

# NAVIG: Natural Language-guided Analysis with Vision Language Models for Image Geo-localization

Anonymous ACL submission

## Abstract

Image geo-localization is the task of predicting the specific location of an image and requires complex reasoning across visual, geographical, and cultural contexts. While prior Vision Language Models (VLMs) have the best accuracy at this task, there is a dearth of high-quality datasets and models for analytical reasoning. We first create NAVICLUES, a high-quality dataset derived from GeoGuessr, a popular geography game, to supply examples of expert reasoning from language. Using this dataset, we present NAVIG, a comprehensive image geo-localization framework integrating global and fine-grained image information. By reasoning with language, NAVIG reduces the average distance error by 14% compared to previous state-of-the-art models while requiring fewer than 1000 training samples. Our dataset and code are available at <https://anonymous.s.4open.science/r/Navig-8788>.

## 1 Introduction

Image geo-localization—the task of predicting the location where an image was taken (Hays and Efros, 2008)—remains a challenging multimodal problem. For example, to say that Figure 1 is a picture from Darlington (in England) requires reading the name of the hotel to determine possible candidates and excluding—for instance—the Croft hotel in Ontario based on the architecture. Directly predicting the exact location or coordinates of an image (Weyand et al., 2016; Haas et al., 2023; Cepeda et al., 2023) is difficult for computer vision models and requires extensive training on large datasets of image-location pairs.

In contrast, human experts infer locations by reasoning. For example, in a GeoGuessr<sup>1</sup> game video, an expert player, *zi8gzag*, explained how he identified a location in Korea: the presence of single yellow road lines and the language on the

<sup>1</sup><http://www.geoguessr.com>



### Reasoning

The climate appears temperate with lush greenery suggesting a region with moderate rainfall, the architecture includes brick buildings and stone structures typical of Northern European styles. The overall environment is peaceful and rural, typical of the countryside in the UK, the specific style of the buildings and road infrastructure aligns with those in the Yorkshire region.

### Map Search



The Croft Hotel, Northallerton Road, Darlington, North Yorkshire, England, United Kingdom", "lat": "54.4824", "lon": "-1.5561"

### Guidebook



The chevrons are black with white arrows in the United Kingdom.

### Location Prediction:

(54.4824, -1.5561), Darlington, U.K.

Figure 1: In image geo-localization, models need to find both cultural and geographical clues to infer correct locations. External tools like maps and guidebooks can also be helpful by providing extra knowledge.

road signs suggest an Asian region; the large spikes atop concrete poles help narrow it down to Japan and Korea, and the black and yellow guardrails rule out Japan. While recent research has integrated textual knowledge (Luo et al., 2022) and explicit clues (Zhang et al., 2024; Mendes et al., 2024; Li et al., 2024) with Vision Language Models (VLMs) to enhance their accuracy, the reasoning in these models is often limited to a few words related to landmarks and does not provide a concrete analysis, as human experts would.

To date, these models' reasoning remains more superficial than humans' for two reasons: (1) **Lack of high-quality reasoning datasets:** Existing geo-

040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053

054 tagged datasets lack linguistic reasoning elements, 103  
055 while constructing a dataset that involves reason- 104  
056 ing based on image details is resource-intensive. 105  
057 (2) **Complexity of diverse information retrieval:** 106  
058 Images often contain rich details, such as road 107  
059 signs, texts, and building styles, requiring addi- 108  
060 tional tools for accurate retrieval and interpretation. 109

061 To address these questions, we introduce NAVI- 110  
062 CLUES, a detailed and high-quality reasoning 111  
063 dataset for image geo-localization, and NAVIG, a 112  
064 framework that combines both visual analysis and 113  
065 external knowledge to perform analytical reason- 114  
066 ing. Inspired by the popular geographical game 115  
067 GeoGuessr, we construct NAVICLUES with over 116  
068 2000 instances from five experienced YouTubers, 117  
069 recording their process of analyzing image details 118  
070 to infer locations, which trains VLMs to generate an- 119  
071 alytical reasoning that mimics professional human 120  
072 players. With tools like public maps and expert- 121  
073 written guidebooks, we design a pipeline that dives 122  
074 into fine-grained details and retrieves relevant infor- 123  
075 mation, enhancing the accuracy of geo-localization. 124  
076 We evaluate NAVIG against state-of-the-art models 125  
077 on two open benchmark datasets using five levels of 126  
078 prediction, and conduct ablation studies to investi- 127  
079 gate the contribution of each component of NAVIG. 128  
080 NAVIG outperforms previous state-of-the-art mod- 129  
081 els by a 14% reduction in average distance error 130  
082 while using less than 1,000 training samples. We 131  
083 further illustrate the analyzing process of NAVIG 132  
084 by providing examples of both successful and chal- 133  
085 lenging cases. We release our dataset and frame- 134  
086 work to advance the use of reasoning in the field of 135  
087 image geo-localization. 136

## 088 2 Collecting NAVICLUES: Linking Places 137 089 to Images 138

090 This section explains how we process the reasoning 139  
091 of GeoGuessr players to construct NAVICLUES. In 140  
092 addition, we analyze their reasoning and identify 141  
093 fifteen key clues humans use in geo-localization. 142

### 094 2.1 Data Collection 143

095 Despite previous efforts to create datasets contain- 144  
096 ing image-location pairs and reasoning insights 145  
097 from guidebooks (Hays and Efros, 2008; Vo et al., 146  
098 2017; Astruc et al., 2024; Luo et al., 2022; Li et al., 147  
099 2024), there is still a lack of datasets that capture 148  
100 the analytical reasoning process used to deduce 149  
101 locations from image details. To train NAVIG to 150  
102 generate reasoning (Section 3.1), we use the data

103 from GeoGuessr, a popular game where players 104  
105 infer locations from street views, which preserve 106  
107 experts’ knowledge and strategies for image geo- 108  
109 localization. We mine game data from “play along” 110  
111 videos of five popular YouTubers, along with tran- 112  
113 scripts of their reasoning during gameplay. 114

115 **Data Mining.** In a typical GeoGuessr game, 116  
117 there are multiple rounds of guessing the location 118  
119 from a new image. To segment the video tran- 120  
121 script, we identify the timestamps of each round’s 122  
123 result pages by using Qwen-VL (Bai et al., 2023) 124  
125 to match the buttons and extract the correspond- 126  
127 ing scores. For images, we retrieve images from 128  
129 Google StreetView (GSV) API based on the coor- 129  
130 dinates of each round, omitting any unavailable 130  
131 ones.<sup>2</sup> Following Haas et al. (2024), we capture 131  
132 images from four different directions and combine 132  
133 them to create 360-degree panoramic views (Ta- 133  
134 ble 1), which contain the same details as in the 134  
135 games. For reasoning data, we split the transcripts 135  
136 by round timestamps. The raw dataset contains 136  
137 2637 images and respective locations. 137

138 **Data Processing.** To ensure data quality, we 139  
140 apply several processing steps: (1) we manually 140  
141 review and remove games where the visual con- 141  
142 tent differs between the players’ view and the GSV 142  
143 data, such as games that allow movement or feature 143  
144 satellite view; (2) we exclude games with poor rea- 144  
145 soning quality, including those with a time limit of 145  
146 less than 30 seconds, transcripts shorter than 100 146  
147 words, or incorrect answers (where the GeoGuessr 147  
148 Score is lower than 3400, approximately 575 km); 148  
149 (3) answer-guided reasoning (Mendes et al., 2024) 149  
150 generates the step-by-step reasoning process from 150  
151 both the transcripts and images, guided by the cor- 151  
152 rect locations and details mentioned by humans. 152

153 After processing, our final dataset, NAVICLUES, 153  
154 consists of 1120 panorama images, each asso- 154  
155 ciated with a corresponding location (continent, 155  
156 country, and coordinates), reasoning process, and 156  
157 scores. This dataset is useful for analyzing human 157  
158 geo-localization strategies and training models for 158  
159 image-based tasks. Prompts, example data, and 159  
160 other details are in the Appendix B. 160

161 **Guidebook Data.** For SEARCHER (Section 3.2), 161  
162 we collect guidebook data from two popular Ge- 162  
163 oGuessr community sources,<sup>3</sup> which include tips 163  
164

<sup>2</sup>As GSV updates the images from time to time, some coordinates are deleted.

<sup>3</sup><https://somerandomstuff1.wordpress.com/2019/02/08/geoguessr-the-top-tips-tricks-and-techniques> and <https://www.plonkit.net>

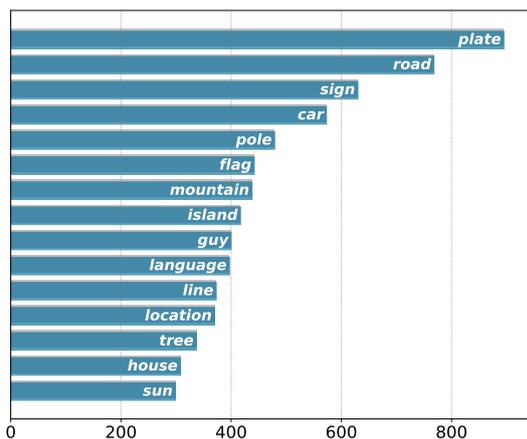


Figure 2: Top clues in human reasoning. Humans identify *roads, cars, poles*, and linguistic clues—specifically the languages on *plates, signs* and *houses*.

from top players of GeoGuessr. Following Luo et al. (2022) and Li et al. (2024), each instance in our dataset contains a clue (e.g., *The chevrons are black with white arrows in the U.K.*) paired with an image. The final dataset contains 6227 image-clue pairs that have distinguishable geographical features, which could be further integrated into geo-localization frameworks as additional knowledge (Examples are in Appendix B).

## 2.2 Data Analysis

We further analyze the reasoning of human players and investigate common patterns among human experts in geo-localization tasks. We use spaCy (Honibal and Johnson, 2015) for noun extraction, allowing us to explore the specific information these experts focus on. We manually filter out irrelevant words and phrases, retaining only content pertinent to geographic reasoning.

Through this process, we identify fifteen core clues frequently mentioned by experts highly associated with geo-localization. These clues reflect common analytical patterns and reasoning strategies (Figure 2). The keywords cover cultural clues (e.g., language, flag, road, house) and natural geographical features (e.g., mountain, island, tree). This distribution conforms to categories in the guidebooks and further guides the implementation of our framework in Section 3.

## 3 NAVIG: Localizing Images with Reasoning and Tools

This section presents NAVIG: image geo-localization with reasoning about cultural and geographical clues and using external tools. NAVIG



**Location**  
Australia, Benalla. -36.5510941, 145.9863798

**Query**  
<image> Given an image, craft a brief and cohesive reasoning path that deduces this location based on the visual clues present in the image. Using a tone of exploration ...

**Response**  
The scene displays **a bright, sunny day**, indicating a temperate climate, **the trees** lining the streets suggest a region with sufficient rainfall, the architecture shows low-rise buildings, typical of smaller towns rather than large urban centers, there are bilingual (or unilingual) English signs, helping to narrow down the location to an English-speaking country, **road markings and signs** indicate that **driving is on the left**, pointing towards countries like Australia or New Zealand, ... the specific layout and style are consistent with towns in Victoria, Australia, further narrowing down the location to Benalla in Victoria, given the visual and cultural context presented.

Table 1: Fine-tuning data format. The reasoning process leverages visual information in the images to deduce the correct location, such as **weather, vegetation, signage**, and **driving orientation**. The complete prompts and response is shown in the Appendix.

(Figure 3) has three components: (1) REASONER, which focuses on generating a reasoning process that analyzes the general information present in the image (Section 3.1); (2) SEARCHER, which uses additional knowledge sources and tools to explore the details (Section 3.2); and (3) GUESSER, which takes the concatenated outputs from both analyzers to determine the final location, which can be configured to any granularity of locations (Section 3.3).

### 3.1 Training VLMS to REASON about Image Locations

Recent VLMS can—sometimes—reason about the location of an image. (Li et al., 2024). However, the reasoning is limited to only a few words and does not help localization (Zhang et al., 2024). To enhance VLMS to reason location-relevant information in images, we create NAVICLUES and fine-tune VLMS using it to build REASONER. The reasoning includes geographical information such as climate, vegetation, language, and driving orientations (Table 1). This approach enables models to deduce locations from geographically pertinent details, expanding the depth and applicability.

After training, REASONER can generate a ra-

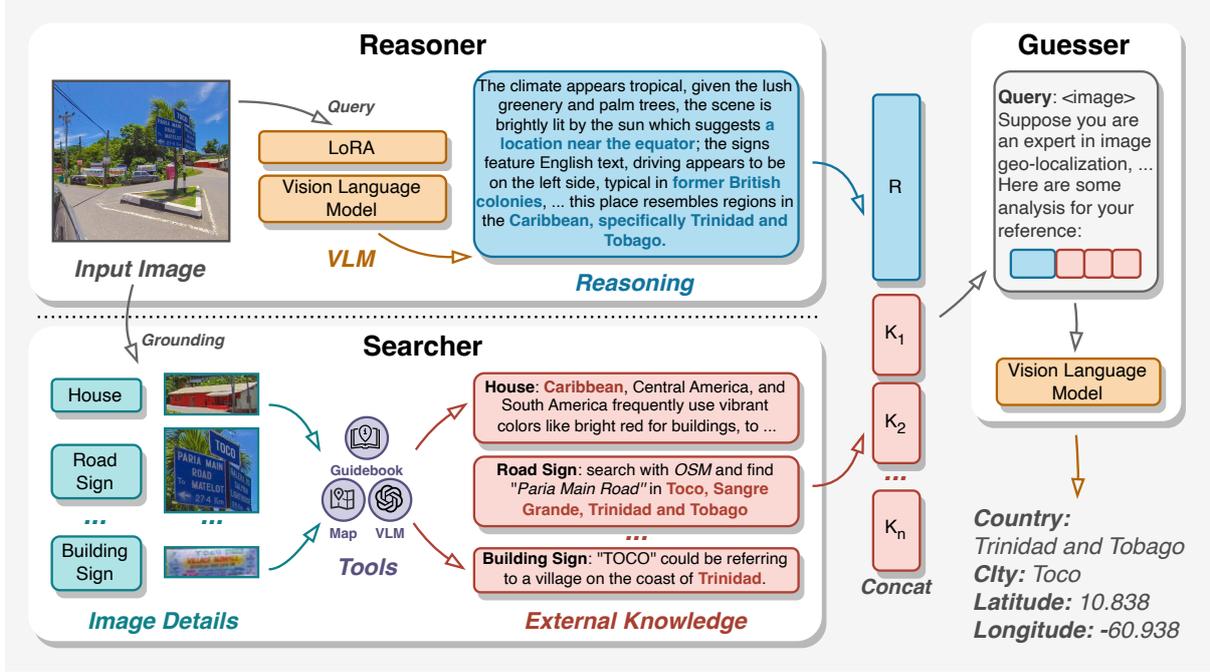


Figure 3: The framework of NAVIG. It comprises three main components: the REASONER, which handles general reasoning; the SEARCHER, which leverage external knowledge for detail-specific analysis, and the GUESSER, which combines outputs from both analyzers to generate predictions.

tionale for images, where given an image  $I$ , the fine-tuned VLM produces a reasoning  $R$ . However, as the reasoning relies solely on VLMs constrained by their parameterized knowledge, it lacks the information to understand specific details. For instance, human experts can search maps for texts on building or road signs and consult guidebooks to identify the house style of a particular country, which goes beyond the intrinsic knowledge within VLMs. To emulate this process, an additional module, SEARCHER, integrates external tools, enabling more accurate interpretation of nuanced details within the reasoning.

### 3.2 SEARCHING Image Details

The SEARCHER module extracts fine-grained details from images to enhance the reasoning by integrating relevant knowledge. It initially crops the image using grounding models, generates queries, and uses tools to retrieve external knowledge.

**Grounding Image Details.** As highlighted in Section 2.2, human experts often concentrate on specific elements in images, such as signs, houses, and roads, which provide crucial location-based clues. To emulate this process, given an image  $I$  and a predefined set of elements  $E = \{e_1, e_2, \dots, e_n\}$ , SEARCHER uses GroundingDino (Liu et al., 2023) to crop the im-

age according to these elements. Since each image may contain multiple instances of a certain element, the resulting set of cropped images is defined as  $C = \{c_{i,j} \mid e_i \in E, j \in [1, m_i]\}$ , where  $m_i$  represents the number of element  $e_i$  in  $I$ . Specifically, we select *road sign*, *building sign*, and *house* from Figure 2 as elements, which align well with GroundingDino, since alternatives could yield overly large figures or uninformative results. Each cropped image  $c_{i,j}$  is used as a query for specific tools. Additionally, if  $c_{i,j}$  is a *sign* that contains rich text information, text-based queries are generated with Optical Character Recognition (OCR) using Qwen2-VL (Wang et al., 2024). Therefore, the query set  $Q$  is defined as:

$$Q = \bigcup_{i,j} \{c_{i,j}, \text{OCR}(c_{i,j}) \text{ if } c_{i,j} \in \text{signs}\}$$

**Tools.** The query set  $Q$  is then fed into a Tool Set  $T$ , which retrieves relevant knowledge. We use three tools for information retrieval. (1) The *GeoGuessr Guidebook* contains rich information for locating images (Section 2.1). Following prior research (Luo et al., 2022; Zhou et al., 2024), we frame *Guidebook* using as a Retrieval-Augmented Generation problem. Given an input image (e.g., a house as in Figure 3), we retrieve the most similar images. (2) *Map*. The map is a critical tool

in image geo-localization, where textual information in images, such as names on signs, can directly pinpoint a location. With text-based queries, we use the OpenStreetMap<sup>4</sup> API for location retrieval, providing the top three search results, each with the place name and multi-level location details. (3) VLM. We use an additional VLM as a tool by prompting it to identify details that might be overlooked in the REASONER, adding further analytical insights. The VLM can generate descriptions for details to narrow down the potential locations (Figure 3). Each tool  $t$  in the Tool Set  $T$  contributes to the retrieval of additional knowledge  $K$ :

$$K = \bigcup_{t \in T} t(Q)$$

Further implementation details about tool parameters are available in the Appendix A.

### 3.3 GUESSING the Final Location

The GUESSER uses all prior information to generate the final prediction. It concatenates the reasoning  $R$  from the REASONER with the external knowledge  $K$  retrieved by the SEARCHER, forms them into a prompt template  $p$  along with the image  $i$ , and makes the location prediction with a VLM:

$$\hat{y}_{\text{loc}} = \text{VLM}_p(I, \text{concat}(R, K))$$

where  $\hat{y}_{\text{loc}}$  is the model’s generated location. Note that  $p$  is configurable, allowing for flexible adjustments to the output format based on specific requirements, such as predicting locations at various levels (*e.g.*, country, city, and coordinates).

## 4 How Well Does NAVIG Reason Image Locations?

This section demonstrates the effectiveness of NAVIG through a series of experiments. We compare it against prior state-of-the-art image geo-localization models and other baseline approaches (Section 4.2), conduct ablation experiments to evaluate the contributions of each module in NAVIG (Section 4.3), and provide qualitative examples to highlight successful and challenging cases for further discussion (Section 4.4).

### 4.1 Experimental Setup

**Implementation.** We use three open-source models in NAVIG, MiniCPM-V (Yao et al., 2024), LLaVA (Liu et al., 2024), and Qwen2-VL (Wang

et al., 2024). These models serve as VLMs for REASONER, SEARCHER, and the GUESSER components within the NAVIG framework. (1) For REASONER, Low-Rank Adaptation (LoRA) (Hu et al., 2022) fine-tunes models using the NAVICLUES dataset. We use *minicpm-v-2.6*, *llava-1.6-vicuna-7b*, and *qwen2-vl-7b* due to their advanced performance and mid-range size for training costs. (2) For SEARCHER, we select the top three cropped clues as the basis for generation (*e.g.*, if multiple houses are cropped, only the three with the highest similarity scores will be analyzed). We use CLIP (Radford et al., 2021) to encode both guidebook images and query images, construct a database using FAISS (Johnson et al., 2019), and retrieve guidebook data by the Euclidean distance  $d$  between image embeddings, returning associated text if  $d$  is below a threshold  $d_t$  (set to 30). We prompt the GUESSER to predict locations at the coordinates level. Training hyperparameters, model configurations, and prompts are in Appendix A.

**Baselines.** We compare NAVIG with two types of baselines: (1) *Geo-localization Models*: we select top-performing open-source models from prior research in image geo-localization, including  $G^3$  (Luo et al., 2022), GeoCLIP (Cepeda et al., 2023), and StreetCLIP (Haas et al., 2023). (2) *Vision Language Models*: we select vanilla MiniCPM-V, LLaVA, Qwen2-VL as baselines, consistent with the backbone models used in NAVIG. The prompts for these VLM baselines are identical to those in NAVIG, with only the analyses removed. We do not include commercial closed-source models, as training on these models is not feasible for a fair comparison.

**Dataset and Metrics.** Following previous work (Hays and Efros, 2008; Astruc et al., 2024; Haas et al., 2024), we evaluate our framework on two public datasets, including GWS5K sampled from GWS15K (Clark et al., 2023) due to cost constraints, and Im2GPS3k (Hays and Efros, 2008). First, we computed the haversine distance between the predicted and ground truth coordinates. For models limited to city level outputs, we use the coordinates of the predicted city as their predictions. Next, we evaluated the prediction accuracy—the percentage of guesses that fall within a distance threshold from the correct location—at five geographic levels: Street (1 km), City (25 km), Region (200 km), Country (750 km), and Continent (2,500 km). In addition, we calculated the average error distance and GeoGuessr Score, a metric from the

<sup>4</sup><https://www.openstreetmap.org/>

Model	Continent 2,500 km	Country 750 km	Region 200 km	City 25 km	Street 1 km	Distance↓	Score↑
$G^3$	50.9	14.6	2.3	0.1	0.0	4,341	1,304
GeoCLIP	78.2	46.5	17.1	3.5	0.4	2,099	2,613
StreetCLIP	79.4	43.4	13.4	1.7	0.3	2,060	2,543
MiniCPM-V	27.1	15.9	6.7	1.6	0.1	7,320	909
LLaVA	43.9	23.1	7.0	1.2	0.0	5,096	1,418
Qwen2-VL	89.4	66.7	31.8	6.1	0.1	1,124	3,344
<b>NAVIG</b>							
- MiniCPM-V	71.5	44.1	16.9	3.5	0.3	2,956	2,413
- LLaVA	74.7	39.4	12.0	1.9	0.3	2,243	2,354
- Qwen2-VL	<b>91.1</b>	<b>66.9</b>	<b>31.9</b>	<b>6.7</b>	<b>0.7</b>	<b>965</b>	<b>3,389</b>

Table 2: Accuracy and scores on GWS5k. The data from Continent to Street represents the accuracy (%) at each level. The three sections are geo-localization models, VLMS, and NAVIG. **Bold** font indicates the best performance. NAVIG (Qwen2-VL) achieves the highest accuracy across all metrics.

Model	ROUGE F1		
	R1	R2	RL
<b>REASONER</b> (MiniCPM-V)	51.0	<b>14.8</b>	<b>24.6</b>
MiniCPM-V	46.4	12.6	22.1
<b>REASONER</b> (LLaVA)	49.8	13.9	24.0
LLaVA	44.7	10.8	21.8
<b>REASONER</b> (Qwen2-VL)	<b>51.4</b>	14.6	24.3
Qwen2-VL	45.2	12.3	22.1

Table 3: ROUGE F1 scores for reasoning generated by models and humans (%). REASONER models reason more similarly to humans.

Model	Country	City	Street	Score↑
<b>NAVIG</b>				
- MiniCPM-V	56.0	18.0	0.0	2,863
- LLaVA	48.0	14.0	0.0	2,690
- Qwen2-VL	<b>86.0</b>	32.0	4.0	<b>4,202</b>
<b>Human Players</b>	76.0	<b>48.0</b>	<b>42.0</b>	3,757

Table 4: Performance between humans and NAVIG. The data from City to Street represents accuracy (%). Our best model beats humans with a higher overall score but still struggles to achieve fine-grained accuracy.

original GeoGuessr game that quantifies guess accuracy, with a scoring range of 0 to 5000. Details about metric computation are in Appendix C.

## 4.2 Main Experiments

**Accuracy.** We compare NAVIG with state-of-the-art image geo-localization models and Vision Language Models (GWS5k results in Table 2). (1) Generally, within the framework of NAVIG, Qwen2-VL achieved the highest accuracy across all metrics, beating specialized geo-localization models trained on domain-specific datasets, despite its relatively compact size of only 7 billion parameters. (2) All VLMS generate effective analytical reasoning, which is trained with only around 1,000 samples. These findings underscore the quality of training data and the efficacy of the NAVIG framework. Similar results on Im2GPS3k are in Appendix D.

**Reasoning.** We evaluate the quality of the linguistic reasoning generated by the model on a reserved test set of 50 human games. To measure the alignment between model-generated and human reasoning, we compute their ROUGE scores (Lin, 2004), which illustrate whether the model sim-

ulates human reasoning. REASONER achieves higher ROUGE scores across all models and metrics after training (Table 3). We further examine the effectiveness of reasoning through ablation.

**Comparison with Humans.** We also compare NAVIG’s performance against human players in 50 randomly sampled GeoGuessr games, focusing on common metrics for country, city and street level predictions. NAVIG outperforms humans in overall scores (Table 4), although humans excel at finer-grained predictions by iteratively cross-referencing maps and comparing terrain and features within the game. This highlights a future direction to use non-textual features to refine map-based searches.

## 4.3 Ablation Study

To illustrate the contributions of each component in NAVIG, we ablate the reasoning training, the impact of REASONER, and SEARCHER. Table 5 presents the three VLMS’ accuracy on GWS5k. In this setup, NAVIG represents our framework, “w/o training” denotes results with the same prompt but without training on NAVICLUES, “w/o Macro” and “w/o Micro” refer to the results without the REASONER and SEARCHER modules, respectively.

Model	Country	City	Street
NAVIG (MiniCPM-V)	<b>44.1</b>	<b>3.5</b>	<b>0.3</b>
- w/o training	- 3.3	- 0.4	- 0.2
- w/o REASONER	- 10.2	- 0.7	- 0.0
- w/o SEARCHER	- 0.3	- 0.3	- 0.2
- MiniCPM-V	- 14.9	- 0.5	- 0.2
NAVIG (LLaVA)	39.4	<b>1.9</b>	<b>0.3</b>
- w/o training	- 25.8	- 1.2	- 0.3
- w/o REASONER	- 20.2	- 0.8	- 0.0
- w/o SEARCHER	<b>+ 0.4</b>	- 0.2	- 0.2
- LLaVA	- 16.3	- 0.7	- 0.3
NAVIG (Qwen2-VL)	66.9	<b>6.7</b>	<b>0.7</b>
- w/o training	- 6.0	- 0.9	- 0.5
- w/o REASONER	- 4.0	- 0.6	- 0.2
- w/o SEARCHER	<b>+ 0.1</b>	- 0.9	- 0.5
- Qwen2-VL	- 0.2	- 0.6	- 0.6

Table 5: Ablation results of NAVIG on the GWS5k dataset. Each component contributes to model accuracy, with their removal leading to notable declines across Country, City, and Street levels.

**Results.** (1) Each module contributes to improving the model’s accuracy. (2) Surprisingly, when the model is prompted to generate reasoning processes in a zero-shot setting, the reasoning can be misleading, resulting in decreased final prediction accuracy. This highlights the necessity of training the model with NAVICLUES. (3) REASONER plays a critical role in coarse-grained localization, with improvements at the country level and decrease without it, as the reasoning in the training dataset is limited to the country and city level. (4) SEARCHER substantially enhances fine-grained reasoning. Achieving precise street-level localization on the GWS dataset is highly challenging, but the SEARCHER narrows the scope within 1 km for images containing textual information by using map searches (Table 2). Results on Im2GPS3k are in Appendix D, which is consistent with GWS.

#### 4.4 Qualitative Analysis

This section examines how the analytical reasoning derived from images contributes to NAVIG’s inference process. As shown in Figure 4 (top), NAVIG closely examines details within the image, such as the temperate climate, orientation of driving cars, and “Lower Mill” to determine the location. This detailed reasoning narrows down the possible range, while integration with OpenStreetMap data further aids the model in finding the restaurant, with an error distance of under 1 meter, improving its estimate by 144 km.

However, image elements can also mislead the

model. In Figure 4 (middle), the model fixates on a shop name in the image, “KLICK”, which can be interpreted as a German word. This leads the reasoning process astray, resulting in an incorrect localization. OpenStreetMap can also lead to false predictions when there are places with the same name, such as “Bradesco”, a well-known Brazilian bank (Figure 4 (bottom)). The reasoning makes image geo-localization models more interpretable by revealing how image elements influence decisions.

## 5 Related Work

**Image geo-localization** Image geo-localization falls into three methods: (1) *Retrieval-based methods* retrieves the most similar images (Hays and Efros, 2008; Zhu et al., 2023). Various retrievers (Vo et al., 2017; Pramanick et al., 2022; Haas et al., 2023) and gallery types (Cepeda et al., 2023) have been proposed. (2) *Classification-based methods* divide geographical maps into distinct classes—such as countries, cities, or geographical cells—and train models to classify the location of images into these categories. Researchers have proposed different model structures (Radford et al., 2021; Wu and Huang, 2022) and map division strategies (Weyand et al., 2016; Theiner et al., 2022; Haas et al., 2024) to improve accuracy. Despite this, these methods are limited by the size and scale of the defined granularity. (3) *Generation-based methods* use visual understanding and generation in Vision Language Models (VLMs) to directly generate the location or coordinates for geo-localization. By aligning visual content with rich text descriptions and reasoning (Jia et al., 2024; Li et al., 2024; Zhang et al., 2024), along with incorporating external knowledge through Retrieval-Augmented Generation (Luo et al., 2022; Zhou et al., 2024), these methods have achieved state-of-the-art performance. However, challenges persist in effectively using VLMs, including limited reasoning data for model training, a reliance on constrained external knowledge sources, such as image galleries.

**Visual Reasoning** In Visual Reasoning, models need to derive solutions from image details to answer questions, which requires both visual understanding and reasoning capabilities (Hudson and Manning, 2019; Gupta and Kembhavi, 2023). Recently, with relevant techniques like In-context Learning, Chain of Thought, and tool using, Vision Language Models (VLMs) have demonstrated exceptional performance in visual reasoning

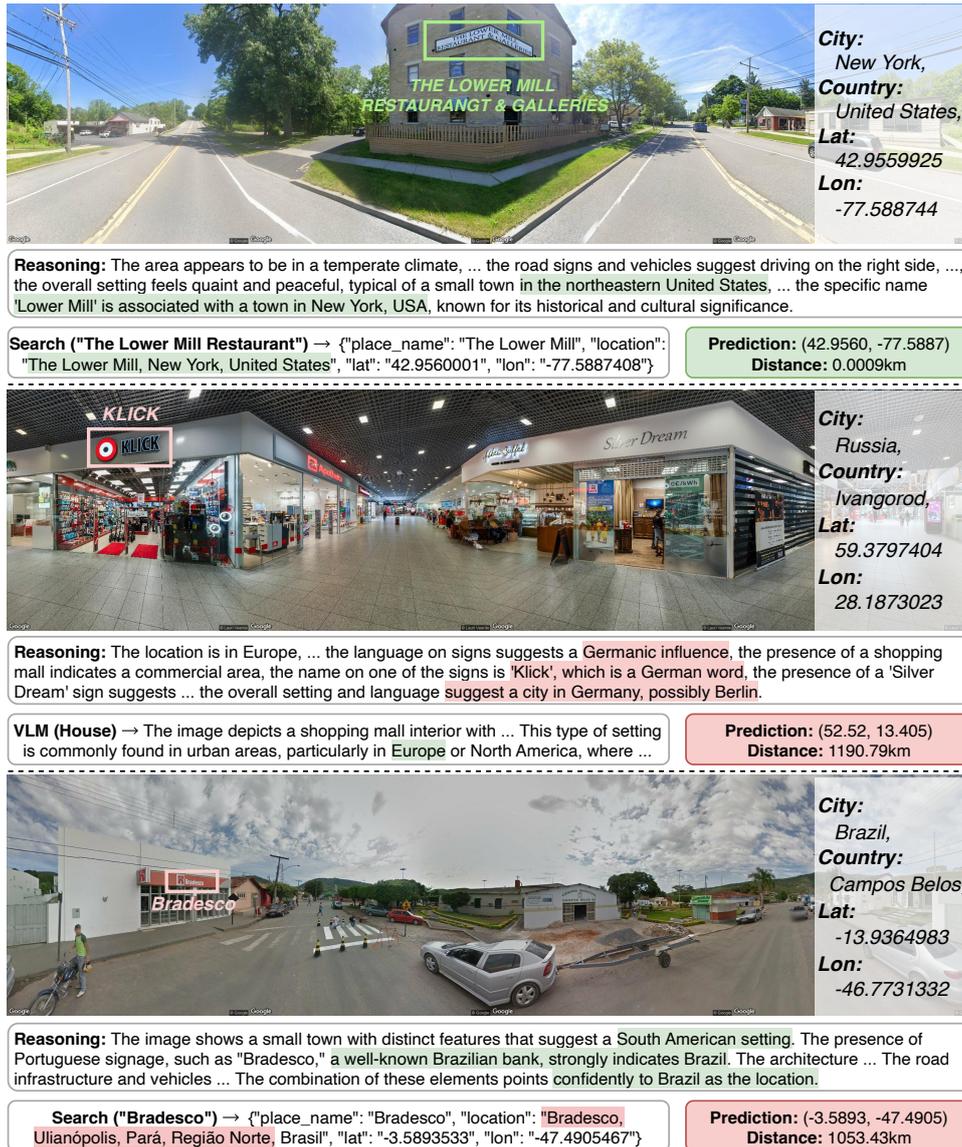


Figure 4: **Top:** The model uses visual details and OpenStreetMap to accurately determine the location. **Middle:** The model is misled by linguistic elements—the shop name, resulting in an incorrect inference. **Bottom:** The model found a namesake when using OpenStreetMap.

484 tasks (Alayrac et al., 2022; Lu et al., 2023).

485 Generally, key directions for enhancing the vi-  
486 sual reasoning of VLMs in specific tasks include:  
487 (1) *High-quality complex reasoning data*, which  
488 researchers have shown to be particularly effective  
489 in improving the performance of VLMs (Du et al.,  
490 2023; Chen et al., 2023); (2) *Vision grounding*,  
491 which enables models to ground in the details of  
492 the image and perform step-by-step reasoning (Qi  
493 et al., 2024; Wu and Xie, 2023; Zhang et al., 2024);  
494 and (3) *Tool using*, which aid the model by lever-  
495 aging tools to retrieve additional knowledge (Yang  
496 et al., 2023; Marino et al., 2021; Chen et al., 2022).  
497 Unlike traditional end-to-end methods that map im-  
498 ages directly to locations, we treat geo-localization

499 as a complex reasoning task that deduces the loca-  
500 tion with language in a pipeline.

## 6 Conclusion 501

502 We introduce a novel framework NAVIG and a  
503 reasoning dataset NAVICLUES, designed to per-  
504 form image geo-localization through detailed vi-  
505 sual reasoning and supplementary knowledge re-  
506 trieval. We demonstrate the effectiveness of our ap-  
507 proach through comprehensive comparisons with  
508 state-of-the-art models and ablation studies. Fu-  
509 ture developments could include expanding more  
510 tools and using these tools to improve results for  
511 finer-grained predictions.

## 512 Limitations

513 **Datasets.** In this work, we utilize data from hu- 562  
514 man players in the GeoGuessr game to train Vision 563  
515 Language Models for performing geographic rea- 564  
516 soning on images. The copyright and usage rights 565  
517 of the images are subject to that of Google Street 566  
518 View. However, the dataset size of NAVICLUES 567  
519 is limited due to the scarcity of available data on 568  
520 YouTube and the noise in the collected data. To sim- 569  
521 ulate the perspective of players in the GeoGuessr 570  
522 game, we use stitched panoramic images as the 571  
523 input to the model. Furthermore, nearly all im- 572  
524 ages in the data from GeoGuessr are street views, 573  
525 despite our efforts to ensure a geographically bal- 574  
526 anced distribution of data across countries. Conse- 575  
527 quently, the training data is limited to certain sizes 576  
528 and types, which might constrain its performance.

529 Future work could consider expanding the training 577  
530 dataset by incorporating images of different sizes, 578  
531 and types to further enhance the performance of 579  
532 image geo-localization tasks with better reasoning. 580  
533 **Models.** Due to cost constraints, we conducted 581  
534 our experiments using a limited number of medium- 582  
535 sized open-source models (around 7B parameters). 583  
536 This choice may result in performance that is not 584  
537 as competitive as larger models. Additionally, we 585  
538 evaluated only a limited set of tools and grounding 586  
539 words. Identifying more geographic features such 587  
540 as cars, road markings, and poles would require 588  
541 more precise recognition methods and more so- 589  
542 phisticated model designs, which could potentially 590  
543 improve performance. We employed a pipeline 591  
544 approach to construct our model, aiming to maxi- 592  
545 mize the performance of each component at every 593  
546 stage. Future work could also explore building 594  
547 finer-grained datasets to train end-to-end model 595  
548 training for better performance. 600

549 **Experiments.** Due to the limitations of NAVI- 601  
550 CLUES, the reasoning process is constrained to 602  
551 the country or city level. In practice, the level 603  
552 of reasoning could significantly impact the final 604  
553 results. Adding additional experiments, such as 605  
554 training with data constructed at the coordinates 606  
555 level, could potentially improve overall accuracy 607  
556 and help address more research questions. 608

## 557 Ethical Considerations

558 In this work, we use the data from GeoGuessr play- 609  
559 ers on YouTube to train our models. We carefully 610  
560 process the data and remove the personal informa- 611  
561 tion of the players, and all the data are only used 612

613 for academic purposes. While the task of image 614  
615 geo-localization has the potential to enable innova- 616  
617 tive applications in fields such as navigation and 618  
619 tourism, the misuse of these models could also lead 619  
620 to risks such as privacy breaches and surveillance. 620  
621 In our work, we ensured that all training and testing 621  
622 data came from publicly available sources, with no 622  
623 involvement of private or personal images or loca- 623  
624 tion data. Currently, as shown in our experiments, 624  
625 these models have not yet reached a level of pre- 625  
626 cision to accurately predict coordinates-level loca- 626  
627 tions. For the future development of this field, it is 627  
628 crucial for researchers to ensure that these models 628  
629 are used within appropriate boundaries to prevent 629  
630 the leakage of private information. 630

## 631 References

- 632 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, 633  
634 Antoine Miech, Iain Barr, Yana Hasson, Karel 634  
635 Lenc, Arthur Mensch, Katherine Millican, Malcolm 635  
636 Reynolds, Roman Ring, Eliza Rutherford, Serkan 636  
637 Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, 637  
638 Marianne Monteiro, Jacob L. Menick, Sebastian 638  
639 Borgeaud, Andy Brock, Aida Nematzadeh, Sahand 639  
640 Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, 640  
641 Oriol Vinyals, Andrew Zisserman, and Karén Si- 641  
642 monyan. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. 642
- 643 Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, 643  
644 Constantin Aronsohn, Nacim Bouia, Stephanie Fu, 644  
645 Romain Loiseau, Van Nguyen Nguyen, Charles 645  
646 Raude, Elliot Vincent, et al. 2024. [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. 646
- 647 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, 647  
648 Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, 648  
649 and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). 649
- 650 Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and 650  
651 Mubarak Shah. 2023. [Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. 651
- 652 Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, 652  
653 Feng Zhu, and Rui Zhao. 2023. [Shikra: Unleashing](#) 653

616	<a href="#">multimodal llm’s referential dialogue magic</a> . <i>ArXiv preprint</i> , abs/2306.15195.	
617		
618	Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z Pan, Ningyu Zhang, and Wen Zhang. 2022. <a href="#">Lako: Knowledge-driven visual question answering via late knowledge-to-text injection</a> . In <i>Proceedings of the 11th International Joint Conference on Knowledge Graphs</i> , pages 20–29.	
619		
620		
621		
622		
623		
624	Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. 2023. <a href="#">Where we are and what we’re looking at: Query based worldwide image geo-localization using hierarchies and scenes</a> . In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023</i> , pages 23182–23190. IEEE.	
625		
626		
627		
628		
629		
630		
631		
632	Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. 2023. <a href="#">What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning</a> . <i>ArXiv preprint</i> , abs/2311.01487.	
633		
634		
635		
636		
637		
638	Tanmay Gupta and Aniruddha Kembhavi. 2023. <a href="#">Visual programming: Compositional visual reasoning without training</a> . In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023</i> , pages 14953–14962. IEEE.	
639		
640		
641		
642		
643		
644	Lukas Haas, Silas Alberti, and Michal Skreta. 2023. <a href="#">Learning generalized zero-shot learners for open-domain image geolocation</a> . <i>ArXiv preprint</i> , abs/2302.00275.	
645		
646		
647		
648	Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. 2024. <a href="#">Pigeon: Predicting image geolocations</a> . In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 12893–12902.	
649		
650		
651		
652		
653	James Hays and Alexei A. Efros. 2008. <a href="#">IM2GPS: estimating geographic information from a single image</a> . In <i>2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA</i> . IEEE Computer Society.	
654		
655		
656		
657		
658		
659	Matthew Honnibal and Mark Johnson. 2015. <a href="#">An improved non-monotonic transition system for dependency parsing</a> . In <i>Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing</i> , pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.	
660		
661		
662		
663		
664		
665	Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. <a href="#">Lora: Low-rank adaptation of large language models</a> . In <i>The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022</i> . OpenReview.net.	
666		
667		
668		
669		
670		
	Drew A. Hudson and Christopher D. Manning. 2019. <a href="#">GQA: A new dataset for real-world visual reasoning and compositional question answering</a> . In <i>IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019</i> , pages 6700–6709. Computer Vision Foundation / IEEE.	671
		672
		673
		674
		675
		676
		677
	Pengyue Jia, Yiding Liu, Xiaopeng Li, Xiangyu Zhao, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, and Dawei Yin. 2024. <a href="#">G3: An effective and adaptive framework for worldwide geolocalization using large multi-modality models</a> . <i>ArXiv preprint</i> , abs/2405.14702.	678
		679
		680
		681
		682
		683
	Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. <a href="#">Billion-scale similarity search with gpus</a> . <i>IEEE Transactions on Big Data</i> , 7(3):535–547.	684
		685
		686
	Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. <a href="#">GeoReasoner: Geo-localization with reasoning in street views using a large vision-language model</a> . In <i>Proceedings of the 41st International Conference on Machine Learning</i> , volume 235 of <i>Proceedings of Machine Learning Research</i> , pages 29222–29233. PMLR.	687
		688
		689
		690
		691
		692
		693
	Chin-Yew Lin. 2004. <a href="#">ROUGE: A package for automatic evaluation of summaries</a> . In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.	694
		695
		696
		697
	Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. <a href="#">Llava-next: Improved reasoning, ocr, and world knowledge</a> .	698
		699
		700
	Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. <a href="#">Grounding dino: Marrying dino with grounded pre-training for open-set object detection</a> . <i>ArXiv preprint</i> , abs/2303.05499.	701
		702
		703
		704
		705
	Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. <a href="#">Chameleon: Plug-and-play compositional reasoning with large language models</a> . In <i>Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023</i> .	706
		707
		708
		709
		710
		711
		712
		713
	Grace Luo, Giscard Biamby, Trevor Darrell, Daniel Fried, and Anna Rohrbach. 2022. <a href="#">G3: Geolocation via guidebook grounding</a> . In <i>Findings of the Association for Computational Linguistics: EMNLP 2022</i> , pages 5841–5853, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	714
		715
		716
		717
		718
		719
	Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. <a href="#">KRISP: integrating implicit and symbolic knowledge for open-domain knowledge-based VQA</a> . In <i>IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021</i> , pages 14111–14121. Computer Vision Foundation / IEEE.	720
		721
		722
		723
		724
		725
		726

727	Ethan Mendes, Yang Chen, James Hays, Sauvik Das,	Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin	782
728	Wei Xu, and Alan Ritter. 2024. <a href="#">Granular privacy</a>	Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu,	783
729	<a href="#">control for geolocation with vision language models.</a>	Ce Liu, Michael Zeng, and Lijuan Wang. 2023. <a href="#">Mm-</a>	784
730	<i>ArXiv preprint</i> , abs/2407.04952.	<a href="#">react: Prompting chatgpt for multimodal reasoning</a>	785
		<a href="#">and action.</a> <i>ArXiv preprint</i> , abs/2303.11381.	786
731	Shraman Pramanick, Ewa M Nowara, Joshua Gleason,	Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang,	787
732	Carlos D Castillo, and Rama Chellappa. 2022.	Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li,	788
733	<a href="#">Where in the world is this image? transformer-based</a>	Weilin Zhao, Zhihui He, et al. 2024. <a href="#">Minicpm-v:</a>	789
734	<a href="#">geo-localization in the wild.</a> In <i>European Conference</i>	<a href="#">A gpt-4v level mllm on your phone.</a> <i>ArXiv preprint</i> ,	790
735	<i>on Computer Vision</i> , pages 196–215. Springer.	abs/2408.01800.	791
736	Ji Qi, Ming Ding, Weihang Wang, Yushi Bai, Qingsong	Gengyuan Zhang, Yurui Zhang, Kerui Zhang, and	792
737	Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao	Volker Tresp. 2024. <a href="#">Can vision-language models</a>	793
738	Dong, et al. 2024. <a href="#">Cogcom: Train large vision-</a>	<a href="#">be a good guesser? exploring vlms for times and</a>	794
739	<a href="#">language models diving into details through chain of</a>	<a href="#">location reasoning.</a> In <i>Proceedings of the IEEE/CVF</i>	795
740	<a href="#">manipulations.</a> <i>ArXiv preprint</i> , abs/2402.04236.	<i>Winter Conference on Applications of Computer Vi-</i>	796
		<i>sion</i> , pages 636–645.	797
741	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya	Zhongliang Zhou, Jieliu Zhang, Zihan Guan, Mengxuan	798
742	Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-	Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai.	799
743	try, Amanda Askell, Pamela Mishkin, Jack Clark,	2024. <a href="#">Img2loc: Revisiting image geolocation</a>	800
744	Gretchen Krueger, and Ilya Sutskever. 2021. <a href="#">Learn-</a>	<a href="#">using multi-modality foundation models and image-</a>	801
745	<a href="#">ing transferable visual models from natural language</a>	<a href="#">based retrieval-augmented generation.</a> In <i>Proceed-</i>	802
746	<a href="#">supervision.</a> In <i>Proceedings of the 38th International</i>	<i>ings of the 47th International ACM SIGIR Confer-</i>	803
747	<i>Conference on Machine Learning, ICML 2021, 18-24</i>	<i>ence on Research and Development in Information</i>	804
748	<i>July 2021, Virtual Event</i> , volume 139 of <i>Proceedings</i>	<i>Retrieval</i> , pages 2749–2754.	805
749	<i>of Machine Learning Research</i> , pages 8748–8763.		
750	PMLR.		
751	Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth.	Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah,	806
752	2022. <a href="#">Interpretable semantic photo geolocation.</a> In	Xiaohui Shen, and Heng Wang. 2023. <a href="#">R2former:</a>	807
753	<i>Proceedings of the IEEE/CVF Winter Conference on</i>	<a href="#">Unified retrieval and reranking transformer for place</a>	808
754	<i>Applications of Computer Vision</i> , pages 750–760.	<a href="#">recognition.</a> In <i>Proceedings of the IEEE/CVF Con-</i>	809
		<i>ference on Computer Vision and Pattern Recognition</i> ,	810
755	Nam N. Vo, Nathan Jacobs, and James Hays. 2017.	pages 19370–19380.	811
756	<a href="#">Revisiting IM2GPS in the deep learning era.</a> In <i>IEEE</i>		
757	<i>International Conference on Computer Vision, ICCV</i>		
758	<i>2017, Venice, Italy, October 22-29, 2017</i> , pages 2640–		
759	2649. IEEE Computer Society.		
760	Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhi-		
761	hao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin		
762	Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei		
763	Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang		
764	Zhou, Jingren Zhou, and Junyang Lin. 2024. <a href="#">Qwen2-</a>		
765	<a href="#">vl: Enhancing vision-language model’s perception</a>		
766	<a href="#">of the world at any resolution.</a> <i>ArXiv preprint</i> ,		
767	abs/2409.12191.		
768	Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016.		
769	<a href="#">Planet-photo geolocation with convolutional neural</a>		
770	<a href="#">networks.</a> In <i>Computer Vision–ECCV 2016: 14th Eu-</i>		
771	<i>ropean Conference, Amsterdam, The Netherlands,</i>		
772	<i>October 11-14, 2016, Proceedings, Part VIII 14,</i>		
773	pages 37–55. Springer.		
774	Meiliu Wu and Qunying Huang. 2022. <a href="#">Im2city: im-</a>		
775	<a href="#">age geo-localization via multi-modal learning.</a> In		
776	<i>Proceedings of the 5th ACM SIGSPATIAL Interna-</i>		
777	<i>tional Workshop on AI for Geographic Knowledge</i>		
778	<i>Discovery</i> , pages 50–61.		
779	Penghao Wu and Saining Xie. 2023. <a href="#">v*:</a> Guided vi-		
780	<a href="#">sual search as a core mechanism in multimodal llms.</a>		
781	<i>ArXiv preprint</i> , abs/2312.14135.		

## A Implementation Details

### A.1 Training Parameters

We trained the REASONER on Nvidia RTX 6000 Ada (48G), with CUDA 12.4, Transformers 4.45.1, and Pytorch 2.1.2.

Parameters	Value
Max Length	2048
LoRA Rank	8
LoRA Alpha	32
Optimizer	AdamW
Adam Beta1	0.9
Adam Beta2	0.95
Learning Rate	1e-4
Warmup Ratio	0.05
LR Scheduler Type	cosine
Batch Size	1
Weight Decay	0.1

Table 6: Training parameters for REASONER.

### A.2 Other Parameters.

For reproducibility, we also provide the parameters used in other modules and VLMs within our framework.

**GroundingDino.** We utilize GroundingDino to crop detailed information from the images, such as signs and houses. We observe variation in the features of images across different datasets. For instance, the GWS5k dataset focuses on street scenes, and other datasets contain considerable noise (*e.g.*, animals). Consequently, to reduce noise that could potentially affect model performance, we empirically set the thresholds as follows: *Box-Threshold* = 0.5 and *Text-Threshold* = 0.5 for GWS5K, and *Box-Threshold* = 0.8 and *Text-Threshold* = 0.6 for Im2GPS3k.

**Retrieval-Augmented Generation.** We employ CLIP as the image encoder for guidebook clues, using ViT-B-32 as the vision encoder. The guidebook database is deployed with FAISS, and similarity is calculated using Euclidean Distance. The number of most relevant retrieved images,  $k$ , is set to 3, with a similarity threshold of 30.

**OpenStreetMap.** We use the *Nominatim Search API* to process map searches, which takes text queries, and return the most relevant results along with the place name, address, and coordinates.

**Vision Language Models.** We use Vision-Language Models in our framework for reasoning and location inference. The three models are *minicpm-v-2.6*, *llava-1.6-vicuna-7b*, and

*qwen2-vl-7b*. Each model is configured with a temperature of 0 and an output length of 2048.

### A.3 Prompts for VLMs.

In Table 7 and Table 8, we present the prompts used in NAVIG for Vision Language Models. Four distinct prompts are employed: (1) the *Data Processing Prompt*, which employed an answer guided reasoning generation method to prompt VLMs in extracting step-by-step reasoning from YouTube transcripts; (2) the *REASONER Prompt*, which is the same as the query in the training data, prompting VLMs to generate a coherent reasoning process to infer the location within an image; (3) the *SEARCHER Prompt*, which generates additional knowledge from image details, and (4) the *GUESSER Prompt*, which synthesizes all prior information to make a final prediction.

#### *Data Processing Prompt*

<image> Given an image and the known location details (Country: country, Latitude: lat, Longitude: lon), and an expert’s analysis of the location (transcript), craft a brief and cohesive reasoning path that deduces this location based on the visual clues present in the image. Begin your reasoning without revealing that you know the exact location, using a tone of exploration and inference. Carefully analyze and link observations of natural features (climate, vegetation, terrain), man-made structures (roads, buildings, signage), and distinct landmarks. Allow these observations to naturally lead you to the correct country, enhancing the accuracy of your deductions. Ensure that while the narrative seems to be guessing, it aligns with the known country, confirming the reliability of your reasoning without stating the specific coordinates. Start the reasoning without any intro, and make sure to make it brief.

Table 7: The prompts used in NAVIG.

## B Data.

In this section, we present the data processing workflows and provide more detailed information on the various types of data used in the system.

### B.1 Data Processing.

**YouTubers.** We utilized the scripts of five professional GeoGuessr players’ YouTube videos as the starting data for our reasoning generation. We thank these five players for their contributions to knowledge dissemination and promotion of image geo-localization: zi8gzag, GeoWizard, GeoPeter, Geogasm, and RAINBOLT TWO.

**Data Processing.** We used the Google Street View<sup>5</sup> API to retrieve images for our dataset. We

<sup>5</sup><https://www.google.com/streetview/>

879 selected a resolution of 640×640 pixels (the maxi-  
 880 mum resolution accepted by GSV), a field of view  
 881 (FOV) of 90, and headings of 0, 90, 180, and 270  
 882 degrees to obtain four images. Stitching them to-  
 883 gether produces a complete street view image, pro-  
 884 viding the same amount of information that a Ge-  
 885 oGuessr player would see.

886 Next, we split the videos for retrieving the tran-  
 887 scripts or each round. After a player submits their  
 888 final guess, the game reveals the distance between  
 889 their guessed location and the actual coordinates,  
 890 where the player can choose to either proceed to  
 891 the next round or end the challenge. We use precise  
 892 pixel coordinates in conjunction with OCR technol-  
 893 ogy to detect the presence of the “Next” or “End”  
 894 buttons and split the videos. We sample frames  
 895 at a rate of 1/6 per second to ensure no scene is  
 896 missed. Simultaneously, we extract the GeoGuessr  
 897 Score displayed beside the button. Next, due to  
 898 the noise in the data (with many informal language  
 899 from players), we provide GPT-4o with the correct  
 900 locations for paraphrasing and generating higher  
 901 quality and more coherent data.

---

**REASONER Prompt**

<image> Given an image, craft a brief and cohesive rea-  
 soning path that deduces this location based on the visual  
 clues present in the image. Using a tone of exploration  
 and inference. Carefully analyze and link observations of  
 natural features (climate, vegetation, terrain), man-made  
 structures (roads, buildings, signage), and distinct land-  
 marks. Allow these observations to naturally lead you to  
 the correct country, enhancing the accuracy of your deduc-  
 tions. Start the reasoning without any intro, and make sure  
 to make it brief.

---

**SEARCHER Prompt**

<image> Analyze the {item} images to determine the re-  
 gion with the highest likelihood of finding this type of  
 {item}. For each image, provide only the core reasoning in  
 one sentence. Don’t say you can’t determine, try your best  
 as it’s a geo-localization game

---

**GUESSER Prompt**

<image> <information> Using the provided information  
 as a reference, estimate the location depicted in the image  
 with as much accuracy and precision as possible. Gen-  
 erally, you might use the reasoning to roughly locate the  
 coarse-grained location, and use other information to help  
 you decide more precisely. Use your own knowledge as  
 well. Aim to deduce the exact coordinates whenever fea-  
 sible. Format your response strictly as JSON in the fol-  
 lowing structure: {“country”: “<country\_name>”, “city”:  
 “<city\_name>”, “latitude”: <Latitude Coordinate>, “longi-  
 tude”: <Longitude Coordinate>} Ensure the JSON output  
 is correctly formatted. Provide a well-informed estimate  
 for each value, avoiding any empty fields. Do not include  
 additional information or commentary.

---

Table 8: The prompts used in NAVIG.

**B.2 Data Demonstration.**

In this section, we present examples and key statis-  
 tics for both NAVICLUES and guidebook datasets.

**NAVICLUES.** As shown in Figure 9, the data  
 includes a panoramic image, the corresponding loca-  
 tion, and a high-quality reasoning process that  
 shows how geographical and cultural information,  
 such as vegetation, landmarks, and text on signage,  
 is used to infer the location. We do not require  
 the model to generate specific street-level locations  
 or coordinates directly, as these details could intro-  
 duce excessive hallucination. As shown in Figure 5,  
 the dataset is geographically well-distributed, cov-  
 ering various countries across the globe.

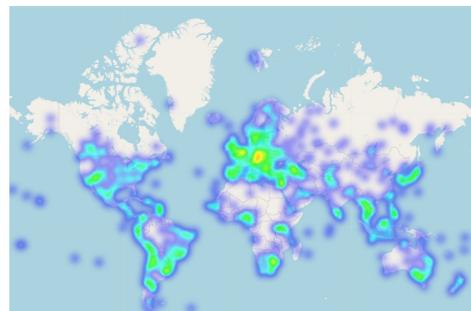


Figure 5: Location distribution of NAVICLUES, cover-  
 ing a wide range of countries around the world.

**Guidebooks.** As shown in Table 10, each guide-  
 book entry consists of an image-text pair, where  
 the text describes the location and features depicted  
 in the image. We utilize RAG in NAVIG to identify  
 the most similar image and use the accompanying  
 text to support the reasoning process.

**Statistical Results.** We present the statistical  
 results of NAVICLUES in Figure 6. Most distances  
 are relatively small, indicating that players are often  
 successful in deducing the correct location. The  
 reasoning process averages 842 words in length,  
 demonstrating high quality with extensive details.

**C Metrics Definition**

**C.1 GeoGuessr Score**

The *GeoGuessr Score* is a metrics introduced in  
 the GeoGuessr game to quantify the accuracy of a  
 guess. It is defined as:

$$\text{GeoGuessr Score} = 5000 \times e^{-\frac{d}{1492.7}} \quad (1)$$

where  $d$  represents the distance between the  
 guessed coordinates and the actual coordinates in  
 kilometers. Therefore, a more accurate guess yields  
 a *GeoGuessr Score* closer to 5,000.

---

**Image**



---

**Location**

Chile. -27.1265479, -109.2876917

---

**Reasoning**

These imposing stone figures are situated on a grassy hillside overlooking the ocean, suggesting a coastal location. The sparse vegetation and the presence of unique megalithic structures indicate a historical and possibly isolated environment. The statues, known as Moai, are characteristic of an island rich in archaeological history. Such a distinct culture is associated with an island famous for these figures, which is set far away from mainland destinations. This points toward the Polynesian island in the southeastern Pacific known for these world-renowned statues, aligning closely with a location associated with Chile.

---

**Image**



---

**Location**

China. 22.27992557066081, 114.1648415981852

---

**Reasoning**

Tall, sleek skyscrapers dominate the cityscape, suggesting a prominent urban environment. The distinct yellow license plates on vehicles are a key hint, typically associated with regions influenced by British rule. The street signs feature both English and Chinese characters, which points to a bilingual city. The presence of a distinctive red double-decker bus is highly characteristic of a city with British influence. All these clues, combined with the modern architecture and dense urban vibes, strongly suggest this is Hong Kong, likely within its central business district.

---

**Image**



---

**Location**

Italy. 42.71658113303754, 13.01051150781922

---

**Reasoning**

The landscape features rolling hills and a dense cover of greenery, indicative of a temperate climate typically found in parts of Southern Europe. The road layout with its narrow curve and the types of vehicles suggest a European setting. The presence of Italian-language signage such as Cascia and recognizable Italian road markers, like narrow front plates, strongly points to Italy. The architecture, with its rustic and modest buildings in the distance, complements the rural Italian countryside vibe. Hence, these visual cues collectively affirm the location as Italy, likely in a more central to southern region given the sign for Cascia.

---

Table 9: Demonstration of NAVICLUES.

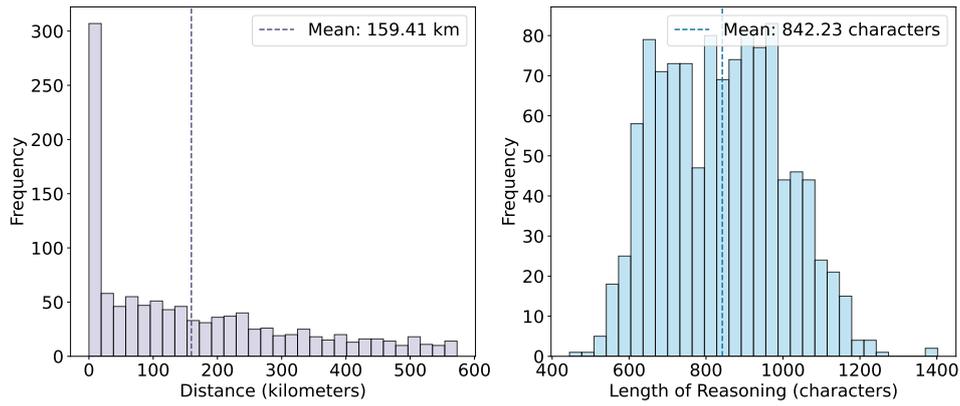


Figure 6: Distribution of Distance (left) and Length of Reasoning (right) in NAVICLUES.

**Image**



**Text**

(Toptips) Houses in Brittany, a western region of France, are coloured white with dark roofs.

**Image**



**Text**

(Plonkit) Tunisia has a fairly unique stop sign with Arabic and Latin script.

Table 10: Examples of clues in the two guidebooks, Toptips and Plonkit. The data is constructed as image-text pairs in the guidebooks.

**Image**



**Location**

Australia, Benalla. -36.5510941, 145.9863798

**Query**

<image> Given an image, craft a brief and cohesive reasoning path that deduces this location based on the visual clues present in the image. Using a tone of exploration and inference. Carefully analyze and link observations of natural features (climate, vegetation, terrain), man-made structures (roads, buildings, signage), and distinct landmarks. Allow these observations to naturally lead you to the correct country, enhancing the accuracy of your deductions. Start the reasoning without any intro, and make sure to make it brief.

**Response**

The scene displays a **bright, sunny day**, indicating a temperate climate. **the trees** lining the streets suggest a region with sufficient rainfall, the architecture shows low-rise buildings, typical of smaller towns rather than large urban centers, there are bilingual (or unilingual) **English signs**, helping to narrow down the location to an English-speaking country, **road markings and signs** indicate that **driving is on the left**, pointing towards countries like Australia or New Zealand, the presence of both cars and greenery suggests a suburban or rural setting, notable artwork on a building wall adds a cultural touch, indicating community engagement with public art, the overall scene feels relaxed, supporting the idea of a smaller, laid-back town such as those found in regional Australia, the specific layout and style are consistent with towns in Victoria, Australia, further narrowing down the location to Benalla in Victoria, given the visual and cultural context presented.

Table 11: Fine-tuning data format. Note how the reasoning process leverages visual information in the images to deduce the correct location, such as **weather**, **vegetation**, **signage**, and **driving orientation**.

Model	Continent 2, 500 km	Country 750 km	Region 200 km	City 25 km	Street 1 km	Distance↓	Score↑
$G^3$	47.3	23.9	6.0	1.6	0.0	4,938	1,451
GeoCLIP	82.3	66.5	48.0	<b>32.2</b>	<b>13.0</b>	1,762	3,402
StreetCLIP	68.2	51.2	29.6	19.0	4.2	3,161	2,640
MiniCPM-V	33.2	27.8	22.4	15.9	2.3	6,624	1,433
LLaVA	61.2	43.2	25.9	16.5	2.6	3,387	2,338
Qwen2-VL	75.0	65.0	48.9	29.9	5.3	2,483	3,237
<b>NAVIG</b>							
- MiniCPM-V	68.5	51.7	36.5	23.1	3.0	3,149	2,726
- LLaVA	70.4	47.8	26.8	16.7	2.8	2,851	2,592
- Qwen2-VL	<b>84.0</b>	<b>68.3</b>	<b>49.1</b>	28.9	5.5	<b>1,631</b>	<b>3,482</b>

Table 12: Performance on Im2GPS3k.

## C.2 Haversine Distance

We calculate the Haversine Distance of the models with the following formulas:

$$\Delta = \sqrt{\sin^2\left(\frac{\delta_{\text{lat}}}{2}\right) + \cos(\text{lat}_{\text{cor}})\cos(\text{lat}_{\text{pred}})\sin^2\left(\frac{\delta_{\text{lon}}}{2}\right)} \quad (2)$$

$$d = 2r \cdot \arcsin(\Delta) \quad (3)$$

where:

- $r$  is the Earth’s radius, which we set as 6,371,
- $\delta_{\text{lat}}$  is the difference in latitude between the true and predicted coordinates,
- $\delta_{\text{lon}}$  is the difference in longitude between the true and predicted coordinates,
- $\text{lat}_{\text{cor}}$  and  $\text{lon}_{\text{cor}}$  are the correct coordinates,
- $\text{lat}_{\text{pred}}$  and  $\text{lon}_{\text{pred}}$  are the predicted coordinates.

## D Supplementary Experiments

In this section, we present supplementary experiments, including results from the experiments on Im2GPS3k, and SEARCHER details.

As shown in Table 12, NAVIG outperforms prior models on Im2GPS3k in terms of Average Distance and GeoGuessr Score. However, GeoCLIP achieves better performance at the City and Street level, likely due to its training on coordinates datasets. The ablation results demonstrated in Table 13 are consistent with those in Table 5.

We also analyze the usage of each tool across the datasets and the number of grounding images. This analysis illustrates how frequently NAVIG leverages each tool and image detail to deduce locations.

Model	Country	City	Street
<b>NAVIG (MiniCPM-V)</b>	<b>51.7</b>	<b>23.1</b>	<b>3.0</b>
- w/o training	- 1.6	- 1.8	- 0.1
- w/o REASONER	- 10.6	- 3.5	- 0.2
- w/o SEARCHER	- 0.3	- 0.2	- 0.0
- MiniCPM-V	- 23.9	- 7.2	- 0.7
<b>NAVIG (LLaVA)</b>	47.8	<b>16.7</b>	<b>2.8</b>
- w/o training	- 15.3	- 4.7	- 0.8
- w/o REASONER	- 8.1	- 1.3	- 0.1
- w/o SEARCHER	<b>+ 0.1</b>	- 0.2	- 0.2
- LLaVA	- 4.5	- 0.2	- 0.1
<b>NAVIG (Qwen2-VL)</b>	68.3	28.9	<b>5.5</b>
- w/o training	- 4.3	- 1.2	- 0.3
- w/o REASONER	- 2.9	+ 0.5	- 0.1
- w/o SEARCHER	<b>+ 0.1</b>	- 0.0	- 0.2
- Qwen2-VL	- 3.3	<b>+ 1.0</b>	- 0.2

Table 13: Ablation results on Im2GPS3k.

As shown in Table 14 and Table 15, houses are the most frequently identified items in the testing dataset, as images often contain multiple houses. In contrast, signs, though less common, play a critical role by generating queries for OSM. The distribution of items directly influences the frequency of tool usage for knowledge retrieval.

Dataset	house	road sign	building sign
GWS5k	3,451	20	104
Im2GPS3k	465	52	24

Table 14: The frequency of how each item is grounded.

Dataset	$N$	RAG	MAP	VLN
GWS5k	5,000	128	70	1,978
Im2GPS3k	2,997	213	21	493

Table 15: The usage of each tool in each dataset, where  $N$  denotes the size of the dataset.