Evaluating Precise Geolocation Inference Capabilities of Vision Language Models

Anonymous submission

Abstract

The prevalence of Vision-Language Models (VLMs) raises important questions about privacy in an era where visual information is increasingly available. While foundation VLMs demonstrate broad knowledge and learned capabilities, we specifically investigate their ability to infer geographic location from previously unseen image data. This paper introduces a benchmark dataset collected from Google Street View that represents its global distribution of coverage. Foundation models are evaluated on single-image geolocation inference, with many achieving median distance errors of $<$ 300 km. We further evaluate VLM "agents" with access to supplemental tools, observing up to a 38.2% decrease in distance error. Our findings establish that modern foundation VLMs can act as powerful image geolocation tools, without being specifically trained for this task. When coupled with increasing accessibility of these models, our findings have greater implications for online privacy. We discuss these risks, as well as future work in this area. [1]

Introduction

With the proliferation of user-data collecting services on the Internet, it seems that privacy is becoming increasingly scarce (Jiang et al. 2022). This issue has only been exacerbated by the advent of Large Language Models (LLMs) and multi-modal Vision-Language Models (VLMs). Research has demonstrated AI's ability to infer user demographics from both text (Staab et al. 2024) and social media images (Tömekçe et al. 2024). Inspired by the popular geolocation game, GeoGuessr, we investigate whether this capability extends to precise location inference of everyday images.

Geolocation inference refers to the ability to determine exact geographical coordinates (latitude and longitude) given an image or series of images (Haas et al. 2024). While research in this field has advanced substantially over the past decade, it still remains a challenging problem. Previous works built custom datasets and architectures that focus on specific regions (cities or countries) and fail to generalize beyond distribution shifts in testing (Suresh, Chodosh, and Abello 2018; Wu and Huang 2022; Berton, Masone, and Caputo 2022; Clark et al. 2023). A recent development is the PIGEON architecture (Haas et al. 2024), achieved remarkable global geolocation performance. It is notable that the PIGEON authors have not released the model's training data and weights, citing the ethical risks of public use. Our work demonstrates that similar capabilities can be achieved through foundation VLMs and limited scaffolding.

While prior studies (Wang et al. 2024; Zhang et al. 2023) have demonstrated geolocation in VLMs to some extent, they limit their evaluation to country classification, rather than precise geographical coordinates. In addition, they fail to test the full potential of VLM capability. Compared to custom architectures, foundation models have shown impressive general reasoning and common sense (Lu et al. 2024), allowing them to use tools and act as autonomous agents. General VLM benchmarks have considered the evaluation of these agents (Li et al. 2024), but haven't explored them in the context of geolocation.

To address these research gaps, we collect a dataset of images from Google Street View to benchmark image geolocation ability. An evaluation of single-image location inference is performed on popular VLM families. Finally, we test if enabling model tool-use improves accuracy.

The main contributions of this paper are as follows:

- **Comprehensive dataset for geolocation:** Our dataset contains 1602 images taken from Google Street View. The images depict various levels of urbanization and represent a diverse set of countries where Street View data is available. Each image is labeled with exact latitude, longitude, and API parameters. The addition of this metadata means images can be dynamically updated or fetched for agentic purposes.
- **Base model benchmarks:** We present benchmarks of geolocation on single images for popular foundation VLMs.
- **VLM agent benchmarks:** We further test VLM capabilities by giving foundation models access to tools such as Street View or Google Lens.

Methodology

Dataset

We aim to capture a set of Google Street View images with geographical and urban diversity. Up to 200 unique cities

---

[1] Datasets and code can be found at: https://anonymous.4open.science/r/location-inference-1611

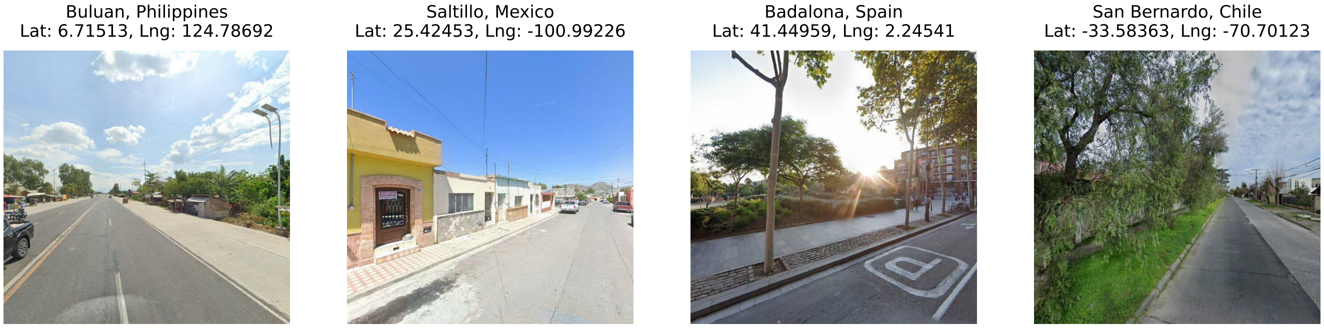

Buluan, Philippines
Lat: 6.71513, Lng: 124.78692
Saltillo, Mexico
Lat: 25.42453, Lng: -100.99226
Badalona, Spain
Lat: 41.44959, Lng: 2.24541
San Bernardo, Chile
Lat: -33.58363, Lng: -70.70123

Figure 1: A snapshot of Google Street View images in the full benchmark dataset.

(a) Population Distribution

Class	Population Range	Count
Small Urban	50,000–200,000	792
Medium Urban	200,000–500,000	601
Metropolitan	500,000–1.5M	344
Large Metropolitan	>1.5M	156

(b) Geographic Distribution

Continent	Count
North America	165
South America	112
Europe	95
Asia	91
Africa	24
Oceania	5

Table 1: Geographical and population distribution of full benchmark dataset.

are selected from every continent and population class (see Table 1). For each city, we attempt to fetch an image from the Google Street View Static API at a random point within a 10 km radius. Up to 5 attempts were made per city, after which the city is scrapped. The final dataset consists of 1602 images from 110 unique countries, representative of global Street View coverage. Table 1 contains details regarding full dataset distribution. A smaller subset of 319 images is also derived by sampling images from all 110 countries.

Images are labeled with the following: latitude, longitude, city, country, and API parameters used. The field-of-view is fixed to 90° and pitch (vertical camera rotation relative to a horizontal plane) is fixed to 0°. The heading (camera rotation relative to true north) is randomized across a uniform distribution from 0° to 360°. This ensures a diversity of views, as Google Street View images are typically taken on roads. Since Google Street View consists of a continuous set of billions of images stitched together, the chance of a model already seeing a particular view is near zero.

Model Evaluation

We benchmark a variety of large- and small-parameter VLMs on the full 1602 images dataset. The following families of models are tested: GPT, Gemini, Claude, and LLAVA.

To evaluate image geolocation capabilities, we construct a system prompt allowing models to assume the role of a competitive GeoGuessr player. Models are instructed to provide a Chain-of-Thought (CoT), describing the reasoning process and visual elements of interest in the image. Finally, the models provide a guess, containing predicted country, city, latitude, and longitude.

Distance error of predicted coordinates is scored in kilometers (km) using the Haversine formula:

$$d = R\,\mathrm{arctan2}(\sqrt{\mathrm{hav}\ \theta}, \sqrt{1 - \mathrm{hav}\ \theta})$$

where R is the Earth's mean radius and θ is the central angle of two points on a sphere. We define:

$$\mathrm{hav}\ \theta = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\left(\phi_1\right)\cos\left(\phi_2\right)\sin^2\left(\frac{\Delta\lambda}{2}\right)$$

where ϕ are the latitudes of both points and λ are the longitudes of both points.

Model CoTs are analyzed by tagging them with one of 12 categories. These categories are meant to encapsulate common visual elements that are used by the model to inform guesses. More information about the categories, along with their descriptions, can be found in Appendix A.

VLM Agents

Considering the versatile capabilities of foundational models and growing prevalence of AI agents, we evaluate the effects of additional scaffolding to the models. These agents are benchmarked on our 319 image data subset. The following two tool-use cases are implemented:

- **VLM + Street View:** Similar to the game GeoGuessr, the agent is able to "look around" the environment by requesting more images from the same camera location. This is done by giving the agent control over the "heading" and "pitch" parameters of the Google Street View Static API. The agent will make a prediction, choose new API parameters, and receive an additional Street View image. It can then refine its prediction by examining all images thus far. This process is evaluated over 5 iterations.

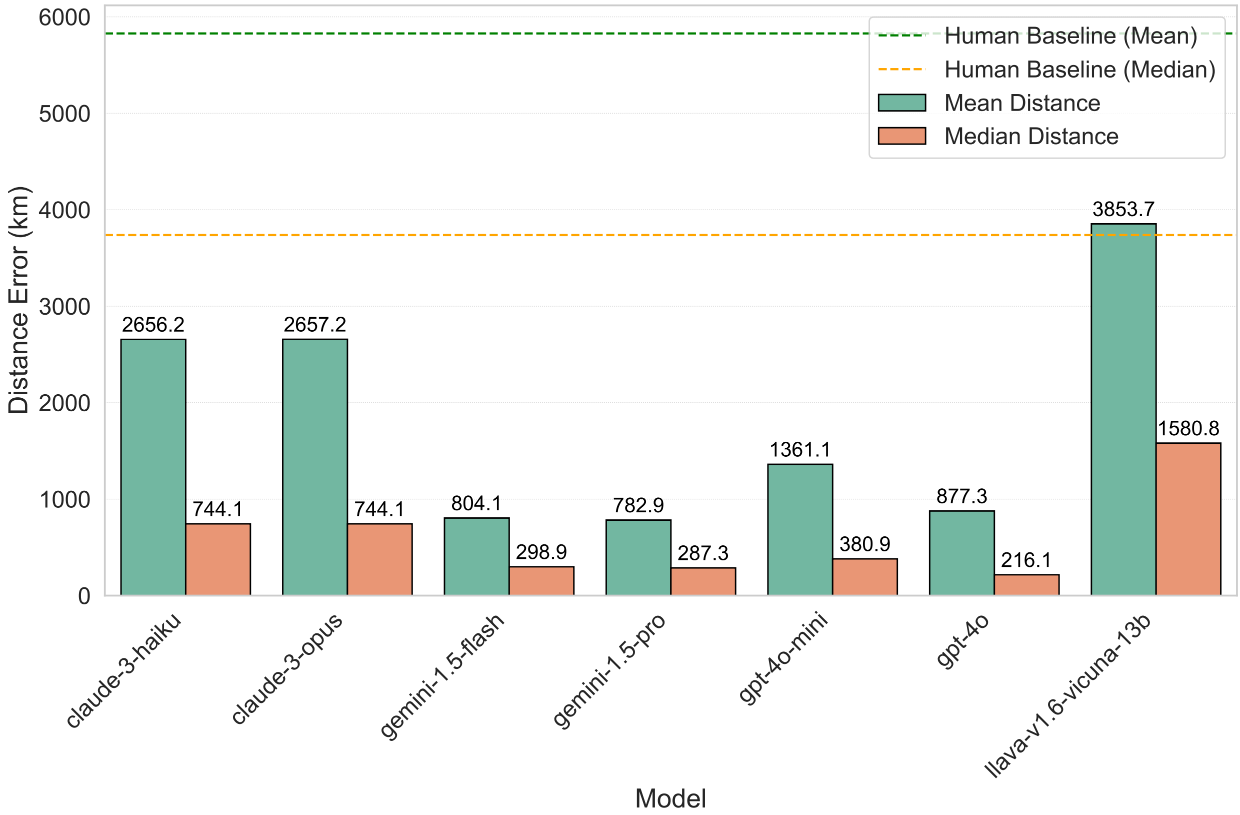


Figure 2: Overview of distance error for all benchmarked models and human baselines.

- **VLM + Google Lens:** The agent's guesses are augmented with reverse-image search via Google Lens. The agent will make a prediction, after which the top 10 results (each consisting of a thumbnail, title, and link) from the Google Lens API are provided. The agent can use this additional information to inform its second guess.

Human Baseline

To obtain a meaningful performance reference for both base-models and agents, we establish an average human baseline for image geolocation. We develop a web-based annotation interface that presents users with images from the full benchmark and Street View photospheres from the benchmark subset. Human baseline data was collected from 8 unique volunteers, including the authors.

Results

VLM Geolocation

Table 2 shows the accuracy in correctly guessing city/-country for all benchmarked models. GPT-4o performs best for city prediction, achieving an accuracy of 0.1199, while Gemini 1.5 Pro performs best for country prediction, achieving an accuracy of 0.8146. Figure 2 shows distance error for each model and human baselines. GPT-4o has the lowest median error, at 216.1 km. It is notable that mean values are significantly higher than median, indicating the large error magnitude of outliers.

Model	City Accuracy	Country Accuracy
claude-3-haiku	0.0293	0.5705
claude-3-opus	0.0293	0.5705
claude-3.5-sonnet-20240620	0.0687	0.6891
gemini-1.5-flash	0.0674	0.7341
gemini-1.5-pro	0.0899	**0.8146**
gpt-4o-mini	0.0774	0.7097
gpt-4o	**0.1199**	0.7603
llava-v1.6-vicuna-13b	0.0119	0.3689

Table 2: Country and city accuracy for base model image geolocation predictions.

Categories. For most visual elements, the model's category-specific performance is close to its overall median performance. More detailed analysis of category error can be found in Appendix A.

Agentic Geolocation

VLM + Street View. With access to Street View, an agent is able to move the camera view to areas where it believes it will get the most useful information. After retrieving up to 5 images, most agents have higher performance in all areas compared to the base model. Figure 3 shows distance

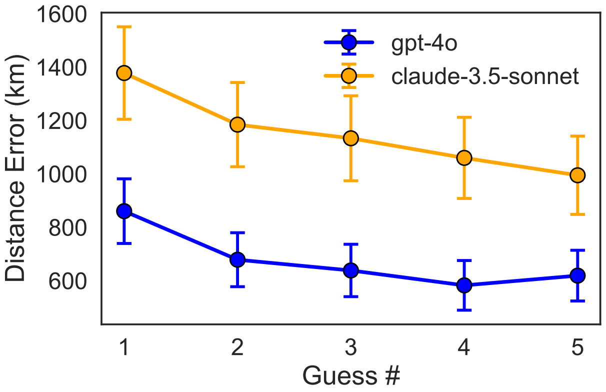


Figure 3: Mean distance error over 5 guesses for the VLM + Street View agent.

error change over guesses. By the 5th prediction, GPT-4o achieved a 38.2% decrease in median error, and Claude 3.5 Sonnet achieved a 35.2% decrease in median error. GPT-4o Mini, a small-parameter model, only had a 14.3% decrease in median error. Full statistics for all base models benchmarked can be found in Appendix B.

As shown in Figure 4, agent performance is still lower than the current state-of-the-art image geolocation architecture, PIGEON. However, in only 3 guesses, the agent is already able to surpass human GeoGuessr players at the "Champion Division" (Haas et al. 2024). This division includes the top 0.01% of competitive players. Our findings demonstrate that an agent with relatively simple and low-cost scaffolding can substantially improve geolocation ability and surpass experts.

VLM + Google Lens. With access to Google Lens, an agent would be able to get more context about specific items found in the image through reverse-image search. Table 3 shows accuracy of predictions made before and after Google Lens results. We observed that the informed guess has 87.1% higher median distance error and 100% lower city accuracy. We theorize that the large amount of noise and irrelevant search results negatively impacted the efficacy of the agent. The large gap between foundation VLMs and reverse-image search tools is highlighted for geolocation tasks.

Conclusion and Future Work

This work presents a novel benchmark of the precise geolocation capabilities of VLMs. We present a comprehensive dataset of images labeled with geographical coordinates, representative of a variety of geographical and urban environments around the world. Models are benchmarked on single-image inference, resulting in surprising levels of accuracy. Finally, we implement VLM agents that are able to access tools. In some cases, these agents substantially improved in accuracy with relatively simple scaffolding. With just 3 views, an agent was able to outperform the top tier of GeoGuessr experts.

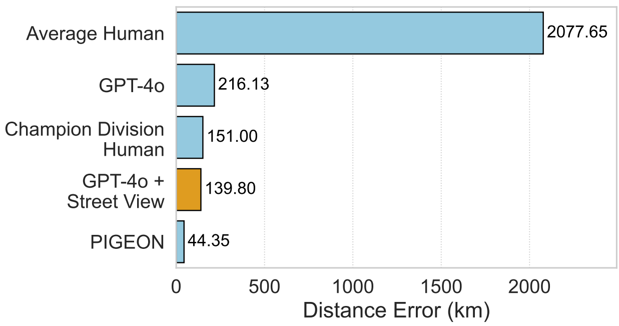


Figure 4: Comparison of geolocation performance for humans and AI. Both human baselines are collected with access to full 360 degree photospheres (Haas et al. 2024).

Guess	Avg Dist	Median Dist	Ctry Acc	City Acc
Initial	792.99	196.68	0.8037	0.1196
With Lens	1469.55	367.97	0.0368	0.0000

Table 3: VLM + Google Lens before and after retrieving top search results.

Some limitations of our work include limited representation of certain geographical areas. While images are representative of global Google Street View coverage, there is a lack of adequate imaging in regions such as Russia, China, and Africa. In addition, Street View has generally less availability in rural areas with limited human development. Another limitation is the reliability of CoTs collected from model responses. CoT reasoning is not always faithful (Lanham et al. 2023), and specific elements of the image that the model uses for its prediction may not be explicitly mentioned.

Our work indicates potential risks of privacy due to precise image geolocation capabilities in VLMs. In an era of global communication and social media, visual information about people and institutions is ubiquitous. We have demonstrated that readily-available commercial VLMs and simple, low-cost agents have the capability to obtain sensitive location information with just a few images. Malicious actors may not only exploit online media for doxxing, but also to infer demographics such as ethnicity, age, and voting patterns. (Kosinski, Stillwell, and Graepel 2013) These actors can range from lone criminals to governments seeking to perform surveillance or suppression of online speech. (Kaur et al. 2021; Gaffin 2012)

For future work, we look to implement more complex agents that infer location in realistic settings. For example, we can imagine an agent that scans dummy social media accounts for images most useful for geolocation . Evaluating these complex agents will lead to a better understanding of real-world threat models of image geolocation. Additionally, we hope this work can lead to AI safety and anonymization frameworks that successfully diminish model performance.

Appendix

A. Chain-of-Thought Categorization

When conducting benchmarks, we prompt models to provide a Chain-of-Thought (CoT) reasoning before guessing. This CoT contain information about various visual elements within the image that are used to inform the guess. We define the following 12 categories of visual elements:

- *Road and infrastructure*: Details related to roads, infrastructure on roads, pavements, or sidewalks.
- *Urban layout and elements*: Features related to street layout, building density, urban planning, etc.
- *Signage*: Traffic signs, shop signs, billboards, etc.
- *Architecture*: Buildings, structures, materials, architectural styles, etc.
- *Traffic and vehicles*: Types of vehicles, license plates, car models, traffic patterns, utility vehicles, etc.
- *Vegetation*: Plants, trees, etc.
- *Environment and climate*: Sky, weather, landscape features, terrain, etc.
- *Lighting and shadows*: Used to guess hemisphere, time of day, season, etc.
- *Recognizable landmarks*: Specific, identifiable places or structures.
- *Language*: Text on signs, buildings, etc.
- *Other cultural elements*: Clothing, festivals, customs, etc. (not including language).
- *Other*: Any other details that don't fit the above categories.

All model CoTs are passed into an annotator, which tags them with any number of the above categories. The annotator model used was Claude 3.5 Sonnet. Sample CoTs and respective category tags can be found in Figure 5.

For GPT-4o, categories that were most often used included: road and infrastructure, urban layout and elements, architecture, and vegetation. Categories used the least often included: lighting and shadows, recognizable landmarks, and cultural elements. Figure 6 depicts category-specific relative distance error for GPT-4o (well performing model), Claude 3 Opus (poor performing model), and human baselines. GPT-4o performs relatively better for categories: signage, recognizable elements, and language. It also performs well on vegetation, which tends to be a commonly encountered but challenging to decipher.

B. VLM + Street View Results

Table 4 displays full results for "VLM + Street View" agent evaluation on three different base models.

References

Berton, G.; Masone, C.; and Caputo, B. 2022. Rethinking Visual Geo-localization for Large-Scale Applications. arXiv:2204.02287.

Clark, B.; Kerrigan, A.; Kulkarni, P. P.; Cepeda, V. V.; and Shah, M. 2023. Where We Are and What We're Looking At: Query Based Worldwide Image Geo-localization Using Hierarchies and Scenes. arXiv:2303.04249.

Gaffin, E. 2012. Friending Brandeis: Privacy and Government Surveillance in the Era of Social Media. *SSRN Electronic Journal*.

Haas, L.; Skreta, M.; Alberti, S.; and Finn, C. 2024. PIGEON: Predicting Image Geolocations. arXiv:2307.05845.

Jiang, Y.; Le, B. D.; Zia, T.; and Gauravaram, P. 2022. Privacy Concerns Raised by Pervasive User Data Collection From Cyberspace and Their Countermeasures. arXiv:2202.04313.

Kaur, P.; Dhir, A.; Tandon, A.; Alzeiby, E. A.; and Abohassan, A. A. 2021. A systematic literature review on cyberstalking. An analysis of past achievements and future promises. *Technological Forecasting and Social Change*, 163: 120426.

Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15): 5802–5805.

Lanham, T.; Chen, A.; Radhakrishnan, A.; Steiner, B.; Denison, C.; Hernandez, D.; Li, D.; Durmus, E.; Hubinger, E.; Kernion, J.; Lukošiūtė, K.; Nguyen, K.; Cheng, N.; Joseph, N.; Schiefer, N.; Rausch, O.; Larson, R.; McCandlish, S.; Kundu, S.; Kadavath, S.; Yang, S.; Henighan, T.; Maxwell, T.; Telleen-Lawton, T.; Hume, T.; Hatfield-Dodds, Z.; Kaplan, J.; Brauner, J.; Bowman, S. R.; and Perez, E. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702.

Li, J.; Lu, W.; Fei, H.; Luo, M.; Dai, M.; Xia, M.; Jin, Y.; Gan, Z.; Qi, D.; Fu, C.; Tai, Y.; Yang, W.; Wang, Y.; and Wang, C. 2024. A Survey on Benchmarks of Multimodal Large Language Models. arXiv:2408.08632.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. arXiv:2310.02255.

Staab, R.; Vero, M.; Balunović, M.; and Vechev, M. 2024. Beyond Memorization: Violating Privacy Via Inference with Large Language Models. arXiv:2310.07298.

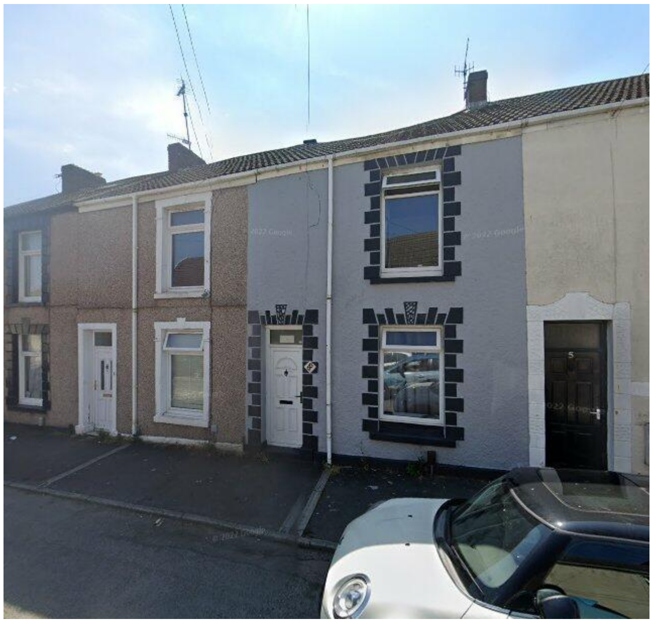
Suresh, S.; Chodosh, N.; and Abello, M. 2018. DeepGeo: Photo Localization with Deep Neural Network. arXiv:1810.03077.

Tömekçe, B.; Vero, M.; Staab, R.; and Vechev, M. 2024. Private Attribute Inference from Images with Vision-Language Models. arXiv:2404.10618.

Wang, Z.; Xu, D.; Khan, R. M. S.; Lin, Y.; Fan, Z.; and Zhu, X. 2024. LLMGeo: Benchmarking Large Language Models on Image Geolocation In-the-wild. arXiv:2405.20363.

Wu, M.; and Huang, Q. 2022. IM2City: image geo-localization via multi-modal learning. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '22, 50–61. New York, NY, USA: Association for Computing Machinery. ISBN 9781450395328.

Zhang, G.; Zhang, Y.; Zhang, K.; and Tresp, V. 2023. Can Vision-Language Models be a Good Guesser? Exploring VLMs for Times and Location Reasoning. arXiv:2307.06166.



Example 1 - Abertawe, United Kingdom

Model CoT:

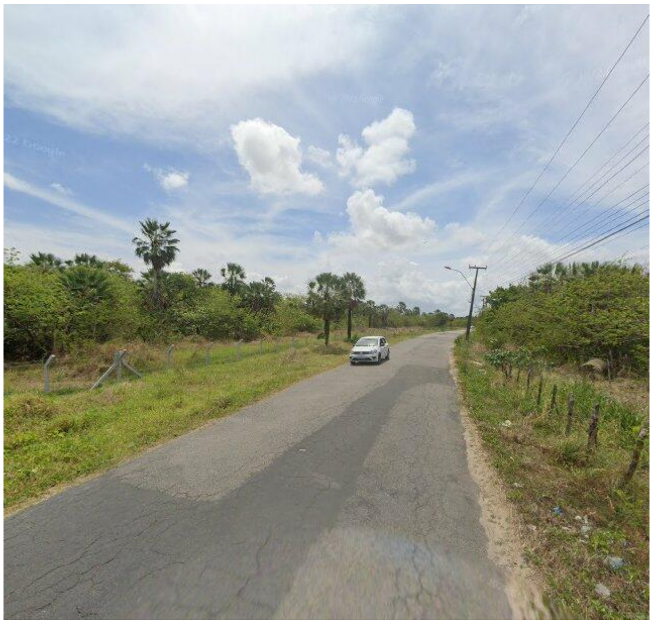
To determine the location of this image, let’s analyze the visual clues:

1. **Architecture:** The terraced houses with brick and stone detailing are typical of the UK, particularly in Wales and Northern England. The style is consistent with working-class housing from the late 19th to early 20th century.
2. **Car:** The presence of a Mini Cooper, a popular car in the UK, supports the idea that this is a British location.
3. **Street:** The narrow street and lack of front gardens are common in older urban areas in the UK.
4. **Weather and Lighting:** The lighting suggests a temperate climate, which aligns with the UK.
5. **House Numbers:** The numbering style and door designs are typical of British homes.

Based on these observations, the location is likely in a residential area in Wales, possibly in a town like Merthyr Tydfil or Swansea, where such housing is common.

Here’s my best guess for the coordinates...

Tagged Categories: Road and infrastructure, Urban layout and elements, Architecture, Traffic and vehicles



Example 2 - Caucaia, Brazil

Model CoT:

To determine the location of this image, let’s analyze the visual clues:

1. **Vegetation:** The presence of palm trees and the type of vegetation suggest a tropical or subtropical climate, likely in a region close to the equator.
2. **Road and Infrastructure:** The road is paved but appears to be in a rural area. The power lines and the style of the poles are common in many parts of Latin America and Southeast Asia.
3. **Sky and Weather:** The sky is clear with some clouds, indicating a warm climate.
4. **Car:** The car is driving on the right side of the road, which is common in most countries, but it helps rule out places like the UK, Australia, and Japan.
5. **Fencing and Landscape:** The fencing and open landscape suggest agricultural or undeveloped land.

Based on these observations, a likely region is northeastern Brazil, which fits the climate, vegetation, and road style.

Here’s my best guess for the coordinates...

Tagged Categories: Road and infrastructure, Traffic and vehicles, Vegetation, Environment and climate

Figure 5: Sample GPT-4o CoTs from full benchmark images. In both examples, the model predicts within 25 km of actual location.

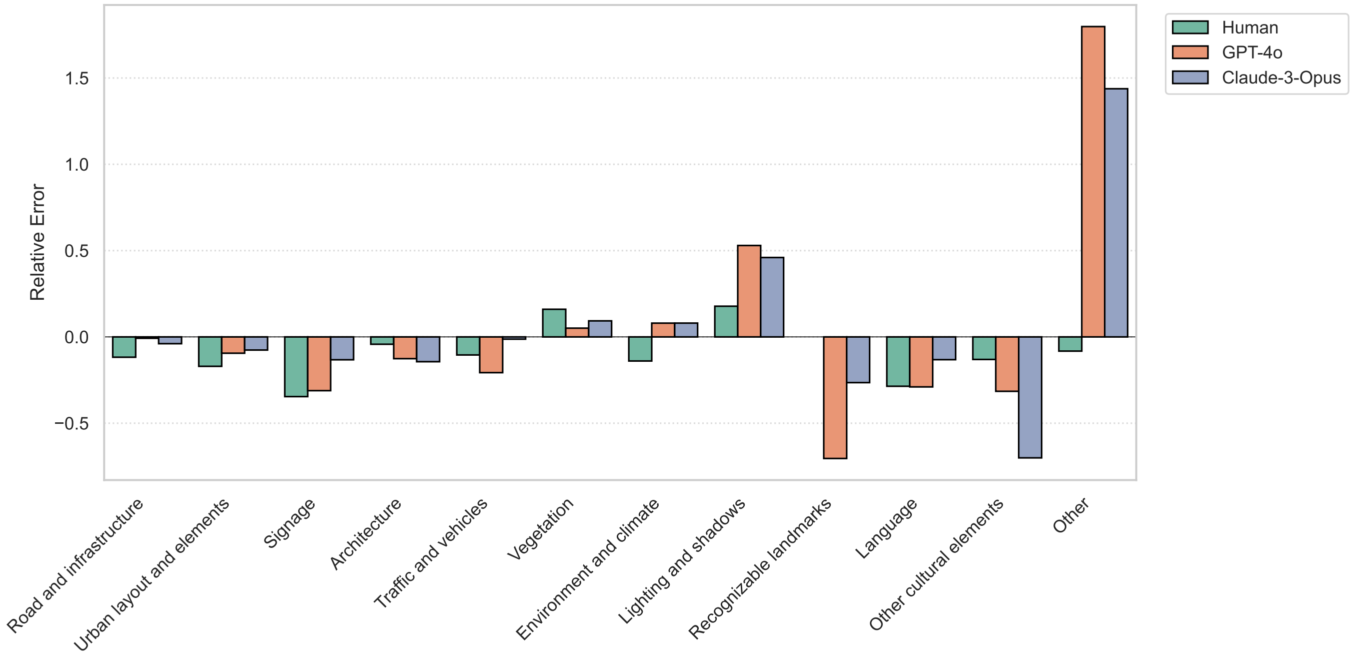


Figure 6: Relative distance error for each visual category. Relative error is calculated by dividing the difference of category-specific mean and overall median by overall mean.

(a) gpt-4o

Guess #	Avg Dist (km)	Median Dist (km)	Country Acc	City Acc
1	860.60	226.30	0.7642	0.1069
2	678.68	170.85	0.7862	0.1415
3	638.41	150.84	0.7830	0.1447
4	**582.65**	**127.29**	**0.7893**	**0.1572**
5	619.09	139.80	0.7830	0.1572

(b) gpt-4o-mini

Guess #	Avg Dist (km)	Median Dist (km)	Country Acc	City Acc
1	1527.64	443.51	0.7086	0.0397
2	**1308.64**	**433.82**	**0.7285**	0.0397
3	1308.53	433.89	0.7285	0.0397
4	1308.68	433.96	0.7285	0.0397
5	1308.64	434.03	0.7285	0.0397

(c) claude-3.5-sonnet-20240620

Guess #	Avg Dist (km)	Median Dist (km)	Country Acc	City Acc
1	1377.92	394.28	0.6047	0.0676
2	1184.61	343.12	0.6453	0.1115
3	1133.54	313.25	0.6757	0.0845
4	1060.09	269.30	**0.6858**	0.1115
5	**995.00**	**255.40**	0.6824	**0.1216**

Table 4: Full statistics for VLM + Street View agent.