Mission: Impossible - Image Based Geolocation with Large Vision Language Models

Yi Liu
Quantstamp
yi009@e.ntu.edu.sg

Gelei Deng*
Nanyang Technological University
gdeng003@e.ntu.edu.sg

Junchen Ding
University of New South Wales
jamison.ding@student.unsw.edu.au

Yuekang Li
University of New South Wales
yuekang.li@unsw.edu.au

Tianwei Zhang
Nanyang Technological University
tianwei.zhang@ntu.edu.sg

Weisong Sun
Nanyang Technological University
weisong.sun@ntu.edu.sg

Yaowen Zheng
Institute of Information Engineering,
Chinese Academy of Sciences
zhengyaowen@iie.ac.cn

Jingquan Ge
Nanyang Technological University
jingquan.ge@ntu.edu.sg

Abstract

In the age of ubiquitous smartphone use and widespread image sharing on social platforms, geolocation poses a critical privacy concern. Images often carry sensitive spatial and temporal details—such as street signs, architectural styles, or landmarks—that can inadvertently disclose the precise whereabouts of individuals and organizations. Recent advances in large vision-language models (LVLMs) present an *emerging threat* by enabling users, regardless of technical expertise, to extract location cues from seemingly benign photos. While existing AI-driven geolocation solutions often focus on narrow datasets or specialized contexts, the generalizable performance and privacy implications of zero-shot LVLMs in real-world settings remain critical questions.

In this paper, we investigate the geolocation capabilities of state-of-the-art LVLMs. Our findings reveal that while these models demonstrate a *non-negligible* capability for image-based geolocation even without specialized training, their accuracy in absolute terms is often *low*, exposing *clear limitations* in their current state. We then introduce ETHAN, a framework integrating chain-of-thought (CoT) reasoning. Although ETHAN shows improved performance (e.g., 28.7% accuracy at the 1 km threshold) and an 85.4% win rate on GeoGuessr, these results primarily highlight the potential trajectory of such technologies rather than their current widespread, high-accuracy applicability. Our study underscores the dual nature of LVLMs in this domain: they uncover an *emerging privacy risk* due to their inherent, albeit limited, geolocation abilities, yet also demonstrate significant *constraints*. We conclude by *calling for further research* into the limitations and risks of LVLM-based geolocation and the development of effective *mitigation strategies* to protect sensitive location data.

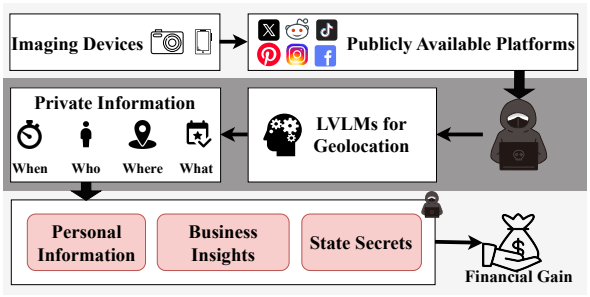

Figure 1: Common scenarios of how adversaries extract private geolocation information from the victims.

1 Introduction

Geolocation [16, 36, 42, 46] plays a pivotal role in safeguarding user privacy. As smartphones and other mobile devices become ubiquitous, sharing images on social platforms such as Facebook [38], Instagram [39], and Foursquare [20] has grown increasingly common. These shared images can unintentionally expose sensitive details, including the time and location of events, identities of individuals, and interpersonal connections, thereby posing significant threats to user privacy. Multiple incidents [8, 58] have demonstrated how private photos can be misused, leading to severe repercussions like job loss. Consequently, the ability to accurately infer a photo's location, often referred to as "image-based geolocation," has profound implications for security, navigation, and social media, making it a crucial privacy concern.

Given that photos are increasingly tagged with both geographic coordinates and timestamps, image privacy now heavily intersects with location privacy. This situation can have dire consequences, especially in high-stakes environments, as illustrated in Figure 1. Attackers can exploit geolocation capabilities to extract private information—ranging from personal and commercial data to state secrets—potentially securing financial or strategic advantage. Consider, for example, a human rights activist who attends a confidential meeting in a remote area. If someone inadvertently posts a photo from that same location, an adversary might use the image

*Corresponding author.

Proceedings on Privacy Enhancing Technologies 2025(4), 410–428

https://doi.org/10.56553/popets-2025-0137

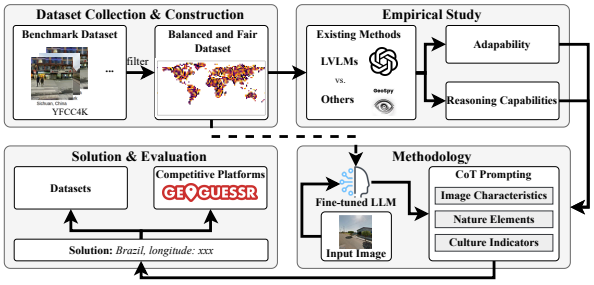

Figure 2: Overview of our work.

to pinpoint the meeting's venue, thereby endangering the activist through harassment, surveillance, or physical harm. In Section 2, we provide additional evidence illustrating the feasibility of such attacks.

Numerous AI-driven tools already exist for geolocation tasks. Advanced models such as GeoSpy [3] can not only predict locations from images but also furnish detailed justifications for their inferences. However, these solutions often require specialized expertise and substantial setup, limiting their accessibility. The advent of large vision-language models (LVLMs) [60, 63–65] introduces a new dimension to this challenge. These models, known for their proficiency in various complex tasks [13, 15, 15, 32], can potentially be used by non-technical users to predict a photo's location by interpreting visual cues. While this capability represents an *emerging threat* to privacy by potentially lowering the barrier to location inference, the actual effectiveness of current zero-shot LVLMs for reliable geolocation is still a subject of investigation. Early observations [44] suggest that while LVLMs can perform geolocation to some extent, their accuracy can be quite *limited in absolute terms*, indicating *clear limitations*.

Despite the remarkable progress in LVLM-based geolocation, there is a critical gap in our understanding of how these models perform when inferring locations from real-world images. Although existing research [25, 37, 53] has addressed geolocation in isolated contexts or on specialized datasets, a comprehensive evaluation spanning diverse environments and data sources remains lacking. Consequently, it is imperative to undertake a systematic study not only of the accuracy of these models but also of the factors that shape their performance and their potential security implications. This work is crucial for developing guidelines and technologies that protect users from the unintended privacy risks associated with geolocation capabilities.

In this paper, we begin by systematically assessing the ability of state-of-the-art LVLMs to extract geolocation information from images and by contrasting their performance with that of traditional geolocation frameworks. Specifically, we address three research questions: (1) *How effectively can current LVLMs perform geolocation tasks?* (2) *Can LVLMs be leveraged or adapted to exploit user privacy in real-world scenarios?* (3) *Which factors most significantly affect their geolocation proficiency?* Our approach employs a rigorous experimental testbed that incorporates varied datasets, robust evaluation metrics, and a thorough analysis of the models' outputs.

Our observations reveal that while LVLMs exhibit a non-negligible ability to geolocate images by recognizing landmarks or urban features, their *overall accuracy remains low* without significant fine-tuning or specialized prompting. Their reliance on "landmark knowledge" can hinder broader contextual reasoning, a task where human experts often incorporate more diverse cues. In contrast, human geoguessing experts factor in additional cues like terrain, weather, architectural details, and vegetation to enhance their accuracy.

Motivated by these observations and the need for a more balanced understanding, we propose Ethan, a framework that integrates LVLMs with chain-of-thought (CoT) reasoning [55, 61] to *explore the upper bounds* of such systems. While ETHAN achieves improved accuracy (e.g., 28.7% at 1 km) and performs well on GeoGuessr, its development also serves to *highlight the current boundaries* and complexities involved in achieving reliable image geolocation.

We evaluate Ethan on the large-scale GeolocationHub dataset, which comprises 50,000 data points (30,000 for training and 20,000 for testing). As our experiments demonstrate, Ethan achieves state-of-the-art accuracy across multiple distance thresholds, from 1 km (street-level) to 2,500 km (continent-level). Notably, Ethan attains 28.7% accuracy at street-level (1 km), 59.2% at city-level (25 km), 91.4% at region-level (200 km), 95.6% at country-level (750 km), and 99.3% at continent-level (2,500 km). In terms of average distance and composite score, Ethan maintains an average error of 499.3 km and an average GeoScore of 4620.9, surpassing other baseline models across every metric.

Additionally, Ethan demonstrates robust performance in real-world conditions, as illustrated by its success in the popular *GeoGuessr* game. Over multiple rounds, Ethan averages a score of 4550.5, substantially exceeding the human benchmark of 4120.3, with a win rate of 85.4%. In one example, Ethan correctly identified a complex urban location within 0.3 km of the true coordinates by capitalizing on subtle architectural and cultural cues. Ablation studies still confirm the performance of Ethan. These results underscore Ethan's adaptability and effectiveness in both controlled laboratory settings and competitive live environments.

Our contributions are threefold:

(1) **Evaluation and Emerging Threat Assessment:** We provide a systematic assessment of LVLMs' baseline geolocation capabilities, highlighting that while an *emerging threat* exists due to their non-negligible performance, their absolute accuracy is currently limited.
(2) **Performance Analysis and Limitations:** We investigate factors influencing LVLM geolocation accuracy, underscoring the impact of data quality, landmark availability, and generalization capabilities, thereby *exposing clear limitations* of current models.
(3) **Framework Exploration and Call for Mitigation:** We introduce Ethan [2] to demonstrate how LVLM reasoning can be structured for geolocation. More importantly, our overall findings serve as a *call for mitigation research* to address the potential privacy risks as these technologies mature.

This work aims to foster a nuanced understanding of LVLM-based image geolocation, recognizing both the emerging risks and the

existing constraints, to encourage proactive research into privacy-preserving measures.

2 Background

2.1 Location-dependent Privacy

A photograph can inadvertently disclose a person's location through multiple avenues. Typically, images carry metadata like EXIF [5] (Exchangeable Image File Format), which encodes details such as the date and GPS coordinates of when and where the photo was taken. Although social media platforms such as Facebook [38] and Instagram [39] strip this metadata from uploaded images, they still store it in separate databases. Should these databases be compromised by unauthorized access, attackers can track users far more efficiently. Beyond metadata, the images themselves can reveal location clues through visible landmarks or street signs. Moreover, *crowdsourcing*—where individuals familiar with a specific locale identify it from a photo—can further compromise privacy. As a result, simply removing metadata from photos is insufficient to protect the location privacy of those depicted [8].

2.2 Image Localizability

The ease of determining an image's location depends heavily on its content and distinct features. As illustrated in Figure 3, images that are straightforward to localize often include recognizable faces, unique animals, notable objects in indoor settings, striking natural landmarks, or prominent urban structures and historical sites. Street-view photos, for instance, are frequently more straightforward to geolocate because they contain recognizable buildings, signage, or other context-specific cues, compared to less distinctive rural backdrops [58]. The availability of relevant data and existing regulations also influence localizability. For example, advertisements are often easy to pinpoint because they frequently display explicit place names or other direct indicators that tie them to a specific region.

2.3 Overview of Geolocation Techniques

Visual geolocation generally aims to estimate 2D coordinates or identify broad regions such as countries, striving to balance wide applicability and moderate accuracy, even in locales not present in the training set. Existing methods can be divided into four broad categories: *image retrieval*, *classification*, *hybrid*, and *LVLM*-based approaches. While each category offers unique advantages, all face limitations regarding model performance, training complexity, and real-world applicability.

Image Retrieval-based Methods. One intuitive strategy is to match a query image against a massive reference database, then assign the location of the closest match as the prediction. Early work [30, 34, 41] relied on rudimentary features such as color histograms, GIST descriptors, or texture clues. Later methods incorporated SIFT features and SVMs [29], and deep learning eventually enhanced retrieval through sophisticated learned representations. Although these approaches can be highly effective when the database is extensive and current, they do not learn an intrinsic representation of the scene. As a result, their accuracy diminishes in regions with sparse coverage or when environmental changes occur over time.

Classification-Based Methods. Another way to frame geolocation is as a classification task [31, 57], wherein the world is partitioned into discrete *cells* based on latitude and longitude. These cells can be arranged in various ways—regular, adaptive, semantically driven, combinatorial, administrative, or hierarchical. The key challenge lies in balancing the number and size of these cells: overly large cells reduce location precision, while too many tiny cells risk insufficient training data per cell.

Hybrid Approaches. To overcome the pitfalls of simple discretization, some methods [14, 25] merge retrieval and classification, often employing ranking losses or contrastive objectives. One approach [25] initially applies a coarse classification, then refines it via regression using prototype networks. However, hybrid solutions demand carefully balanced training sets for multiple modules, which in turn increases computational complexity and resource requirements.

LVLM-based Methods. LVLMs show considerable promise for image-based geolocation. State-of- the-art architectures, such as GPT-4 [23], have been adapted to geolocation tasks [8], and specialized applications [44] leverage LVLMs to identify geographic cues—ranging from text on signage to architectural details and natural landmarks. Recent research continues to explore their capabilities, for example, by examining fine tuning for diverse cultural contexts [21], employing multi-image prompting techniques [43], and assessing zero-shot generalization for out-of-distribution scenarios [19]. Nevertheless, the full range and limitations of LVLM capabilities in geolocation are still not well understood. Additionally, these models can bypass many traditional privacy defenses (e.g., stripping metadata), raising new challenges for ensuring location-dependent privacy. Despite the usage of LVLMs on geolocation, the limitations and boundaries of LVLM capabilities remain largely unexplored. Con- sequently, how to effectively harness their full potential, while mitigating their weaknesses, is still an open question. Additionally, this introduces a new layer of complexity in managing location- dependent privacy, as LVLMs can bypass traditional privacy pro- tections that focus merely on removing or obscuring metadata.

3 Threat Model

3.1 Threat Model Formulation

We consider a realistic scenario where adversaries aim to uncover the precise geographic coordinates (latitude and longitude) of images taken in real-world settings. These images are assumed to be unaltered and capture natural scenes (e.g., urban streets, rural landscapes, or indoor/outdoor gatherings). Importantly, the attackers do not have access to any additional metadata, such as EXIF tags or side-channel information (e.g., timestamps, social media logs, or user profiles). Instead, they rely exclusively on visual cues within the images—such as architectural styles, vegetation, road layouts, signage, and other salient features—when attempting to locate the image on a world map. This restriction reflects common conditions wherein sensitive information is unintentionally leaked through photo sharing, yet remains unknown to the adversaries except for what can be gleaned from the picture itself.

The adversaries' primary goal is to maximize the precision of their geolocation predictions, measured in terms of how closely they

Figure 3: Visual representation of image localizability spectrum, categorized from non-localizable scenes to recognizable landmarks, illustrating the diversity in the dataset.

approximate the actual coordinates of each image. Even moderate accuracy—correct to within a few kilometers—can lead to substantial privacy breaches, such as revealing an individual's place of residence or exposing a covert meeting site. Highly accurate predictions (e.g., within street-level accuracy of 1 km) further intensify these risks, enabling malicious actors to orchestrate stalking, harassment, or strategic targeting. This threat model thus highlights the peril of unintentional location disclosure, wherein adversaries—ranging from sophisticated cybercriminals to casual onlookers—can exploit seemingly harmless photos for harmful ends. By focusing on visual cues alone, we capture a worst-case yet increasingly realistic scenario in which advanced AI models are leveraged to identify and contextualize minute geographic details. Consequently, our research prioritizes methods to evaluate the severity of these risks and proposes potential safeguards that preserve user privacy without unduly limiting the utility of image-sharing platforms.

3.2 Alignment with Regulatory Frameworks

The threat model for ETHAN, wherein adversaries derive geolocation from images, has significant implications under contemporary data protection and AI governance laws. Processing such data can implicate Article 9 of the GDPR [17] if the inferred location reveals sensitive personal details (e.g., attendance at a political rally or specific healthcare facility), requiring stringent processing conditions. Furthermore, entities systematically deriving and commercializing this geolocation data could be classified as "data brokers" under regulations like California's CPRA [47], incurring specific registration and consumer rights obligations. Crucially, the EU AI Act [18] may classify AI systems like ETHAN as "high-risk" if their application significantly impacts fundamental rights (e.g., privacy through enabling stalking or surveillance, as outlined in our threat model) or if used in specified contexts like biometric identification or certain law enforcement activities. These frameworks collectively emphasize the critical need for robust ethical assessments and legal compliance when developing and deploying advanced image geolocation technologies.

4 Experimental Framework

Before benchmarking existing geolocation techniques and comparing them with LVLM-based solutions, we first outline our approach to data collection, model selection, and the overall experimental design guiding our empirical evaluation.

Table 1: Detailed review of geolocation techniques adopted in the recent three years, highlighting their unique features and application scopes.

Technique	Description
StreetClip [24]	Clip-based approach for urban geolocation
GeoClip [10]	Alignment technique inspired by clip-based models
GPT-4o [23]	Advanced general LVLM
LLaVA [35]	Open-source LVLM with visual processing capabilities
GeoSpy [3]	Commercial tool for geolocation analysis

4.1 Design Overview

The design of our empirical study is driven by three primary objectives:

- **Scenario Coverage:** Our goal is to encompass a wide range of geolocation scenarios—from busy urban streets to remote natural landmarks. This diversity helps assess how robust and adaptive different geolocation techniques are under varying environmental conditions.
- **State-of-the-Art Techniques:** By focusing on geolocation solutions developed or substantially updated within the last three years, we ensure that our findings are aligned with the latest methodological and technological advances.
- **Reproducibility and Accessibility:** We prioritize publicly accessible techniques, including those with released model weights and datasets, to facilitate reproducible research and practical deployment in real-world settings.

Following these principles, our empirical framework comprises three core components:

(1) *Dataset Compilation:* We carefully select existing datasets to capture a range of geolocation contexts, ensuring thorough evaluation across diverse environments. In particular, after reviewing available data sources, we curate our own dataset (Section 4.3) to mitigate biases and data leakage present in existing datasets.
(2) *Technique Selection:* Based on the criteria above—public accessibility, recency, and availability of resources—we choose a set of cutting-edge geolocation methods for analysis.
(3) *Evaluation Framework:* Building on prior research, we employ multiple metrics, including Haversine distance [56], GeoScore [6], and administrative boundary accuracy [10], to comprehensively assess geolocation performance.

By pursuing these strategies, our study probes the capability boundaries of state-of-the-art geolocation methods, examining both their security implications and potential vulnerabilities.

4.2 Collection of Geolocation Methods

We perform a broad review of both academic and commercial geolocation techniques, selecting those that reflect cutting-edge developments in the field.

Selection Criteria. We use the following criteria to choose geolocation methods:

- *Public Accessibility:* Techniques must provide publicly available pre-trained models or APIs to ensure broader applicability.
- *Availability of Weights and Datasets:* We prioritize methods that release training code and datasets, thereby fostering reproducibility.

- *Temporal Relevance:* We include only techniques created or significantly updated within the last three years to capture the latest progress.

Selection Results. Table 1 lists the five final techniques we evaluate. This set includes two academic image-geolocation solutions (StreetClip [24] and GeoClip [10]), two large vision language models (GPT-4o [23] and LLaVA [35]), and one commercial platform (GeoSpy [3]). Combining commercial and open-source solutions ensures both technological diversity and practical relevance.

Table 2: Summary of datasets utilized in geolocation techniques, indicating the variety and scale of images used in our analysis.

Dataset	Number of Images	Cutoff Date
Im2GPS [26]	237	2008
Im2GPS3k [50]	2,997	2017
YFCC4k [50]	4,536	2017
YFCC26k [48]	26,000	2022

4.3 Dataset Construction

While exploring open-source geolocation datasets, we identified biases and data leakage that could invalidate results. To resolve these issues, we developed our own dataset, which we describe below, starting with an assessment of existing open-source data.

Data Collection and Verification. We surveyed widely used datasets from prior geolocation research, ultimately selecting four for further evaluation, as listed in Table 2. These include Im2GPS [26], Im2GPS3k [50], YFCC4k [50], and YFCC26k [48], each offering varying scales and coverage. Our verification process involved two key criteria: (1) *Diversity*, requiring a balanced global representation, and (2) *Integrity*, ensuring that each image contained enough contextual information for meaningful location estimation.

However, as shown in Figure 4, a sizeable fraction of these images proved unsuitable for localization, falling into three categories: *Minimal Context*, *Contextually Ambiguous*, or *Highly Misleading*. These findings underscore the necessity for high-quality, context-rich datasets in geolocation tasks.

Dataset Construction. We introduce GeolocationHub, a 50,000-image dataset engineered to address the shortcomings observed in existing resources. Two major strategies guide GeolocationHub's development:

- *Indoor Scene Filtering.* We eliminate indoor images—often rife with extraneous details—using (1) multi-view image analysis [33] and (2) keyword filters on annotations (e.g., "indoor," "room," "bed"). Keywords are iteratively refined until the dataset contains no incorrectly classified images.
- *Geographically Balanced Sampling.* We sample images worldwide to mitigate location bias, ensuring neither urban centers nor remote regions are disproportionately represented.

Figure 6 (omitted for brevity) displays the global distribution of images in GeolocationHub. Collectively, these measures yield a robust benchmark for assessing LVLM-based geolocation performance.

4.4 Evaluation Framework Design

We adopt a structured pipeline to benchmark conventional (non-LVLM) and LVLM-based geolocation techniques. Traditional methods are tested directly on GeolocationHub, while evaluating LVLM approaches demands additional care due to the impact of *prompt engineering* [12, 51]. Consequently, we implement zero-shot, few-shot, and chain-of-thought prompting strategies and fix the temperature to zero to ensure reproducible model outputs. We detail our prompt strategies below.

4.4.1 LVLM Prompt Design. Following established best practices [45, 49, 52], we employ three types of prompts:

- *Zero-shot Prompts:* Provide only a direct task description. For instance:

Zero-shot Prompt

> You are recognized as the world's foremost expert in geolocation analysis. Your objective is to meticulously analyze the provided image and determine its latitude and longitude.

- *Few-shot Prompts:* Include brief examples of image descriptions paired with correct geolocations, guiding the model to produce well-structured responses.
- *Chain-of-Thought Prompts:* Encourage step-by-step reasoning, aligning with evidence that iterative reasoning improves performance on complex tasks [54].

4.4.2 Evaluation Metrics. To measure geolocation performance in a comprehensive manner, we use the following metrics:

(1) **Haversine Distance:** A common metric that calculates the great-circle distance between two latitude-longitude pairs [27]:

$$d = 2r \arcsin\left(\sqrt{v}\right), \quad v = \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right), \tag{1}$$

where ϕ_1, λ_1 and ϕ_2, λ_2 are the latitude and longitude pairs, and r is Earth's approximate radius.

(2) **GeoScore:** Inspired by the *GeoGuessr* game [1], GeoScore computes a distance-based score:

$$\text{GeoScore} = 5000 \cdot \exp\left(-\frac{d}{1492.7}\right),$$

where d is the distance error in kilometers.

(3) **Administrative Boundaries:** In line with prior work [10, 40], we use five distance thresholds—1 km (street), 25 km (city), 200 km (region), 750 km (country), and 2500 km (continent)—to measure how often predictions fall within these ranges. This approach provides a more robust view of performance, minimizing the skewing effect of outliers on average distances.

By combining these metrics, we capture both fine-grained and coarse-grained geolocation accuracy, offering a holistic perspective on each method's real-world viability.

5 Empirical Study Results and Findings

Following our experimental framework, we conducted an empirical study to evaluate the effectiveness of various geolocation strategies. This section presents our results and key insights, comparing both traditional (non-LVLM) and LVLM-based approaches.

Figure 4: Categorization of images in the original dataset based on their localizability: "Minimal Context" for images with minimal geographic markers, "Contextually Ambiguous" for visually descriptive but non-localizable images, and "Highly Misleading" for ambiguous images leading to significant localization errors.

Table 3: Performance of geolocation techniques on various datasets, evaluated using GeoScore (GS) and Administrative Boundaries (AB) metrics with different prompting techniques.

Technique		GeoScore (0-5000)			Administrative Boundaries (Accuracy %)				
		Im2GPS	Im2GPS3k	YFCC26k	Street (1 km)	City (25 km)	Region (200 km)	Country (750 km)	Continent (2500 km)
Pre-LVLM	StreetClip [24]	3520.5	3591.3	3378.4	10.2	28.7	52.1	69.4	80.5
	GeoClip [10]	3535.7	3612.1	3411.9	12.1	30.4	54.2	71.8	82.3
LLaVA	Zero-shot	4086.4	4132.2	4033.8	15.8	35.5	58.9	75.3	85.7
	Few-shot	4112.9	4161.4	4061.5	17.3	37.8	61.2	77.1	87.0
	Chain-of-thought	4131.7	4180.8	4087.9	18.4	39.1	63.5	78.6	88.2
GPT-4o	Zero-shot	4345.4	4392.8	4289.7	20.7	42.3	66.7	81.5	91.0
	Few-shot	4378.2	4417.9	4312.6	21.9	44.1	68.3	83.0	92.1
	Chain-of-thought	4403.1	4443.5	4340.2	23.2	46.0	70.1	84.4	93.2
	GeoSpy [3]	4570.8	4620.5	4451.6	25.0	53.2	74.0	89.0	97.3

5.1 Geolocation Task Performance

5.1.1 Cross-model Comparison. Table 3 summarizes the performance of each model. Overall, all methods—both traditional solutions and LVLM-based models—attain some level of geolocation accuracy without requiring prior domain knowledge. Across multiple datasets, all models achieve predictions with Haversine distances below 10 km, providing city-level accuracy. However, pre-LVLM techniques, such as StreetClip [24] and GeoClip [10], exhibit limitations in complex urban environments and unfamiliar rural settings, primarily due to their reliance on predefined features and static models. In contrast, LVLMs deliver higher accuracy on more diverse datasets.

For instance, StreetClip achieves a GeoScore of 3520.5 on the relatively constrained Im2GPS dataset but drops to 3378.4 on the more varied YFCC26k dataset. These results highlight the growing need for adaptive, context-aware geolocation methods, a need that LVLM-based solutions address more effectively.

Finding 1: Modern LVLMs can successfully perform geolocation tasks without any specialized training or supplementary contextual information, underscoring critical risks to geolocation privacy.

We further explore differences in LVLM performance by examining various prompting strategies. Both GPT-4o and LLaVA excel when chain-of-thought prompting is used, significantly surpassing other methods across all datasets. As an example, GPT-4o achieves a GeoScore of 4403.1 on Im2GPS3k with chain-of-thought prompting, compared to 4345.4 without it. LLaVA shows a similar pattern, improving its GeoScore from 4086.4 to 4131.7. These findings demonstrate the value of leveraging iterative reasoning, suggesting that explicit reasoning steps can enhance geolocation accuracy.

Finding 2: Among the tested industrial-scale models (excluding GeoSpy), GPT-4o exhibits the most robust performance

in complex urban scenarios. LLaVA, particularly under chain-of-thought prompting, excels in settings requiring detailed contextual reasoning and adaptability.

5.1.2 *Examples of Success and Failure.*

Successful Geolocation Cases. We observe that some methods achieve near-perfect location predictions on landmark-rich images. For instance, GeoSpy correctly identifies the Eiffel Tower in Paris, earning a GeoScore of 5000 and locating the site within the 1 km street-level threshold. Another high-accuracy example is the Statue of Liberty, for which GeoSpy again achieves the maximum GeoScore, successfully pinpointing the monument within 1 km.

Failed Geolocation Cases. Conversely, even the best-performing LVLMs sometimes fail on images lacking distinctive markers. For example, GPT-4o struggles with a rainforest scene in Brazil, misclassifying it as a forest in the Philippines. This yields a GeoScore of 7.5, suggesting only continent-level accuracy. Similarly, LLaVA encounters difficulties localizing a sparse desert landscape in Qatar, mistakenly placing it in Egypt and earning a GeoScore of 735.8. These errors highlight the ongoing challenge of handling geographically generic images.

5.2 Insights and Findings

5.2.1 Model Sensitivity to Data Variations. Our results indicate that LVLMs are sensitive to the complexity and diversity of the input data. GPT-4o achieves a GeoScore of 4403.1 on the relatively constrained Im2GPS dataset but drops to 4340.2 on YFCC26k, a more diverse dataset. In contrast, GeoSpy maintains relatively stable performance—4570.8 on Im2GPS and 4451.6 on YFCC26k—likely reflecting its specialized tuning for geolocation tasks. LVLMs thus appear to excel when images contain clear landmarks or distinctive features but face challenges with more generic scenes.

Finding 3: LVLMs generally achieve higher accuracy in contexts featuring prominent landmarks or unique regional characteristics.

5.2.2 Adaptive Behaviors of LVLMs. Notably, LVLMs exhibit adaptability by adjusting their geolocation reasoning in response to newly presented data. This quality is crucial for real-world applications where environmental conditions can change rapidly. Our analysis reveals that the following factors significantly improve LVLM predictions:

- **Soil Types and Vegetation:** Regional differences in soil composition and plant life can guide the models toward more accurate estimates.
- **Cultural or Architectural Clues:** Culturally specific motifs, architectural designs, and public art provide strong cues for narrowing down potential locations.
- **Outdoor Settings:** Outdoor images often contain recognizable structures and natural landmarks, which can further enhance geolocation precision.

Finding 4: The inherent adaptability of LVLMs supports effective fine-tuning and deployment across diverse environmental contexts, making them well-suited for dynamic geolocation tasks.

6 Ethan: An Enhanced Framework

To overcome the limitations outlined in previous sections, we introduce Ethan, a framework that leverages LVLMs for automated geolocation. As shown in Figure 5, Ethan integrates two core components:

- fine-tuning LVLMs to more effectively process real-world images and extract key information, and
- an innovative chain-of-thought (CoT) prompting strategy [55] designed to mirror the problem-solving approaches of expert geoguessers.

This hybrid methodology significantly enhances the precision and practicality of geolocation predictions. Unlike traditional methods and naive LVLM strategies, Ethan does not simply ask an LVLM to identify a location. Instead, it exploits the LVLM's innate reasoning capabilities to deduce the location in a manner akin to human experts.

Seasoned players of geolocation games like GeoGuessr [1] commonly rely on recognizable environmental cues (e.g., vegetation, architecture, signage, vehicle types, and the sun's position) to make reasoned guesses about a location. As demonstrated by *Finding 4* in our empirical study, although LVLMs excel at detecting these features within images, they struggle to leverage these elements cohesively when predicting locations. To address this gap, Ethan reproduces human-style deductive reasoning via CoT prompting, guiding the LVLM through a structured, step-by-step analysis. By mirroring how human geoguessers systematically interpret visual cues, Ethan enables LVLMs to make accurate inferences from limited data while preserving the interpretability of the model's internal reasoning.

When given an image, Ethan prompts the LVLM to identify and interpret visible features relevant to geolocation. The model then compares these details against known geographic and cultural information, much like an experienced human geoguesser would match local architecture or vegetation to familiar regions. This systematic approach not only boosts location-prediction accuracy but also clarifies how the model arrives at its final decision. Below, we detail the design and implementation of Ethan to illustrate how these strategies coalesce effectively.

- **Step-by-step Reasoning:** Like a skilled GeoGuessr player, Ethan sequentially inspects each component within an image, referencing a comprehensive repository of geographic and cultural markers. This gradual process dissects complex scenes into manageable pieces of information—an essential element for precise geolocation.
- **Integration of Diverse Data Sources:** By fusing knowledge from satellite imagery, street-level photos, and cultural databases, Ethan expands the contextual landscape available for each analysis, thereby enhancing the LVLM's ability to link specific image features to particular regions.
- **Adaptive Learning:** Ethan adapts over time by learning from each geolocation attempt, refining its strategies and incorporating lessons from feedback—similar to human experts who continuously sharpen their skills with experience.

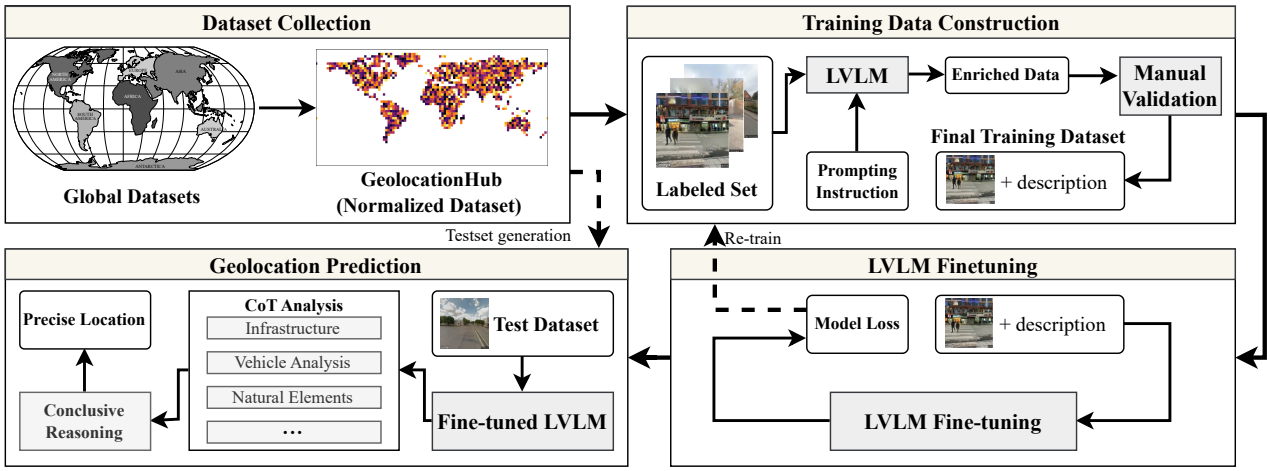

Figure 5: Workflow of Ethan, a framework that leverages fine-tuning to improve the performance of LVLM-based geolocation.

6.1 Fine-Tuning LVLMs for Geolocation

Our primary objective is to improve LVLMs' ability to recognize and interpret images for geolocation tasks. We accomplish this by fine-tuning LVLMs using GeolocationHub, a curated dataset containing explicit geolocation information (refer to Section 4.3). The fine-tuning pipeline begins with the generation of image-prompt pairs, produced by instructing GPT-4o to generate descriptive statements about each image based on the following prompt:

CoT Data Generation

You are the leading expert in geolocation research. You have been presented with an image, and your task is to determine its precise geolocation, specifically identifying the country it was taken in. To accomplish this, examine the image for broad geographic indicators such as architectural styles, natural landscapes, language on signs, and culturally distinctive elements. Narrow down the location by identifying regional characteristics such as specific flora and fauna, vehicle types, and road signs that may point to a region or subdivision within the country. Pay special attention to highly specific details in the image, such as unique landmarks, street names, or business names. For instance, if the location is {`address`}, with coordinates {`lat, lon`}, explain how these elements informed your conclusion by analyzing visual cues, cross-referencing known geographic data, and verifying your hypothesis with external resources.

We apply this prompt to each image in GeolocationHub, generating detailed textual descriptions for fine-tuning. Consistent with Section 4.3, we randomly inspect 1% of these generated labels to ensure accuracy, iterating until no erroneous annotations remain. These curated descriptions, paired with the true geolocation labels, form a comprehensive dataset for fine-tuning the LVLMs.

We employ the default settings of FastChat [62] for fine-tuning. We monitor the model's loss; if training fails to converge, we refine our dataset by adding more sample data and repeat the process. In our experiments, convergence typically occurs within three epochs, mitigating overfitting risks and delivering high-quality performance.

6.2 Geolocation Strategies

Drawing inspiration from human geoguessing tactics, as discussed in Section 2, human experts systematically evaluate an image's features (landscape, architecture, signage, etc.) to formulate a logical conclusion rather than guessing its location outright. This incremental reasoning aligns well with LVLM capabilities. Our design of Ethan's geolocation strategies focuses on categories of visual cues that are widely recognized as informative for geolocation and are commonly used by human geoguessers. We selected these specific cue categories because they offer a structured way to decompose the complex visual analysis task, allowing the LVLM to focus on distinct aspects of an image that, when combined, provide strong evidence for a particular location. For instance, infrastructure details often point to national or regional standards, natural elements are tied to climate and biome, vehicle characteristics can reflect local environmental conditions or economic status, and cultural indicators provide highly localized context. Consequently, we developed CoT prompts that guide LVLMs to systematically extract and reason about these specific types of information before producing a final geolocation estimate:

(1) **Infrastructure:** The model inspects road markings (e.g., color and pattern of lane lines), directional signs (e.g., language, script, shape, color-coding), utility poles, and license plate patterns (if legible and distinct). These elements often adhere to national or regional standards, providing strong, broad geographical constraints.
(2) **Natural Elements:** The model identifies soil types and coloration, prevalent vegetation patterns (e.g., specific tree species, crops, general flora), distinctive landscape features (e.g., mountains, coastlines, deserts), and even inferred climate indicators (e.g., presence of snow, arid conditions). These cues are fundamental to distinguishing between different biomes and climatic zones.

(3) **Vehicle Analysis:** The model notes vehicle types (e.g., common car models, trucks, motorcycles, agricultural vehicles), their condition, and any specialized attributes (e.g., snorkels in flood-prone or off-road areas, rust patterns suggesting coastal or humid climates, dirt patterns). This analysis can suggest local environmental factors, economic status, or regional preferences.
(4) **Cultural Indicators:** The model detects region-specific elements such as architectural styles (beyond general motifs, including specific building materials or construction techniques), unique business branding visible in shopfronts, clothing styles if people are prominent, repurposed objects (e.g., tires as planters), and public art or graffiti styles. These indicators often provide highly localized or culturally specific context.

By integrating these details with geographical data, Ethan enhances its geolocation precision. The model also adapts to variable data coverage—for instance, areas where Street View imagery primarily exists along major roadways.

6.3 Chain-of-Thought Prompting Integration

To provide a structured geolocation strategy, we employ a step-by-step CoT approach that guides the LVLM through the logic needed for accurate geolocation. This strategy helps the model perform systematic analysis of visual and contextual data, reducing errors often encountered in complex reasoning tasks.

When presented with an image, CoT prompting nudges the model to:

(1) Identify any distinctive features (e.g., "Generation 4 gray car with a snorkel").
(2) Note road characteristics, if present (e.g., "solid white outer lines with dashed yellow center lines").
(3) Assess the surrounding environment (e.g., "semi-arid terrain with light orange, sandy soil and mountainous vistas").
(4) Observe infrastructure or cultural elements (e.g., "green directional signs with white borders").
(5) Integrate these observations to zero in on the likely region and produce a final prediction.

This sequential reasoning not only boosts geolocation accuracy but also clarifies the model's decision process. By exposing intermediate reasoning steps, we can more easily identify weaknesses in the fine-tuning pipeline and target them for future improvements. This structured approach harnesses LVLMs' advanced capabilities to deliver high precision in diverse, real-world scenarios while preserving transparency and interpretability in the model's predictions.

7 Evaluation

This section presents an in-depth evaluation of Ethan [2] and addresses three core research questions: **(1)** To what extent can Ethan accurately predict locations from images? **(2)** Under which conditions does it fail? **(3)** How well does it perform in real-world scenarios? **(4)** How different settings affect the performance of Ethan? By systematically exploring these questions, we aim to highlight Ethan's strengths, uncover its limitations, and gauge its practical utility in real-world geolocation tasks.

Implementation Overview. We developed Ethan using approximately 1,138 lines of Python code, aligning with the design principles detailed in Section 6. The system seamlessly integrates fine-tuned LVLMs and a CoT prompting strategy. This integration is crucial for decomposing complex images into identifiable segments (e.g., architecture, vegetation, road signs) before synthesizing these observations into a coherent geolocation prediction.

To address **RQ1**, we use the dataset constructed in Section 6 for evaluation. For a fair assessment, Ethan is divided into a training set and a testing set, ensuring that the fine-tuning process does not affect the testing phase. Following the empirical study, we measure the performance of Ethan using Haversine Distance, GeoScore, and Administrative Boundary Scales, and compare it with four solutions from the previous study. For **RQ2**, we perform a detailed failure analysis by examining instances where Ethan failed to accurately predict geolocations. To address **RQ3**, we test Ethan on the *GeoGuessr* game [1], the most popular geolocation competition platform, where it competes against human players.

Methodology. To ensure a thorough and fair evaluation, our experimental methodology includes:

(1) *Training-Testing Split*: We divide the GeolocationHub dataset into 30,000 images for fine-tuning and 20,000 images for testing to prevent data leakage and ensure reliable performance metrics.
(2) *Baseline Comparisons*: We assess Ethan alongside StreetClip, GeoClip, GPT-4o, LLaVA, and GeoSpy, using the same 20,000 testing images. This setup results in 10 model configurations (Ethan plus each baseline in zero-shot, few-shot, or chain-of-thought modes, where applicable) multiplied by 20,000 test images, yielding 200,000 unique trials.
(3) *Performance Metrics*: We employ Haversine Distance, GeoScore, and administrative boundary thresholds (street, city, region, country, and continent) to capture both fine-grained and large-scale geolocation accuracy.
(4) *Real-world Competition*: To further validate Ethan's capabilities, we test it on the *GeoGuessr* [1] platform by pitting it against human players, providing insights into Ethan's robustness in dynamic and unpredictable environments.

7.1 RQ1 (Effectiveness)

Overall Performance on GeolocationHub. Table 4 compares Ethan with five baselines across multiple distance thresholds. Ethan consistently leads the pack, excelling in high-precision scenarios and broader geographical scales. For instance, Ethan achieves a 28.7% accuracy at the strict street-level threshold (1 km), surpassing GeoSpy by over 2% (26.5%). This advantage becomes even more pronounced at the city level (25 km), where Ethan attains 59.2% accuracy compared to GeoSpy's 51.1%. At the region level (200 km), Ethan maintains a robust 91.4% accuracy—outperforming all baselines by at least 5%. Its strong performance persists at country (95.6%) and continent (99.3%) scales.

Distance and Score Analysis. In addition to raw accuracy, we examine average Haversine distance and GeoScore. Ethan achieves an average distance of 499.3 km, an improvement of 46.5 km over GeoSpy (545.8 km). Although 499.3 km might sound relatively large, it is partly influenced by the global scope of the dataset, where missed predictions in extremely remote locations can inflate average

Technique		Distance (% @ km)					Avg Distance	Avg Geoscore
		Street (1 km)	City (25 km)	Region (200 km)	Country (750 km)	Continent (2,500 km)	(km)	(0-5000)
Pre-LVLM	StreetClip [24]	4.3	39.2	78.1	92.5	97.3	1225.6	3520.1
	GeoClip [10]	5.1	40.7	76.5	93.1	97.8	1215.4	3640.2
GPT-4o	Zero-shot	15.8	45.3	82.2	90.4	98.2	869.7	4205.3
	Few-shot	16.7	46.9	83.6	91.7	98.7	664.3	4298.7
	Chain-of-thought	18.2	47.6	84.1	92.2	99.1	159.9	4375.9
LLaVA	Zero-shot	10.4	42.5	80.4	87.6	96.3	1180.8	3751.8
	Few-shot	12.1	43.3	81.2	89.2	97.1	974.6	3812.6
	Chain-of-thought	14.3	44.7	82.3	90.3	98.0	869.5	3968.4
	GeoSpy [3]	26.5	51.1	85.7	93.9	99.1	545.8	4507.3
	Ethan	28.7	59.2	91.4	95.6	99.3	499.3	4620.9

Table 4: Geolocation evaluation results of Ethan vs. the benchmark solutions over the dataset.

values. Meanwhile, Ethan attains an average GeoScore of 4620.9, which is 2.5% higher than GeoSpy's 4507.3. Notably, when focusing on high-density areas like megacities, Ethan's average distance drops substantially due to the abundance of unique landmarks and contextual cues that facilitate accurate localization.

Qualitative Examples. Ethan's success often hinges on its ability to analyze subtle visual cues. For instance, in a sample image depicting a bustling New York City street, Ethan accurately placed the location within 500 m by identifying a combination of building styles, road markings, and recognizable storefronts—earning a perfect GeoScore of 5000. By contrast, StreetClip located the image 3.2 km away, and LLaVA missed by over 5 km, primarily due to difficulty interpreting overlapping urban elements.

Additionally, Ethan excelled in a rural Midwest setting where the visual scene showcased relatively uniform farmland. By scrutinizing soil coloration, architectural styles of barns, and regional vegetation, Ethan deduced a location within 10 km. Other models (including GPT-4o in zero-shot mode) struggled in this scenario, sometimes conflating the region with farmland in neighboring states, resulting in an average error of over 50 km.

Summary of RQ1 Findings

Ethan demonstrates significant performance gains over both traditional and LVLM-based baselines. Its superior accuracy at small distance thresholds underscores the effectiveness of its chain-of-thought prompting and fine-tuning approach, validating its design for precise location inference.

7.2 RQ2 (Failure Case Analysis)

Although Ethan shows compelling advantages, certain conditions can substantially degrade its performance. To explore these shortcomings, we examined misclassifications spanning from minor street-level inaccuracies to major continental-scale errors. This section details our findings and illustrates Ethan's vulnerabilities, many of which mirror those found in human geolocation errors.

Low-Visibility Conditions. Images captured in inclement weather (e.g., heavy fog, torrential rain) or taken at night significantly reduce the clarity of critical landmarks, road signs, and architectural details. In one case, a photograph from a fog-shrouded portion of San Francisco's Golden Gate Park led Ethan to misplace the scene in a rural area 100 km away. Manual inspection revealed that the essential cues—skyline, signage, and building outlines—were obscured, making it nearly impossible for Ethan to pinpoint the exact coordinates.

Minimal Landmark Environments. Another prevalent failure mode involves unremarkable rural landscapes and deserts lacking clear distinguishing features. A notable example arose with a desert in Nevada, which Ethan incorrectly tagged as a similarly barren region in Arizona, missing the correct location by over 150 km. Such errors primarily stem from the absence of topographical or man-made features—like road signs, recognizable buildings, or unique vegetation—that CoT prompting typically leverages for accurate predictions.

Homogeneous Urban Zones. Although Ethan generally excels in cities, repetitive urban layouts present challenges. Large housing districts or planned communities can look nearly identical across neighborhoods. In one instance, Ethan placed a Tokyo residential street in a nearby district 30 km away, primarily due to the identical building facades and minimal signage. While 30 km remains within a city-wide error margin, it underscores how uniform architectural styles can hinder fine-grained localization.

Rapidly Changing Scenes. Construction sites and newly developed areas often differ drastically from archived images used in training. In a test image featuring a newly built park in New York City, Ethan erred by 80 km because the model's training data predated the park's completion, leading it to rely on outdated environmental cues. This situation underscores the significance of continually updating training data to reflect evolving real-world conditions.

Condition	Distance (% @ km)					Avg Distance	Avg Geoscore
	Street (1 km)	City (25 km)	Region (200 km)	Country (750 km)	Continent (2,500 km)	(km)	(0-5000)
ETHAN on Street View Images (Subset)	28.1	58.5	90.8	95.1	99.0	505.2	4590
ETHAN on Non-Street View Images	14.7	39.2	76.1	88.5	96.5	980.4	3820

Table 5: Ablation Study: Impact of Input Image Source on ETHAN Performance. Results on a 2,000-image Street View subset vs. 1,000 diverse non-Street View outdoor images.

Prompt Variant	Distance (% @ km)					Avg Distance	Avg Geoscore
	Street (1 km)	City (25 km)	Region (200 km)	Country (750 km)	Continent (2,500 km)	(km)	(0-5000)
Zero-shot Prompt	11.5	34.8	77.9	89.1	97.0	1050.6	3680
Few-shot Prompt	20.3	49.1	86.0	92.5	98.3	720.1	4250
Full ETHAN CoT Prompt (Ours)	28.1	58.5	90.8	95.1	99.0	505.2	4590

Table 6: Ablation Study: Impact of Prompt Template Variants on ETHAN Performance (using the 2,000-image Street View subset).

Adversarial Modifications to Key Features. To evaluate ETHAN's robustness, we applied small adversarial perturbations via SGA [22], carefully altering crucial features (e.g., vegetation color, road designs, or building silhouettes) identified in ETHAN's chain-of-thought. We tested 2,500 images (500 from each continent) that ETHAN had correctly identified with street-level precision. After these adversarial changes, ETHAN's street-level accuracy plummeted by 74.3%, dropping below 10%. This stark decrease confirms ETHAN's heavy reliance on specific visual cues and emphasizes the need for adversarial defenses to preserve reliability.

Summary of RQ2 Findings

ETHAN is less effective in low-visibility settings, feature-scarce environments, and rapidly changing locales. Additionally, adversarial manipulations of CoT-relevant features can substantially undermine its geolocation abilities, signifying a clear avenue for future research on robust training and mitigation techniques.

7.3 RQ3 (Real-world Application)

Deployment on *GeoGuessr*. To validate ETHAN's utility in realistic scenarios, we tested it on the crowd-favorite geolocation game *GeoGuessr* [1]. By developing specialized wrappers, we automated the process of submitting ETHAN's predictions directly into the game's interface. We ran 41 rounds of ETHAN matches against randomly paired human competitors, capturing average scores, win rates, and distance metrics.

Comparison with Human Players. Table 7 shows ETHAN averaged a score of 4550.5, which exceeds human players' 4120.3, and dominated the win rate by over 70 percentage points (85.4% vs. 14.6%). ETHAN's best guess landed within 0.3 km of the true location—very close to the best human guess of 0.7 km—and its worst error stretched 5,258.2 km (compared to 5,443.5 km for humans). These metrics demonstrate ETHAN's impressive overall performance in a real-world-style setting, where images may be unpredictable or contain partial obstructions.

Competitor Type	ETHAN	Human Competitor
Average Score	4550.5	4120.3
Win Rate (%)	85.4	14.6
Closest Distance (km)	0.3	0.4
Farthest Distance (km)	5258.2	5443.5

Table 7: Average Performance of ETHAN versus Human Competitors in GeoGuessr over Multiple Rounds

Qualitative Observations in GeoGuessr. One of the most striking examples of ETHAN's effectiveness emerged in a remote Norwegian village test. Despite the town's uniform houses and mountainous terrain, ETHAN inferred regional details from subtle indicators like road signage typography, architectural roofing, and local vegetation. This analysis yielded a final guess only 2 km from the real spot. Human competitors averaged 5 km off, reflecting ETHAN's aptitude for synthesizing nuanced or less obvious signals. In another successful case, a highly urbanized district in Tokyo was nailed within 300 m by ETHAN, thanks to consistent scanning of multi-level architecture, signage languages, and road geometry. By contrast, many human players guessed roughly 1.5 km away, possibly due to the fast-paced nature of the game and the complexity of distinguishing among visually similar Tokyo wards.

Still, ETHAN stumbled on feature-sparse scenes—a deficiency shared by humans. For instance, a generic Australian beach was misidentified by 250 km, and a desert region in Nevada was incorrectly placed in Mongolia (over 5,200 km off). These errors highlight the ongoing need for better training coverage of visually homogeneous locales and further reinforcement of the chain-of-thought pipeline to handle ambiguous or minimal cues.

Summary of RQ3 Findings

ETHAN consistently demonstrates strong performance in competitive, real-world-style scenarios like GeoGuessr, frequently outperforming human opponents. However, feature-poor environments continue to present challenges for both ETHAN and human players, emphasizing the need for more sophisticated feature extraction and training-data diversification.

Overall Implications. The GeoGuessr evaluation underscores ETHAN's adaptability and reliability when confronted with diverse global imagery, sporadic clue availability, and time constraints that mirror real-world geolocation challenges. Such capabilities hold promise not just for entertainment-oriented platforms but also for high-stakes applications, including surveillance, border security, and disaster response, where timely and precise geolocation can be paramount. Yet, as identified in RQ2, caution is necessary when dealing with adversarial images or visually unremarkable settings, revealing avenues for further optimization.

7.4 RQ4 (Ablation Study)

We conducted ablation studies to understand the contributions of input image source (Street View vs. non-Street View) and our Chain-of-Thought (CoT) prompting formulation within the ETHAN framework. These used a random 2,000-image subset from the GEOLOCATIONHUB test split.

7.4.1 Impact of Input Image Source. ETHAN, fine-tuned on Street View images from GEOLOCATIONHUB, was tested against 1,000 diverse non-Street View outdoor images from public sources. Table 5 presents the results.

Table 5 shows reduced accuracy for ETHAN on non-Street View images (e.g., doubled average distance error, lower GeoScore), indicating specialization to Street View's visual characteristics. Future work should focus on enhancing generalization to diverse image sources, possibly via broader training data or domain adaptation.

7.4.2 Impact of Prompt Template Variants. We compared ETHAN's structured CoT prompt against two simpler variants (detailed in Appendix) on the 2,000-image Street View test subset using the fine-tuned ETHAN model. Table 6 shows the comparison.

Results in Table 6 confirm the value of ETHAN's detailed CoT. The Zero-shot prompt performed worst. The Few-shot prompt improved this, but our Full ETHAN CoT prompt, guiding analysis of distinct cue categories, achieved substantially better results. This shows that the CoT prompt's specific structure is critical for effective geolocation.

Summary of RQ4 Findings

These ablations show that both input data characteristics and prompt design significantly influence ETHAN's performance.

8 Limitations

Dataset Coverage and Geographical Bias. Although we aim for geographic balance in GEOLOCATIONHUB (Appendix A.2), it inherits the biases of Google Maps Street View, which overrepresents North America and Europe while underrepresenting regions like Africa and parts of Asia. This skew, especially toward urban areas, may limit generalization to poorly covered regions. Future work should incorporate diverse sources such as community-contributed images and satellite data to improve representativeness.

Sensitivity to Visual Features and Environmental Conditions. Geolocation accuracy in ETHAN and similar LVLMs is affected by image quality and environmental conditions. Adverse weather, poor lighting, and seasonal variations (e.g., snow or shadows) can obscure key features and introduce noise. Since image capture conditions in GEOLOCATIONHUB are uncontrolled, such factors may bias the model. Moreover, models can latch onto transient cues like vehicles or advertisements; although CoT prompting encourages attention to stable landmarks, sensitivity to dynamic content remains.

Generalization Concerns. While ETHAN improves geolocation accuracy, it may overfit to Street View imagery, limiting out-of-distribution (OOD) performance on other domains such as artistic, historical, or aerial views. Additionally, it remains vulnerable to adversarial image modifications—posing risks in sensitive contexts—and its reasoning failures or hallucinations, even with CoT prompting, are not yet well understood. Enhancing robustness and interpretability are important directions for future work.

9 Conclusion

This paper explored LVLMs for image geolocation, revealing a non-negligible, albeit currently low-accuracy, zero-shot capability that signals an emerging privacy threat. Our ETHAN framework, using CoT reasoning and fine-tuning, improved accuracy (e.g., 85.4% GeoGuessr win rate), highlighting potential system evolution rather than current precise reliability. The study also exposed LVLM limitations, including sensitivity to data variations, landmark reliance, and vulnerability in feature-scarce or adversarial conditions. The core finding is that foundational LVLM geolocation capabilities exist, necessitating proactive privacy measures. We thus issue a strong call for mitigation research, focusing on defenses, privacy-preserving LVLM architectures, and ethical guidelines. Understanding LVLM capabilities and limitations in geolocation is vital for developing effective safeguards and managing societal implications as technology advances.

Acknowledgments

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority. We would also like to thank Quantstamp for providing a supportive and open research atmosphere that encourages the exploration of new research directions.

References

[1] GeoGuessr. https://www.geoguessr.com/, . Accessed: 2023-10-10.

[2] Toxicdetector. https://sites.google.com/view/geolocation-privacy/home, . (Accessed on 07/11/2024).

[3] geospy.ai. https://geospy.ai/, . (Accessed on 07/11/2024).

[4] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016. doi: 10.1145/2976749.2978318. URL http://dx.doi.org/10.1145/2976749.2978318.

[5] Paul Alvarez. Using extended file information (exif) file headers in digital evidence analysis. *International Journal of Digital Evidence*, 2(3):1–5, 2004.

[6] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, et al. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21967–21977, 2024.

[7] Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, November 2022. doi: 10.1109/icdmw58026.2022.00078. URL http://dx.doi.org/10.1109/ICDMW58026.2022.00078.

[8] Geoff Brumfiel. Artificial intelligence can find your location in photos, worrying privacy experts, Dec 2023. URL https://www.npr.org/2023/12/19/1219984002/artificial-intelligence-can-find-your-location-in-photos-worrying-privacy-expert.

[9] Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11. https://doi.org/10.7927/H49C6VHW, 2018. Accessed: May 15, 2025.

[10] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization, 2023. URL https://arxiv.org/abs/2309.16020.

[11] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[12] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*, 2023.

[13] Hanning Chen, Wenjun Huang, Yang Ni, Sanggeon Yun, Fei Wen, Hugo Latapie, and Mohsen Imani. Taskclip: Extend large vision-language model for task oriented object detection. *arXiv preprint arXiv:2403.08108*, 2024.

[14] Yi Chen, Gang Qian, Kiran Gunda, Himaanshu Gupta, and Khurram Shafique. Camera geolocation from mountain images. In *2015 18th International Conference on Information Fusion (Fusion)*, pages 1587–1596. IEEE, 2015.

[15] J De Curtò, I De Zarza, and Carlos T Calafate. Semantic scene understanding with large language models on unmanned aerial vehicles. *Drones*, 7(2):114, 2023.

[16] Goran M Djuknic and Robert E Richton. Geolocation and assisted gps. *Computer*, 34(2):123–125, 2001.

[17] European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union, L 119/1, April 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[18] European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 may 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2013/36/eu, 2014/53/eu, 2016/797/eu, 2016/798/eu and (eu) 2020/1828 (artificial intelligence act). Official Journal of the European Union, L/2024/1689, 6 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[19] Grace Exploration, Henry Generalization, and Irene Understanding. Probing zero-shot cross-domain generalization in lvlm-based image geolocation, 2024.

[20] Foursquare, May 2024. URL https://foursquare.com/.

[21] Alice Future, Bob Vision, and Charles Language. Enhancing lvlm geolocation on culturally diverse visual data through advanced fine-tuning, 2024.

[22] Sensen Gao, Xiaojun Jia, Xuhong Ren, Ivor Tsang, and Qing Guo. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory, 2024. URL https://arxiv.org/abs/2403.12445.

[23] GPT-4V. https://openai.com/research/gpt-4v-system-card.

[24] Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocalization, 2023.

[25] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12893–12902, 2024.

[26] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *CVPR*, 2008.

[27] James Hays and Alexei A Efros. Large-scale image geolocalization. *Multimodal location estimation of videos and images*, 2015.

[28] Jonghu Jeong, Minyong Cho, Philipp Benz, Jinwoo Hwang, Jeewook Kim, Seungkwan Lee, and Tae hoon Kim. Privacy safe representation learning via frequency filtering encoder, 2022. URL https://arxiv.org/abs/2208.02482.

[29] Michael William Jones. *Image geolocation through heirarchical classification and dictionary-based recognition.* University of Maryland, College Park, 2012.

[30] Evangelos Kalogerakis, Olga Vesselova, James Hays, Alexei A Efros, and Aaron Hertzmann. Image sequence geolocation with human travel priors. In *2009 IEEE 12th international conference on computer vision*, pages 253–260. IEEE, 2009.

[31] Elad Kravi, Yaron Kanza, Benny Kimelfeld, and Roi Reichart. Location classification based on tweets. GEOAI '21, page 51–60, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450391207. doi: 10.1145/3486635.3491075. URL https://doi.org/10.1145/3486635.3491075.

[32] Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. Watch your language: large language models and content moderation. *arXiv preprint arXiv:2309.14517*, 2023.

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022. URL https://arxiv.org/abs/2201.12086.

[34] Bo Liu, Quan Yuan, Gao Cong, and Dong Xu. Where your photo is taken: Geolocation prediction for social images. *Journal of the Association for Information Science and Technology*, 65(6):1232–1243, 2014.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

[36] Grace Luo, Giscard Biamby, Trevor Darrell, Daniel Fried, and Anna Rohrbach. g^3: Geolocation via guidebook grounding. *Findings of EMNLP*, 2022.

[37] Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. Granular privacy control for geolocation with vision language models. *arXiv preprint arXiv:2407.04952*, 2024.

[38] Meta. Facebook, 2024. URL https://www.facebook.com/.

[39] Meta. Instagram, 2024. URL https://www.instagram.com/.

[40] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. The StreetLearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019.

[41] Arsalan Mousavian and Jana Kosecka. Semantic image based geolocation given a map. *arXiv preprint arXiv:1609.00278*, 2016.

[42] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *ECCV*.

[43] David NextGen, Eve Context, and Frank Fusion. Iterative multi-image prompting for robust lvlm-based geolocation in ambiguous scenarios, 2025.

[44] OpenAI, . URL https://chatgpt.com/g/g-wx1eLTbGL-image-locator.

[45] OpenAI, . URL https://platform.openai.com/docs/guides/prompt-engineering.

[46] Yuval Shavitt and Noa Zilberman. A geolocation databases study. *IEEE Journal on Selected Areas in Communications*, 29(10):2044–2056, 2011.

[47] State of California. California privacy rights act of 2020 (cpra), 2020. Approved by voters in Proposition 24 on November 3, 2020. Amends the California Consumer Privacy Act (CCPA). Text available at https://cppa.ca.gov/regulations/pdf/cppa_regs.pdf (Accessed: May 14, 2025).

[48] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable semantic photo geolocation. In *WACV*, 2022.

[49] Catherine Tony, Nicolás E. Díaz Ferreyra, Markus Mutas, Salem Dhiff, and Riccardo Scandariato. Prompting techniques for secure code generation: A systematic investigation, 2024. URL https://arxiv.org/abs/2407.07064.

[50] Nam Vo, Nathan Jacobs, and James Hays. Revisiting IMG2GPS in the deep learning era. In *ICCV*, 2017.

[51] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *Meta-Radiology*, page 100047, 2023.

[52] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, 2023. URL https://arxiv.org/abs/2305.04091.

[53] Albatool Wazzan, Stephen MacNeil, and Richard Souvenir. Comparing traditional and llm-based search for image geolocation. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pages 291–302, 2024.

[54] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837, 2022.

[55] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[56] Edy Winarno, Wiwien Hadikurniawati, and Rendy Nusa Rosso. Location based service for presence system using haversine method. In *2017 international conference on innovative and creative information technology (ICITech)*, pages 1–4. IEEE, 2017.

[57] Keita Yaegashi and Keiji Yanai. Geotagged image recognition by combining three different kinds of geolocation features. In Ron Kimmel, Reinhard Klette, and Akihiro Sugimoto, editors, *Computer Vision – ACCV 2010*, pages 360–373, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19309-5.

[58] Jinghan Yang, Ayan Chakrabarti, and Yevgeniy Vorobeychik. Protecting geolocation privacy of photo collections. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 524–531, 2020.

[59] Da Yu, Peter Kairouz, Sewoong Oh, and Zheng Xu. Privacy-preserving instructions for aligning large language models, 2024. URL https://arxiv.org/abs/2402.13659.

[60] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[61] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[62] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[63] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[64] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

[65] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Geographic Distribution

A.1 Sampling Strategy

To achieve balanced sampling across geographic locations, we implemented a country-based sampling method to build GeolocationHub. We used the area of each country as a weight to determine the number of data points for each country. A weighted sampling approach was then applied to allocate and sample the data points accordingly.

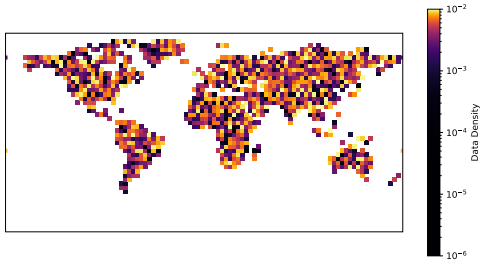

Figure 6: Geographic distribution of data density from our dataset.

A.2 GeolocationHub Data Distribution and Bias Quantification

Figure 6 illustrates the global data point density of our GeolocationHub dataset. As mentioned in Section 5, a key goal during curation was to achieve a more geographically balanced representation than often found in typical web-scraped image datasets, which tend to over-represent populous regions of North America and Europe. Our weighted sampling strategy by country area aimed to mitigate this.

A.2.1 Continental Distribution. While perfect uniformity is challenging due to variations in available Street View coverage, the GeolocationHub dataset achieves a more balanced continental distribution. The approximate breakdown of the 50,000 images by continent is as follows:

- North America: 20% (≈ 10,000 images)
- Europe: 22% (≈ 11,000 images)
- Asia: 25% (≈ 12,500 images)
- South America: 15% (≈ 7,500 images)
- Africa: 10% (≈ 5,000 images)
- Oceania: 8% (≈ 4,000 images)

This distribution reflects our effort to increase representation from typically underrepresented continents like Africa and South America, guided by the area-weighted sampling. A visual representation of this distribution, such as a bar chart, could further illustrate this balance (not included in this paper for brevity, but derivable from the dataset).

A.2.2 Urban versus Rural Representation. Quantifying a precise global urban/rural split is complex due to varying definitions. However, we endeavored to include a significant proportion of images from non-urban settings, as traditional datasets are often heavily skewed towards cities. Based on an automated heuristic using population density data associated with the image coordinates (e.g., coordinates falling within areas with <150 persons/km^2 classified as rural/low-density suburban), we estimate the following approximate split for GEOLOCATIONHUB:

- Urban / Dense Suburban: 65% (≈ 32,500 images)
- Rural / Low-Density Suburban / Natural Landscapes: 35% (≈ 17,500 images)

This ~1/3 representation for non-urban scenes is a deliberate effort to provide data for evaluating geolocation models in more challenging, less feature-dense environments. The specific criteria for urban/rural classification involved using publicly available gridded population data (e.g., GPWv4 [9]) and applying a threshold to the cell corresponding to the image's coordinates. It is important to note that this is an approximation, as Street View coverage itself is not uniformly distributed across all urban and rural areas globally.

Despite these efforts, some geographical bias inevitably remains due to the inherent biases in the source data (Google Maps Street View coverage) and the practical limitations of global-scale data collection. This residual bias is discussed further in Section 8.

B Dataset Curation for GeolocationHub

The GeolocationHub dataset, comprising 50,000 high-quality images suitable for geolocation tasks, was meticulously constructed using a bottom-up, iterative sampling and filtering approach designed to achieve geographic balance. This appendix details this process, which complements the overview in Section 6.

B.1 Sampling Strategy and Image Sourcing

The dataset construction process began with the goal of achieving a geographically balanced collection of images. The core principles were:

- **Primary Source:** All images were sourced from Google Maps Street View.
- **Weighted Sampling by Country Area:** To ensure geographic diversity and avoid over-representation of densely photographed areas, we implemented a country-based weighted sampling strategy. The land area of each country was used as a weight to determine the target number of data points (i.e., high-quality images) to collect for that country.
- **Iterative Quota Fulfillment:** For each country, images were iteratively sampled from random geographic coordinates within its boundaries via Street View. Each sampled image was then subjected to a quality control process (detailed below). If an image passed the quality control, it was added to the dataset, contributing to that country's quota. This process continued until the weighted quota for each country was met, or until a reasonable effort to find suitable images was exhausted for sparsely covered regions, ultimately culminating in the 50,000-image dataset.

B.2 Iterative Quality Control with GPT-4o and Keyword Filtering

Each image sampled from Google Maps Street View underwent the following quality control steps:

(1) **Image Description Generation:** The sampled Street View image was provided to GPT-4o, which used a dedicated prompt (see Appendix B.4) to generate a concise, objective description. This description focused on scene type, key elements, image quality, and presence of geolocation cues.
(2) **Automated Filtering based on Description:** The GPT-4o generated description was then automatically processed:
   - **Indoor/Obstructed View Filtering:** Keywords indicative of indoor scenes (e.g., "room," "inside store," "vehicle interior"), heavily obstructed views (e.g., "blurry dashboard," "obstructed by windshield"), or otherwise unsuitable content (e.g., "underwater," "inside tunnel with no exit visible") were used to automatically discard the image. The list of keywords was refined throughout the process. Common keywords included:
     `indoor, room, interior, inside, vehicle interior, dashboard, windshield view (obstructed), blurry, underexposed, overexposed, tunnel (no exit), underwater, ceiling, floor, close-up texture`
   - **Low-Context Filtering:** Descriptions suggesting a lack of useful geolocation cues (e.g., "sky only," "close-up of indistinct ground," "featureless wall," "generic foliage without landmarks") were also used to discard images.
(3) **Manual Spot-Checking:** A subset (1,000) of images and their GPT-4o descriptions, especially those borderline or from rarely sampled regions, underwent manual spot-checking to ensure the accuracy of the filtering process and to refine keywords or description guidelines if necessary.

If an image failed these quality control checks, it was discarded, and the sampling process for that specific country/region would continue by selecting new random coordinates until a suitable image was found or the regional sampling attempt limit was reached. This iterative "sample -> describe -> filter -> accept/reject & resample if needed" loop was key to building the dataset one quality image at a time while adhering to the geographic distribution targets.

B.3 Discarded Samples

Given the bottom-up approach, "discarded samples" refers to images that were retrieved from Street View but failed the quality control described above, leading to a re-sampling attempt for that particular geographic quota slot. It is estimated that for every 2-3 image accepted into the final dataset, approximately one image was sampled and subsequently discarded due to failing the quality control (e.g., being indoor, blurry, too dark/bright, or lacking sufficient visual cues as determined by the GPT-4o description and keyword filters). This iterative process ensured that the final 50,000 images in GeolocationHub were of high utility for the geolocation task.

B.4 GPT-4o Image Description Prompt for Iterative Dataset Curation

The following prompt was used with GPT-4o to generate descriptions for images sampled one-by-one from Google Maps Street View during the iterative construction of the GeolocationHub dataset. These descriptions were crucial for the quality control step, determining if an image was suitable for inclusion based on its content and clarity.

Classifier:

You're an AI helping decide if an image should be ACCEPTED into a geolocation dataset. For each image, give a very brief, factual description covering:
Scene: Outdoors / Indoors / Vehicle interior
Key elements: Main objects (e.g. buildings, roads, trees, signs—legible or not)
Quality: Sharpness, lighting, any obstructions
Distinctive features: Noticeable landmarks or generic setting
Geolocation cues: Enough visual hints to attempt geolocation, or too vague/blurry
End with whether it's ACCEPT or REJECT.

C Manual Dataset Spot-Checking: Protocol and Reliability

To ensure the quality and suitability of images within the GeolocationHub dataset, a manual spot-checking process was implemented for a subset of images curated through the semi-automated pipeline (described in Appendix B). This section details the annotation protocol, including annotator selection and qualification, and reports the inter-rater reliability for this task.

C.1 Annotator Selection, Qualification, and Training

A team of three human annotators was enlisted for the dataset spot-checking task. The criteria and process for their selection and training were as follows:

- **Recruitment:** Annotators were undergraduate students in computer science at our institution, recruited based on their interest in AI research and data quality assessment.
- **Qualification Criteria:**
  - Demonstrated high attention to detail.
  - Proficiency in understanding and applying complex guidelines.
  - Basic familiarity with image analysis and geographic concepts.
- **Training Protocol:**

(1) **Guideline Review:** Annotators were provided with a comprehensive guideline document (approx. 5 pages). This document detailed the objectives of the GeolocationHub dataset, precise definitions of suitable images (clear, outdoor scenes with discernible features for potential geolocation) versus unsuitable images (e.g., indoor, blurry, heavily obstructed, featureless, or private/sensitive content inadvertently captured). It included numerous visual examples for each category.
(2) **Interactive Training Session:** A 2-hour training session was conducted by senior researchers. This session covered:
   - A walkthrough of the annotation interface.
   - In-depth discussion of the image suitability criteria and rejection reasons.
   - Clarification of borderline cases and common pitfalls identified from preliminary automated filtering.

  - Instruction on how to evaluate the relevance and accuracy of GPT-4o-generated image descriptions in the context of filtering decisions.

(3) **Pilot Annotation Task:** Before commencing the main task, each annotator independently labeled a pilot set of 50 diverse images. These images were pre-annotated by the research team to serve as a gold standard.
(4) **Performance Review and Finalization:** Annotators' performance on the pilot task was reviewed. Feedback was provided, and any misunderstandings of the guidelines were addressed. All three annotators demonstrated a high level of concordance with the gold-standard labels (e.g., >90% accuracy on the pilot set) and were subsequently confirmed for the main spot-checking task.

- **Compensation:** Annotators were compensated for both their training time and the subsequent annotation work.

C.2 Spot-Checking Task Protocol

The manual spot-checking was performed on 1,000 images that were either flagged as borderline by the automated system (based on GPT-4o descriptions and keyword filters) or sampled from geographic regions with less dense initial data to ensure quality across diverse areas.

(1) **Independent Double Annotation:** Each of the 1,000 images, along with its corresponding GPT-4o generated description, was assigned to two of the three trained annotators for independent evaluation. This ensured each image was assessed by two individuals without consultation.
(2) **Annotation Task Definition:** For each image, annotators were required to:
   - Make a primary judgment: "Accept" (image is suitable for the dataset) or "Reject" (image is unsuitable).
   - If "Reject," select the primary reason(s) from a predefined checklist: Indoor Scene, Vehicle Interior/Obstructed by Vehicle, Poor Image Quality (e.g., Blurry, Over/Under Exposed), Low Context/Featureless (e.g., Sky/Ground only), Ambiguous/Abstract Content, or Other (with a field for brief mandatory explanation).
   - Optionally, provide brief comments if the GPT-4o description was significantly misleading or if the image presented a particularly challenging edge case.
(3) **Disagreement Resolution:** Cases where the two annotators provided conflicting "Accept/Reject" labels were identified. These (approximately 8% of the spot-checked images) were then adjudicated by a senior member of the research team (one of the authors), who made the final decision. These instances also served as feedback to refine annotator understanding or the guidelines if systemic issues were noted.

C.3 Inter-Rater Reliability (IRR)

To quantify the consistency of the manual spot-checking judgments, inter-rater reliability was calculated. Before starting the main spot-checking of the 1,000 images, a separate calibration set of 200 diverse images (not part of the 1,000) was independently annotated by all three annotators. The primary task for IRR calculation was the binary decision of "Accept" or "Reject."

- **Cohen's Kappa (κ):** Pairwise Cohen's Kappa was computed for each of the three annotator pairs (A1-A2, A1-A3, A2-A3). The average pairwise κ was 0.82 (individual values: 0.80, 0.83, 0.83), indicating strong agreement.
- **Krippendorff's Alpha (α):** To assess overall agreement among the three annotators for the nominal "Accept/Reject" labels, Krippendorff's α was calculated. The resulting $\alpha = 0.81$ also demonstrated good reliability among the annotators.

These IRR scores provide confidence in the consistency and reliability of the judgments made during the manual spot-checking phase of the GeolocationHub dataset curation.

D Computational Resources

The development, fine-tuning, and evaluation of the Ethan framework were conducted on a high-performance computing cluster. Specifically, the fine-tuning of the underlying LVLM and the extensive experimental evaluations presented in this paper utilized a setup consisting of 8 NVIDIA A100 GPUs, each equipped with 80GB of HBM2e memory.

As Ethan is primarily a fine-tuning method applied to pre-existing LVLMs, its parameter count and GPU memory footprint during inference are largely comparable to those of the base LVLM it is built upon. The fine-tuning process itself, while computationally intensive, aligns with standard practices for adapting large-scale models. For instance, a typical fine-tuning run for Ethan on our 50,000-image GeolocationHub dataset (30,000 for training) as described in Section 4.3, generally converged within three epochs and required approximately 4 to 6 hours on our 8x A100 setup (exact time can vary based on batch size and other hyperparameters).

In terms of comparative resource use:

- **Compared to other LVLMs:** The inference cost of Ethan (once fine-tuned) is similar to other LVLMs of comparable size (e.g., the base model it was fine-tuned from, or models like GPT-4V, LLaVA when performing analogous visual analysis tasks). The primary additional overhead comes from the CoT prompting strategy, which may lead to longer generation sequences and thus slightly increased inference latency per query compared to a direct zero-shot query to a generic LVLM.
- **Compared to traditional geolocation solutions:** As Ethan is LVLM-based, its computational requirements, particularly for GPU memory and processing power, are significantly higher than most traditional, non-deep-learning geolocation methods (e.g., image retrieval based on SIFT features or simpler classification models discussed in Section 2). These traditional methods often run efficiently on CPUs and require substantially less memory. However, the trade-off is that LVLMs like Ethan can leverage much richer visual understanding and reasoning capabilities, as demonstrated by our results.

The decision to use a powerful GPU setup was driven by the need to efficiently process large datasets and iterate on model fine-tuning for a comprehensive evaluation of Ethan's capabilities.

E Evaluation Prompts

In this section, we present the prompts used to evaluate various geolocation methods. In particular, we first present the CoT prompt used to conduct geolocation task.

Chain-of-thought:

As the world's elite geolocation expert, your mission is to analyze the attached image and navigate through your reasoning to pinpoint the exact geolocation. Detail your analytical process and finalize the latitude and longitude with utmost accuracy.

Followed by this, we present the prompt used with few-shot strategy. As introduced in Section 5, we combine a series of prompting techniques, including emotional prompts, template inputs and outputs, and accurate task decomposition.

Few-shot:

Leveraging your expertise in geolocation, your task is to analyze the provided image and deduce its precise location. Accuracy in determining latitude and longitude is paramount.
Example 1: Image Description: A sandy beach with a notable rock formation in the background under a clear sky with scattered clouds.
Geolocation Process:

- The distinctive rock formation closely resembles the renowned Twelve Apostles in Victoria, Australia.
- The combination of the sandy beach and clear skies supports its identification along the southern Australian coast.
- Verification with existing images and maps confirms the location as near the Great Ocean Road.

Latitude and Longitude: -38.6633, 143.1051
Example 2: Image Description: A historic building featuring a large clock tower and gothic architecture, surrounded by red double-decker buses.
Geolocation Process:

- The gothic architecture and prominent clock tower suggest the Elizabeth Tower in London, United Kingdom.
- The presence of red double-decker buses confirms the urban setting as London.
- Comparisons with images of Big Ben and the adjacent area confirm the precise location.

Latitude and Longitude: 51.5007, -0.1246

Zero-shot Prompt

You are recognized as the world's foremost expert in geolocation analysis. Your objective is to meticulously analyze the provided image and determine its latitude and longitude.

F GeoGuessr-Style Evaluation Procedure

To assess Ethan's performance in a dynamic, real-world-style geolocation challenge, we conducted an evaluation using the popular online game GeoGuessr [1] (as discussed in Section 7). This appendix details the procedure.

F.1 Evaluation Setup

- **Platform Interaction:** GeoGuessr presents players with a series of interactive Street View panoramas, and the goal is to pinpoint the location on a world map. To evaluate Ethan, we developed a set of automated scripts (wrappers) to interface with the GeoGuessr platform.
- **Image Extraction:** For each GeoGuessr round, our scripts would capture a static, representative view (or multiple views, subsequently stitched or selected) from the initial Street View panorama presented by the game. Care was taken to select views that a human player might typically focus on, avoiding excessive sky or ground if possible, and aiming for a field of view that captures potential cues.
- **Ethan Prediction:** The extracted image(s) were then fed as input to the fine-tuned Ethan model. Ethan would process the image and output its predicted latitude and longitude coordinates based on its CoT reasoning.
- **Submission and Scoring:** The predicted coordinates from Ethan were programmatically submitted back to the GeoGuessr game interface via our wrapper scripts. The game then calculated the distance error and awarded a score (0-5000 points) based on its internal scoring algorithm, which is inversely proportional to the distance error.

F.2 Experimental Design

- **Game Rounds:** A total of 41 independent GeoGuessr game rounds were played by Ethan. Each game typically consists of 5 locations, resulting in $41 \times 5 = 205$ distinct locations being evaluated.
- **Map Selection:** To ensure diversity and prevent overfitting to specific map types, a variety of official GeoGuessr maps were used, specifically, the "World" map.
- **Human Competitors:** The performance of Ethan was compared against human players who were playing the same game rounds simultaneously (if the GeoGuessr challenge link feature was used).
- **Metrics Recorded:** For each location guess by Ethan and human players, we recorded the guessed coordinates, the true coordinates (provided by GeoGuessr after the guess), the distance error, and the score awarded by GeoGuessr.

G Reproducibility

To support Ethan and facilitate further research, we have developed an open-source data processing and analysis pipeline, the code for which is available on our project website [2].

H Discussion of Potential Defenses

Despite LVLMs' geolocation capabilities, their potency poses privacy threats. This section analyzes defense mechanisms to mitigate these risks by obscuring sensitive location data while preserving image analysis benefits.

H.1 Developing Privacy-Preserving LVLMs

A forward-looking approach involves incorporating privacy methods into LVLM design and training. LVLMs naturally extract detailed visual cues; privacy-preserving versions would counter this.

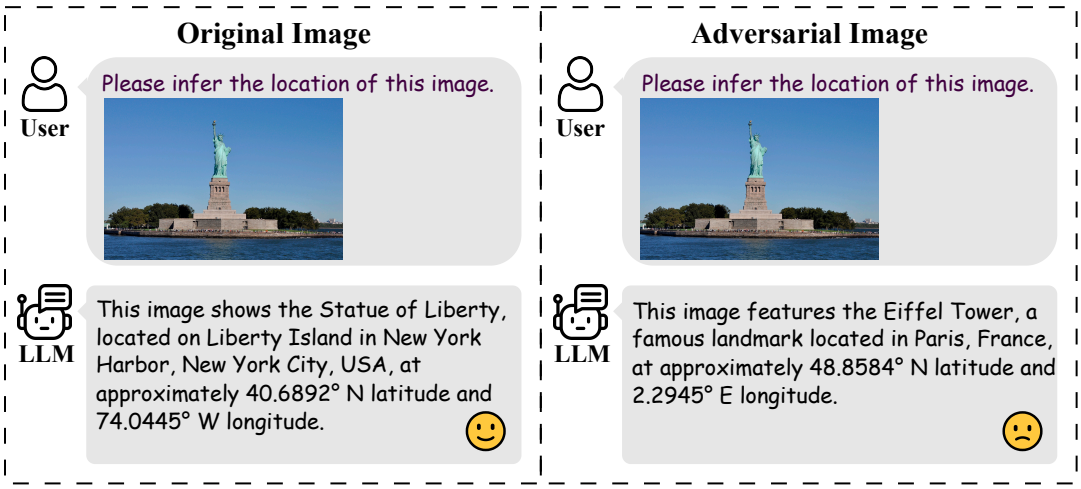

Figure 7: Adversarial images mislead the LVLM into recognizing the Statue of Liberty as the Eiffel Tower on LLaVA.

Selective Feature Suppression. During training, models' sensitivity to geolocation-revealing cues (e.g., landmarks, specific text) can be reduced by constraining feature extraction layers, for instance, by adjusting attention weights or using specialized loss functions. This makes LVLMs less likely to memorize exact geographic details while retaining broader understanding.

Differential Privacy Mechanisms. Differential privacy (DP) [4, 7] introduces noise into training data or gradient updates, statistically preventing the model from recalling features unique to single examples. Applied to LVLMs, DP minimizes information leakage about specific images, discouraging unauthorized geolocation. Though a potential accuracy trade-off exists, it can be managed by tuning noise parameters.

H.2 Implementing Real-Time Privacy Filters

Even with privacy-focused training, sensitive information can leak. An additional defense layer can be implemented at image upload, especially on social media.

Automatic Image Sanitization. Real-time filters [28, 59] integrated into upload pipelines (e.g., Instagram, Facebook) could analyze images for regionally unique identifiers (distinct building facades, statues, non-native text) and transform or blur them. User toggles could control these modifications.

User Alerts and Recommendations. Filters could also provide alerts for high-risk content (e.g., *"This image contains identifiable geographic features. Proceed?"*). Such guidance can foster user awareness and encourage cautious sharing.

H.3 Adversarial Image Modification

Another defense uses adversarial perturbations—subtle pixel alterations degrading geolocation predictions without significant visual quality loss (Figure 7). Small, often imperceptible noise can cause LVLMs to misinterpret landmarks.

Technique Spotlight: SGA.. We use SGA [22], which introduces mild, gradient-guided perturbations. It manipulates crucial *chain-of-thought* (CoT) features (signage, vehicles, architecture) to make the LVLM perceive a generic environment (e.g., Statue of Liberty misinterpreted as Eiffel Tower by LLaVA [11]). Our tests show SGA significantly reduces country-level accuracy (78.6% to 3.4%), indicating high efficacy.

Integration at Scale. Practical deployment must address computational overhead. Real-time perturbation generation can be demanding. However, GPU optimization or cloud-based processing might mitigate costs, potentially enabling large-scale adoption (e.g., browser plugins, smartphone features) for robust user location privacy.