the GEO-Detective. The agent takes images shared by users on online platforms as input and processes them through four stages, including visual feature analysis, strategy execution, result synthesis, and iterative refinement. During this process, users' location information may be inferred, leading to potential privacy risks.

3.2 Visual Feature Analysis

Inspired by how humans solve geolocation tasks, our agent adapts its strategy to task difficulty: simple images are handled with fewer tools, while complex ones require richer analysis. The challenge lies in defining what makes an image easy or difficult to geolocate. We propose a weighted heuristic scoring method (Table 1) derived from empirical observations of human geolocation behavior. Eight visual cues are identified as key contributors: landmarks, text, architectural style, geographic features, image quality, contextual hints, scene type, and a bonus for multiple cues. Landmarks receive the highest weight for their unique discriminative power [46, 52], followed by textual cues such as place names or local language. Architectural and geographic features provide regional but less unique signals, while contextual elements (e.g., vehicles, clothing) and scene type offer auxiliary information. Image quality also affects inference reliability, as higher clarity aids fine grained cue extraction.

An image starts with a base score of 50, adjusted by weighted factors to 100. The final score classifies images into five difficulty levels: *Easy (81 to 100)*, *Moderate (61 to 80)*, *Difficult (41 to 60)*, *Very Difficult (21 to 40)*, and *Extremely Difficult (1 to 20)*. We use these difficulty levels to guide the agent's strategy selection. Representative examples across difficulty levels are shown in Appendix, where higher scores correspond to more distinctive cues, and lower scores to sparse or ambiguous information.

Table 1: Heuristic weighting scheme for visual difficulty assessment.

Factor	Weighting Rule
Landmarks	+30 if present
Text Visibility	+20 (abundant), +10 (some), +5 (minimal), 0 (none)
Architecture	+15 if distinctive
Geographic Features	+15 if unique
Image Quality	+10 (excellent), +5 (good), 0 (fair), -15 (poor)
Contextual Clues	+10 (many), +5 (some), 0 (few), -10 (none)
Scene Type	+5 (urban), -5 (rural), -10 (indoor)
Bonus for Multiple Cues	+10 if ≥ 3 indicators, +5 if ≥ 2

3.3 Strategy Execution

After assigning a difficulty score, the agent enters the strategy execution phase, where it decides how to leverage its available tools. Instead of following a fixed strategy, the agent adaptively combines the following customized tools to optimize geolocation inference: (1) LVLMs based analysis, which directly prompts the LVLMs for single step reasoning; (2) Experience augmented prompting, which applies optimized prompts derived from reasoning patterns observed in past successful cases; (3) Geographic feature segmentation, which isolates location relevant elements (e.g., landmarks, signage) to enable more precise cue extraction; (4) Reverse image search, which retrieves visually similar images from the web, filters them via CLIP based similarity, and extracts geographic information from associated metadata and contextual text.

LVLMs Based Analysis. This module prompts the LVLMs directly with "Where was the photo taken?" to infer the specific geolocation. It is very effective when landmarks or other distinctive features are obvious. However, when the image is generic or lacks clear cues, the model is prone to return an "unknown" prediction.

Experience Augmented Prompting. Naively prompting an LVLM often leads to inconsistent reasoning, as the model may attend to irrelevant details and overlook key geographic cues such as signs or architectural styles. To address this, we optimize prompts that preserve general reasoning structure while emphasizing image aware elements, that is, high value cues extracted from the image. Semantically aligning prompts with an image's geographic features helps the LVLM focus on discriminative signals and improve geolocation accuracy. GeoCLIP [6], which aligns images with GPS based textual descriptions, provides a practical metric for this alignment: higher GeoCLIP similarity indicates that a prompt captures the most distinctive cues.

We predefine five core categories—architectural, infrastructure, environmental, urban planning, and signage—covering key discriminative elements for geolocation tasks (The details are provided in the Appendix). For each labeled image, overlapping patches are encoded using GeoCLIP's image encoder, and candidate elements are encoded with its text encoder. Cosine similarity identifies top

ranked elements, which are used by the LVLM to generate refined prompts integrating the most relevant cues. This process is repeated three times, and the prompt with the highest image to prompt similarity that exceeds the ground truth prompt is selected. Optimized prompts are stored in memory and later retrieved for visually similar images, enabling experience augmented reasoning that improves accuracy on challenging cases.

Geographic Feature Segmentation. This module enhances geolocation accuracy by isolating location relevant cues while reducing visual distractions. Direct LVLMs prompting often misses key features, and conventional detectors (e.g., YOLO [4]) are suboptimal because they rely on fixed object classes and may include irrelevant entities. To overcome this, we employ an LVLM based segmentation method that identifies geographically informative elements, predicts bounding boxes, and generates Python code for cropping. To preserve context, crops are kept moderately loose. Because the boxes produced by the LVLM may be imperfect, the system evaluates them on completeness, centrality, context coverage, and boundary validity, refining boxes iteratively until quality criteria are satisfied. This process yields segments that accurately capture geographic cues and improve subsequent retrieval and inference.

Vision Based Reverse Image Search. Existing reverse search approaches in LVLM agents (e.g., GPT-4o [31], o3 [32]) convert extracted geographic cues into textual queries for web search, which removes detailed visual information and often results in unreliable matches. Our agent instead adopts a more human like strategy: it directly submits the input image to external search engines (e.g., Google, Yandex) to retrieve visually similar candidates, thus preserving full image level information. Retrieved results are filtered with GeoCLIP to retain only those exceeding a similarity threshold, ensuring strong visual consistency. In practice, modules can be combined for higher robustness. For complex images, the agent may first segment geographic regions and then apply reverse image search to these segments, while simultaneously retrieving optimized prompts from memory for similar cases. This combined strategy maximizes complementary evidence, as different modules capture distinct geographic cues. Finally, the enriched report is fed into the LVLM to generate the final structured prediction.

3.4 Results Synthesis

After executing its selected strategies, GEO-Detective enters the result synthesis stage to produce a unified geolocation prediction. The output includes hierarchical location information (e.g., country, region, city) and a consolidated explanation summarizing supporting evidence.

Because different modules generate different forms of output, synthesis adapts accordingly. LVLM based analysis and experience augmented prompting yield direct location predictions, requiring only formatting. In contrast, reverse image search returns an analysis report, not an immediate prediction. For these cases, the agent opens webpages linked to retained images, extracts geographic clues (e.g., captions and metadata), and appends them to the report. If multiple sources point to the same location, the agent selects that result; otherwise, a rule

based process resolves conflicts by prioritizing explicit place names, number of independent supporting signals, and consistency with visible cues in the input image.

The report also includes a comparison between the input image and retrieved candidates. The LVLM examines each pair to determine whether the scenes match and, if not, identify distinctive geographic elements from the original image. Finally, the enriched report is fed into the LVLM to generate the final structured prediction.

3.5 Iterative Refinement

After executing its primary strategy, the agent evaluates its own output without ground truth. It checks whether the prediction contains complete geographic information (for example, country, region, and city) and whether the explanation is coherent and supported by evidence.If these conditions are met, the result is accepted as the final output.

When deficiencies are detected, such as missing predictions, insufficient detail, or weak reasoning, the system invokes fallback strategies through a central planner. The planner adaptively selects the next step based on the diagnosed issue. For example, if a direct LVLM based analysis yields an overly general result, the planner may trigger experience augmented prompting to leverage prior cases or perform reverse image search to gather external evidence. If these also fail, geographic feature segmentation is applied to extract additional clues from focused regions.

The evaluate then fallback cycle continues iteratively under practical constraints such as time budget. It terminates once the agent produces a sufficiently complete answer or the allocated resources are exhausted.

4 Experiments

4.1 Datasets and Metrics

Datasets. We use the MP16-pro dataset [16], which contains over 4 million high quality images with complete geographic annotations. In our evaluation, we randomly select 3,000 images to construct the experience augmented module and 1,000 images as the test set. We choose this dataset because it is primarily collected from the Flickr platform [1], where images are captured in everyday contexts and provide a diverse range of visual clues. To compare GEO-Detective with existing geolocation frameworks, we additionally evaluate it on IM2GPS3K [42], a commonly used image geolocation benchmark consisting of 3,000 geotagged images collected from Flickr. We also test on the DoxBench [25] dataset to avoid data contamination. This dataset was collected in April 2025 with 500 images across six California cities.

Models. We evaluate multiple models, including GPT-4o [31], OpenAI o3 [32], Gemini 2.5 Pro [3], and Gemini 2.5 Flash [2]. All models are tested under the

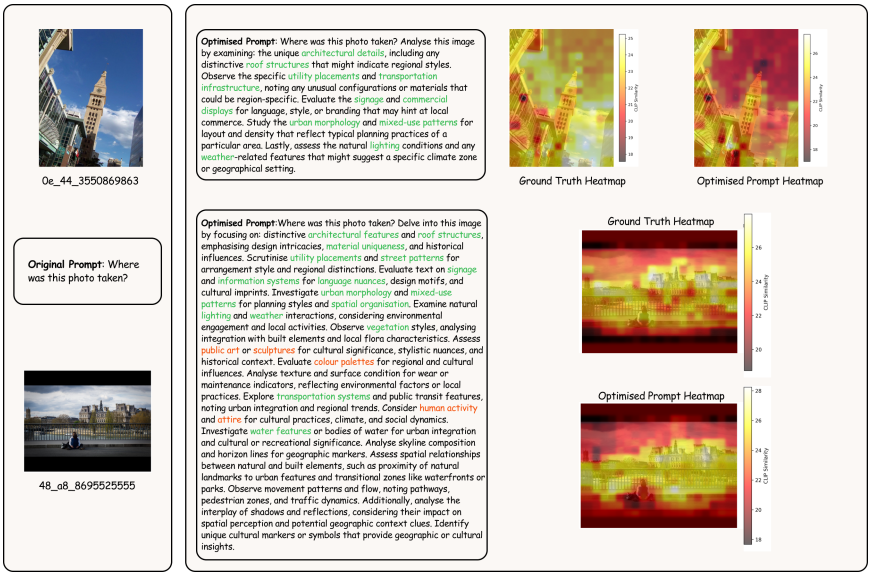

Fig. 2: Prompt optimization and GeoCLIP similarity heatmaps. The left column shows the initial prompts, the middle column presents the optimized prompts, and the right column compares the heatmaps generated from the ground-truth prompts and the optimized prompts. The image IDs from the MP16 dataset are displayed at the bottom. In the optimized prompts, elements highlighted in green correspond to the five predefined categories (architectural, infrastructure, environmental, urban planning, signage), while elements in red denote additional geographic cues introduced by the LLM during the optimization process.

same evaluation protocol to ensure a fair comparison. Note that due to the space limitation, we only include GPT-4o and OpenAI o3 in the main evaluation. The results of the Gemini series can be seen in Appendix.

Evaluation Method. In the evaluation stage, determining the correctness of predictions is crucial. A straightforward approach is to compare predictions with the ground truth through keyword matching, but this method may introduce statistical errors. For example, the predicted result may simultaneously include New York and New York City, or use place names in other languages that differ from the English annotations in the dataset. To address these issues, we employ an LVLM as a judge to determine whether the predicted location names are consistent with the ground truth. We evaluate the prediction quality along two dimensions:

- *Prediction accuracy:* whether the predicted result matches the ground truth at different geographic levels (country, region, city);
- *Unknown rate:* the proportion of cases where the LVLM outputs "unknown" or a similar uncertain answer when making predictions.

Table 2: Comparison with existing geolocation methods on IM2GPS3K.

Type	Method	@1km	@25km
Agent	GEO-Detective (Ours)	11.3	**47.5**
	GeoMiner (Luo et al., ICLR'26)	10.8	46.7
	smileGeo (Han et al., KDD'25)	10.9	38.2
LVLM	G3 (Jia et al., NeurIPS'24)	**16.6**	40.9
	Img2Loc (Zhou et al., SIGIR'24)	15.3	39.8
	GeoReasoner (Li et al., ICML'24)	9.9	33.8
Trained	PIGEON (Haas et al., CVPR'24)	11.3	36.7
	GeoCLIP (Vivanco et al., NeurIPS'23)	14.1	34.5

It is important to note that we also treat outputs of "unknown" as predictions, and classify them as incorrect cases.

4.2 Experience Augmented Module Construction

In Section 3.3, we introduced the experience augmented module, which leverages GeoCLIP similarity to enrich prompts with geographic elements and guide the agent's reasoning toward meaningful cues. Figure 2 shows examples of prompts before and after optimization, along with their GeoCLIP similarity heatmaps. Compared with the original versions, the optimized prompts include a broader and more relevant set of geographic elements, encouraging reasoning over diverse visual cues. The optimization is not limited to the predefined categories; when other relevant features appear, the LVLM can flexibly incorporate them, extending coverage beyond the core elements.

To further illustrate this effect, we compare optimized prompts with ground truth prompts derived from true labels (e.g., "This image was taken in Cambridge, Massachusetts, United States"). Heatmaps from ground truth prompts exhibit scattered activations, often on uninformative regions such as the sky or background. In contrast, optimized prompts yield focused, coherent activations concentrated on discriminative geographic elements—building facades and roofs (Architectural), streets and infrastructure (Infrastructure), and signage or displays (Signage). These results demonstrate that optimized prompts improve semantic alignment between text and image, producing more consistent similarity distributions and attention concentrated on regions truly relevant to geolocation.

4.3 Comparison with Existing Geolocation Frameworks

To position GEO-Detective against existing geolocation approaches, we evaluate it on IM2GPS3K [42], a commonly used benchmark in prior image geolocation studies. We categorize existing approaches into agentic LVLM frameworks, LVLM-based geolocation systems, and trained geolocation models. The compared methods include GeoMiner [25], smileGeo [12], G3 [16], Img2Loc [53], GeoReasoner [19], PIGEON [11], and GeoCLIP [6]. As shown in Table 2, GEO-Detective achieves 11.3% accuracy within 1,km and 47.5% within 25,km. It ob-

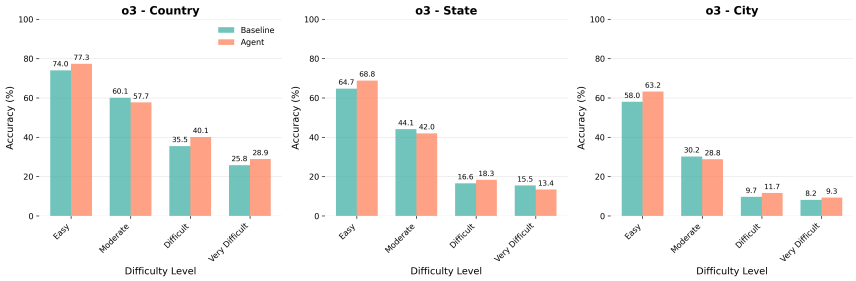

Fig. 3: Accuracy comparison of the baseline LVLM and GEO-Detective across difficulty levels.

Table 3: Unknown rate (%) of o3 under GEO-Detective and the baseline across difficulty levels.

	Baseline	GEO-Detective
Easy (269)	21.9	18.6
Moderate (281)	29.2	26.3
Difficult (349)	45.8	22.6
Very Difficult (97)	55.7	28.9
Extremely Difficult (4)	75.0	50.0

tains the best performance at the 25,km scale among the evaluated agentic LVLM frameworks, outperforming GeoMiner and smileGeo.

4.4 Comparison with Baseline LVLMs

First, we aim to assess whether the proposed agent poses greater privacy risks for shared images compared to a baseline LVLM. As shown in Figure 3, GEO-Detective consistently achieves higher accuracy than the baseline LVLM o3, particularly at higher difficulty levels. For instance, at the country level, accuracy in the "Difficult" setting increases from 35.5% to 40.1%, and in the "Very Difficult" setting from 25.8% to 28.9%. Similar gains appear at the city level, where performance improves from 9.7% to 11.7% on "Difficult" images and from 8.2% to 9.3% on "Very Difficult" images.

In addition, Table 3 shows that GEO-Detective significantly reduces the unknown rate for o3 in challenging cases. At the "Difficult" level, the rate drops from 45.8% to 22.6%, and at the "Very Difficult" level from 55.7% to 28.9%. These results demonstrate that the agent can provide more confident predictions when the task becomes challenging. Although the proposed agents are not always better than LVLMs on simple tasks, their use of multiple strategies becomes valuable in more difficult cases. Meanwhile, this stronger reasoning ability also amplifies privacy risks. Unlike baseline LVLMs, the proposed agent combines multiple strategies with external information retrieval, systematically extracting

Table 4: Country-level accuracy (%) for **o3**. Improvements over baseline are shown in parentheses. Abbreviations: Base = Baseline, EAP = Experience-Augmented Prompting, RS = Reverse Image Search, Seg = Segmentation, Ours = GEO-Detective. Difficulty levels are abbreviated as Easy, Mod. (Moderate), Diff. (Difficult), V.D. (Very Difficult), and Ext. (Extremely Difficult).

Diff.	Base	EAP	RS	EAP+RS	Base+Seg+RS	Ours
Easy_{269}	74.0	78.4 (+4.4)	76.6 (+2.6)	77.0 (+3.0)	74.7 (+0.7)	77.3 (+3.3)
Mod._{281}	60.1	61.6 (+1.5)	50.9 (-9.2)	57.3 (-2.8)	50.5 (-9.6)	57.7 (-2.4)
Diff._{349}	35.5	35.2 (-0.3)	37.8 (+2.3)	40.4 (+4.9)	32.1 (-3.4)	40.1 (+4.6)
V.D._{97}	25.8	15.5 (-10.3)	26.8 (+1.0)	33.0 (+7.2)	15.5 (-10.3)	28.9 (+3.1)
Ext._{4}	0.0	0.0 (+0.0)	0.0 (+0.0)	0.0 (+0.0)	0.0 (+0.0)	0.0 (+0.0)

Table 5: State/Region-level accuracy (%) for **o3**. Improvements over baseline are shown in parentheses.

Diff.	Base	EAP	RS	EAP+RS	Base+Seg+RS	Ours
Easy_{269}	64.7	67.3 (+2.6)	69.5 (+4.8)	71.0 (+6.3)	65.1 (+0.4)	68.8 (+4.1)
Mod._{281}	44.1	43.4 (-0.7)	38.1 (-6.0)	43.4 (-0.7)	35.9 (-8.2)	42.0 (-2.1)
Diff._{349}	16.6	15.8 (-0.8)	17.8 (+1.2)	18.9 (+2.3)	12.9 (-3.7)	18.3 (+1.7)
V.D._{97}	15.5	7.2 (-8.3)	13.4 (-2.1)	15.5 (+0.0)	7.2 (-8.3)	13.4 (-2.1)
Ext._{4}	0.0	0.0 (+0.0)	0.0 (+0.0)	0.0 (+0.0)	0.0 (+0.0)	0.0 (+0.0)

Table 6: City-level accuracy (%) for **o3**. Improvements over baseline are shown in parentheses.

Diff.	Base	EAP	RS	EAP+RS	Base+Seg+RS	Ours
Easy_{269}	58.0	60.6 (+2.6)	63.6 (+5.6)	63.2 (+5.2)	61.3 (+3.3)	63.2 (+5.2)
Mod._{281}	30.2	30.2 (+0.0)	28.8 (-1.4)	30.2 (+0.0)	23.5 (-6.7)	28.8 (-1.4)
Diff._{349}	9.7	9.5 (-0.2)	12.3 (+2.6)	11.7 (+2.0)	8.9 (-0.8)	11.7 (+2.0)
V.D._{97}	8.2	3.1 (-5.1)	9.3 (+1.1)	8.2 (+0.0)	7.2 (-1.0)	9.3 (+1.1)
Ext._{4}	0.0	0.0 (+0.0)	0.0 (+0.0)	0.0 (+0.0)	0.0 (+0.0)	0.0 (+0.0)

hidden geographic cues. These additional signals make it easier for adversaries to infer users' true locations, thereby increasing the threat to user privacy.

4.5 Ablation Study

We performed an ablation study to evaluate the contribution of each module and their privacy implications. Table 4 to Table 6 show the accuracy of GPT o3 at the country, state, and city level geolocation in different difficulty settings.

The Experience Augmented Prompting (EAP) module shows mixed effects. It slightly improves performance in some cases (for example, at the country level under moderate difficulty: from 60.1% to 61.6%) but may reduce accuracy when redundant contextual cues add noise (for example, at the country level

Table 7: Performance comparison between baseline LVLM and GEO-Detective on DOXBENCH.

Model	Difficulty	Baseline			GEO-Detective		
		Country	State	City	Country	State	City
Gemini 2.5 Pro	Easy	100.0	89.7	15.4	100.0	89.7	17.1
	Moderate	99.2	92.6	16.7	98.4	91.0	15.3
	Difficult	94.4	55.6	5.6	100.0	55.6	5.6
Gemini 2.5 Flash	Easy	100.0	89.7	13.7	100.0	91.5	15.4
	Moderate	98.9	89.3	18.4	97.3	86.6	17.8
	Difficult	83.3	50.0	5.6	88.9	50.0	5.6
o3	Easy	72.6	64.1	17.1	72.6	63.2	17.1
	Moderate	85.2	79.2	24.9	79.2	73.7	22.2
	Difficult	61.1	38.9	11.1	72.2	44.4	11.1
GPT-4o	Easy	99.1	88.9	29.1	90.6	81.2	23.9
	Moderate	88.5	83.3	30.7	78.1	71.8	23.6
	Difficult	83.3	44.4	5.6	100.0	55.6	11.1

under very difficult conditions: from 25.8% to 15.5%). The Reverse Search (RS) module provides the most consistent gains in challenging tasks. For example, at the country level under difficult conditions, performance improves from 35.5% to 37.8% and at the city level under difficult conditions from 9.7% to 12.3%. These results show that external visual references help the model infer location when intrinsic clues are limited. However, RS can be unstable in moderate cases (at the state level under moderate difficulty, from 44.1% to 38.1%), where visually similar but irrelevant matches may mislead reasoning. The Segmentation (Seg) module complements RS in difficult settings by isolating geographic subregions and filtering background noise, which improves retrieval reliability. For example, at the country level under difficult conditions, accuracy increases from 35.5% to 40.4% when segmentation is combined with RS.

The agent with autonomous decision making achieves the most consistent improvements. For example, at the country level under difficult conditions, the accuracy increases from 35.5% to 40.1%, and at the city level under difficult conditions from 9.7% to 11.7%. These results confirm that single strategies are not sufficient in complex settings, and an adaptive combination of multiple modules yields stronger performance. However, this adaptability also makes location-related information easier to capture, amplifying privacy risks in challenging scenarios.

Overall, the ablation results show that while EAP and RS may introduce noise or instability in simple cases, the multimodule framework, particularly the autonomous agent, outperforms baseline LVLMs at higher difficulty levels. The results for GPT-4o follow similar trends, showing larger improvements at higher difficulty levels. Detailed results for GPT-4o are provided in the Appendix.

4.6 Assessment of Generalizability

We acknowledge that the dataset used in our primary experiments may pose a potential data contamination risk, as some portion of the data could have been included in the pretraining corpus of the underlying language model. In our main experiments, we adopt a comparative evaluation setup in which both the baseline model and our agent framework are powered by the same underlying LVLM. Under such conditions, any potential contamination would affect both systems equally. Therefore, the performance improvement observed in our agent over the baseline cannot be attributed to pretraining exposure, but rather to the design of the proposed agent framework.

To further mitigate potential contamination concerns, we additionally evaluate our approach on the DOXBENCH dataset. This dataset was collected in April 2025 with 500 images across six California cities and is highly unlikely to overlap with the pretraining data of current LVLMs.

We evaluated four models, Gemini 2.5 Pro, Gemini 2.5 Flash, o3, and GPT-4o, on the DOXBENCH dataset (The results are shown in Table 7). All models achieve high accuracy at the Country and State levels but drop sharply at the City level, typically below 24.0%. Performance variations between GEO-Detective and baseline LVLMs are evident across different difficulty settings.

Most DOXBENCH samples are easy or moderate, containing rich yet sometimes inconsistent geographic elements. For example, a U.S. rural house with British style decorations may mislead models toward Europe. Under the GEO-Detective framework, models explicitly extract and integrate multiple cues, which can amplify such conflicts and slightly reduce accuracy in these cases. In contrast, difficult samples often contain few but distinctive clues. Baseline LVLMs tend to fail when such weak signals dominate, whereas GEO-Detective compels focused reasoning on limited but informative cues, leading to notable accuracy gains. Overall, DOXBENCH results suggest that GEO-Detective performs more robustly when geographic signals are sparse or subtle. However, when images contain potentially misleading or ambiguous cues, the agent's expanded search and integration of multiple signals may also amplify such ambiguity. This behavior is consistent with human geolocation reasoning, where additional but conflicting evidence can sometimes increase uncertainty rather than resolve it.

4.7 Defence Methods

To mitigate the privacy risks of geolocation models, we examine four defense strategies inspired by existing protection mechanisms in LVLMs. **Visual Prompt Injection** [43, 54] adds incorrect geographic labels to an image, creating conflicts between the visual content and the attached text that can mislead model reasoning. **Watermarking** [8] embeds explicit messages such as "Geolocation of this image is prohibited," encouraging the model to decline the task. **Trigger Based Defenses** [20,26] insert small visual symbols into the image, introducing an intervention signal that alters the model's reasoning when detected. Finally, **EXIF Modification** [48] alters or forges location related metadata, preventing

Table 8: Effectiveness of four different defense mechanisms under o3 model.

Baseline (o3)	Country	State	City	Unknown
original	50.0%	34.0%	27.0%	33.0%
Watermark	6.0%	5.0%	4.0%	94.0%
VPI	39.0%	31.0%	27.0%	15.0%
Trigger-based	49.0%	33.0%	29.0%	14.0%
EXIF	52.0%	38.0%	27.0%	30.0%
GEO-Detective (o3)	Country	State	City	Unknown
original	51.0%	34.0%	30.0%	21.0%
Watermark	10.0%	6.0%	5.0%	84.0%
VPI	41.0%	32.0%	27.0%	17.0%
Trigger-based	53.0%	37.0%	29.0%	18.0%
EXIF	51.0%	34.0%	32.0%	23.0%

the direct leakage of coordinates even though the visual content may still enable inference.

Table 8 summarizes the effectiveness of four defense methods on o3 under both the baseline and agent settings. Watermarking is the most effective, raising the unknown rate from 33% to 94% in the baseline model and from 21% to 84% in the agent model, effectively suppressing geolocation outputs. Visual prompt injection and trigger both introduce misleading visual or textual cues, which reduce accuracy by causing the model to incorporate incorrect evidence. In contrast, EXIF modification shows that the agent does not rely on metadata during reasoning. Despite these defenses, the agent remains consistently more robust than the baseline model, which makes its geolocation capability harder to suppress and raises additional privacy concerns.

5 Conclusion

We propose GEO-Detective, an agent-based system for image geolocation that integrates tools to support multistep reasoning. The proposed agent simulates humanlike strategies, extracting subtle geographic cues and leveraging external knowledge to improve inference over baseline LVLMs. We evaluate it on MP16-Pro and DoxBench, across multiple advanced LVLMs. The evaluation shows that GEO-Detective achieves higher accuracy in challenging scenarios, reduces the "unknown" rate by more than half, and increases risks related to location privacy. We study four defense methods and find that watermarking is the most effective, while other methods offer limited protection. Overall, GEO-Detective highlights both the power of agentic reasoning for geolocation and the urgent need for robust privacy defenses. The code will be released at `https://github.com/zxyreal/GEO-Detective`.

References

1. Flickr. `https://flickr.com`, last accessed: 2026-06-30

2. Gemini Flash. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash, last accessed: 2026-06-30
3. Gemini Pro. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro, last accessed: 2026-06-30
4. YOLOv5. https://zenodo.org/record/6222936, last accessed: 2026-06-30
5. John McAfee's location inadvertently revealed by photo metadata. https://www.theguardian.com/world/2012/dec/03/john-mcafee-location-revealed-vice (2012), last accessed: 2026-06-30
6. Cepeda, V.V., Nayak, G.K., Shah, M.: GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. In: Annual Conference on Neural Information Processing Systems (NeurIPS). pp. 8690–8701. NeurIPS (2023)
7. Chen, T., Li, P., Zhou, K., Chen, T., Wei, H.: Unveiling Privacy Risks in Multimodal Large Language Models: Task-specific Vulnerabilities and Mitigation Challenges. In: Annual Meeting of the Association for Computational Linguistics (ACL). pp. 4573–4586. ACL (2025)
8. Deng, A., Cao, T., Chen, Z., Hooi, B.: Words or Vision: Do Vision-Language Models Have Blind Faith in Text? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2025)
9. Future, B.: The Hidden Fingerprint Inside Your Photos. https://www.bbc.co.uk/future/article/20210324-the-hidden-fingerprint-inside-your-photos, last accessed: 2026-06-30
10. Haas, L., Alberti, S., Skreta, M.: Learning Generalized Zero-Shot Learners for Open-Domain Image Geolocalization. CoRR abs/2302.00275 (2023)
11. Haas, L., Skreta, M., Alberti, S., Finn, C.: PIGEON: predicting image geolocations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 12893–12902. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01225, https://doi.org/10.1109/CVPR52733.2024.01225
12. Han, X., Zhu, C., Zhu, H., Zhao, X.: Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In: Antonie, L., Pei, J., Yu, X., Chierichetti, F., Lauw, H.W., Sun, Y., Parthasarathy, S. (eds.) Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025. pp. 814–825. ACM (2025). https://doi.org/10.1145/3711896.3737141, https://doi.org/10.1145/3711896.3737141
13. Hays, J., Efros, A.A.: IM2GPS: estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2008)
14. Huang, J., Huang, J., Liu, Z., Liu, X., Wang, W., Zhao, J.: VLMs as GeoGuessr Masters: Exceptional Performance, Hidden Biases,and Privacy Risks. CoRR abs/2502.11163 (2025)
15. Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In: International Conference on Machine Learning (ICML). pp. 4904–4916. PMLR (2021)
16. Jia, P., Liu, Y., Li, X., Zhao, X., Wang, Y., Du, Y., Han, X., Wei, X., Wang, S., Yin, D.: G3: An Effective and Adaptive Framework for Worldwide Geolocalization Using Large Multi-Modality Models. In: Annual Conference on Neural Information Processing Systems (NeurIPS). pp. 53198–53221. NeurIPS (2024)

17. Jia, P., Park, S., Gao, S., Zhao, X., Li, Y.: GeoRanker: Distance-Aware Ranking for Worldwide Image Geolocalization. CoRR abs/2505.13731 (2025)
18. Jiang, Q., Chen, X., Zeng, Z., Yu, J., Zhang, L.: Rex-Thinker: Grounded Object Referring via Chain-of-Thought Reasoning. CoRR abs/2506.04034 (2025)
19. Li, L., Ye, Y., Jiang, B., Zeng, W.: Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In: Salakhutdinov, R., Kolter, Z., Heller, K.A., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. Proceedings of Machine Learning Research, vol. 235, pp. 29222–29233. PMLR / OpenReview.net (2024), https://proceedings.mlr.press/v235/li24ch.html
20. Liang, J., Liang, S., Luo, M., Liu, A., Han, D., Chang, E., Cao, X.: Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. CoRR **abs/2402.13851** (2024). https://doi.org/10.48550/ARXIV.2402.13851, https://doi.org/10.48550/arXiv.2402.13851
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Annual Conference on Neural Information Processing Systems (NeurIPS). NeurIPS (2023)
22. Liu, X., Zhu, Y., Lan, Y., Yang, C., Qiao, Y.: Safety of Multimodal Large Language Models on Images and Text. In: International Joint Conferences on Artifical Intelligence (IJCAI). pp. 8151–8159. IJCAI (2024)
23. Liu, Y., Deng, G., Ding, J., Li, Y., Zhang, T., Sun, W., Zheng, Y., Ge, J.: Mission: Impossible - image-based geolocation with large vision language models. Proc. Priv. Enhancing Technol. **2025**(4), 410–428 (2025). https://doi.org/10.56553/POPETS-2025-0137, https://doi.org/10.56553/popets-2025-0137
24. Liu, Y., Ding, J., Deng, G., Li, Y., Zhang, T., Sun, W., Zheng, Y., Ge, J., Liu, Y.: Image-Based Geolocation Using Large Vision-Language Models. CoRR abs/2408.09474 (2024)
25. Luo, W., Zhang, Q., Lu, T., Liu, X., Zhao, Y., Xiang, Z., Xiao, C.: Doxing via the Lens: Revealing Privacy Leakage in Image Geolocation for Agentic Multi-Modal Large Reasoning Model. CoRR abs/2504.19373 (2025)
26. Lyu, W., Pang, L., Ma, T., Ling, H., Chen, C.: TrojVLM: Backdoor Attack Against Vision Language Models. In: European Conference on Computer Vision (ECCV). pp. 467–483. Springer (2024)
27. Mai, G., Lao, N., He, Y., Song, J., Ermon, S.: CSP: Self-Supervised Contrastive Spatial Pre-Training for Geospatial-Visual Representations. In: International Conference on Machine Learning (ICML). PMLR (2023)
28. Mendes, E., Chen, Y., Hays, J., Das, S., Xu, W., Ritter, A.: Granular Privacy Control for Geolocation with Vision Language Models. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). p. 17240–17292. ACL (2024)
29. Mink, J., Luo, L., Barbosa, N.M., Figueira, O., Wang, Y., Wang, G.: DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In: USENIX Security Symposium (USENIX Security). pp. 1669–1686. USENIX (2022)
30. Muller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification. In: European Conference on Computer Vision (ECCV). pp. 575–592. Springer (2018)
31. OpenAI: GPT-4o. https://openai.com/index/hello-gpt-4o/, last accessed: 2026-06-30
32. OpenAI: o3. https://platform.openai.com/docs/models/o3, last accessed: 2026-06-30

33. Orekondy, T., Schiele, B., Fritz, M.: Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images. In: IEEE International Conference on Computer Vision (ICCV). pp. 3706–3715. IEEE (2017)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
35. Roy, P.C., Boddeti, V.N.: Mitigating Information Leakage in Image Representations: A Maximum Entropy Approach. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2586–2594. IEEE (2019)
36. Seo, P.H., Weyand, T., Sim, J., Han, B.: CPlaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps. In: European Conference on Computer Vision (ECCV). pp. 536–551. Springer (2018)
37. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. In: Annual Conference on Neural Information Processing Systems (NeurIPS). pp. 38154–38180. NeurIPS (2023)
38. 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. CoRR abs/2502.13759 (2025)
39. Su, J., Kempe, J., Ullrich, K.: Mission Impossible: A Statistical Perspective on Jailbreaking LLMs. CoRR abs/2408.01420 (2024)
40. Suresh, S., Chodosh, N., Abello, M.: DeepGeo: Photo Localization with Deep Neural Network. CoRR abs/1810.03077 (2018)
41. Tömekçe, B.B., Vero, M., Staab, R., Vechev, M.T.: Private Attribute Inference from Images with Vision-Language Models. In: Annual Conference on Neural Information Processing Systems (NeurIPS). pp. 103619–103651. NeurIPS (2024)
42. Vo, N.N., Jacobs, N., Hays, J.: Revisiting IM2GPS in the deep learning era. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 2640–2649. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.286, https://doi.org/10.1109/ICCV.2017.286
43. Wang, H., Ma, Q.: Textural or Textual: How Vision-Language Models Read Text in Images. In: International Conference on Machine Learning (ICML). PMLR (2025)
44. Wang, K., Huang, Y., M., J.O., Gool, L.V., Tuytelaars, T.: An Analysis of Human-Centered Geolocation. In: Winter Conference on Applications of Computer Vision (WACV). pp. 2058–2066. IEEE (2018)
45. Wang, Z., Xu, D., Khan, R.M.S., Lin, Y., Fan, Z., Zhu, X.: LLMGeo: Benchmarking Large Language Models on Image Geolocation In-the-wild. CoRR abs/2405.20363 (2024)
46. Weyand, T., Kostrikov, I., Philbin, J.: PlaNet - Photo Geolocation with Convolutional Neural Networks. In: European Conference on Computer Vision (ECCV). pp. 37–55. Springer (2016)
47. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. CoRR abs/2303.04671 (2023)
48. Xu, H., Wang, H., Stavrou, A.: Privacy Risk Assessment on Online Photos. In: Research in Attacks, Intrusions, and Defenses (RAID). pp. 427–447. Springer (2015)
49. Yang, Y., Wang, S., Li, D., Sun, S., Wu, Q.: GeoLocator: A location-integrated large multimodal model (LMM) for inferring geo-privacy. Applied Sciences (2024)

50. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: ReAct: Synergizing Reasoning and Acting in Language Models. In: International Conference on Learning Representations (ICLR). ICLR (2023)
51. Yerramilli, S., Pande, N., Grover, R., Tamarapalli, J.S.: GeoChain: Multimodal Chain-of-Thought for Geographic Reasoning. CoRR abs/2506.00785 (2025)
52. Zheng, Y., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T., Neven, H.: Tour the World: building a web-scale landmark recognition engine. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1085–1092. IEEE (2009)
53. Zhou, Z., Zhang, J., Guan, Z., Hu, M., Lao, N., Mu, L., Li, S., Mai, G.: Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024. pp. 2749–2754. ACM (2024). https://doi.org/10.1145/3626772.3657673, https://doi.org/10.1145/3626772.3657673
54. Zhu, R.: Text Prompt Injection of Vision Language Models. CoRR abs/2510.09849 (2025)