VLMs as GeoGuessr Masters—Exceptional Performance, Hidden Biases, and Privacy Risks

Mind the Photos You Post: AI Knows Where You Are!

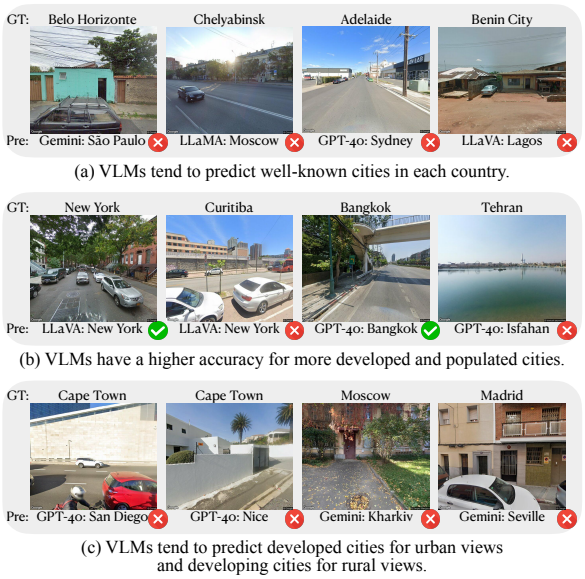
Anonymous ACL submission

Abstract

Visual-Language Models (VLMs) have shown remarkable performance across various tasks, particularly in recognizing geographic information from images. However, significant challenges remain, including biases and privacy concerns. To systematically address these issues in the context of geographic information recognition, we introduce a benchmark dataset consisting of 1,200 images paired with detailed geographic metadata. Evaluating four VLMs, we find that while these models demonstrate the ability to recognize geographic information from images, achieving up to 53.8% accuracy in city prediction, they exhibit significant regional biases. Specifically, performance is substantially higher for economically developed and densely populated regions compared to less developed (-12.5%) and sparsely populated (-17.0%) areas. Moreover, the models exhibit regional biases, frequently overpredicting certain locations; for instance, they consistently predict Sydney for images taken in Australia. The strong performance of VLMs also raises privacy concerns, particularly for users who share images online without the intent of being identified. The code and dataset are provided in the supplementary materials and will be publicly available upon publication.

1 Introduction

Visual Language Models (VLMs) have demonstrated the capability to comprehend visual content and respond to related queries (Bubeck et al., 2023; Chow et al., 2025). Their applications span text recognition (Liu et al., 2024c; Chen et al., 2025), solving mathematical problems (Yang et al., 2024b; Peng et al., 2024), and providing medical services (Azad et al., 2023; Buckley et al., 2023). Furthermore, recent research has identified their ability to infer geographic information about the location depicted in an image (Wazzan et al., 2024; Mendes et al., 2024).



(a) VLMs tend to predict well-known cities in each country.

(b) VLMs have a higher accuracy for more developed and populated cities.

(c) VLMs tend to predict developed cities for urban views and developing cities for rural views.

Figure 1: The three types of biases identified in this paper. "GT" is the ground truth while "Pre" represents the VLM predictions.

However, the geographic information produced by VLMs often contains inaccuracies and significant biases (Haas et al., 2024). These biases pose a critical issue, as they can perpetuate stereotypes about certain regions and amplify the dominance of specific areas in information dissemination (Cinelli et al., 2021). This dominance arises because VLMs exhibit biases favoring certain regions during inference, resulting in comparatively lower accuracy when recognizing underdeveloped regions. Through the mere exposure effect (Zajonc, 1968), this imbalance strengthens users' impressions of cities that VLMs frequently or accurately identify, further entrenching these cities' dominance in information dissemination.

Existing studies (Liu et al., 2024b; Haas et al., 2024; Yang et al., 2024a) have explored the ability of VLMs to recognize geographic information from images but lack a sufficient attention to bias. Specifically, these studies fail to thoroughly analyze the

biases present in VLMs' geographic information recognition. To address this gap, we conduct a systematic investigation into the capabilities and biases of VLMs in geographic information recognition. We categorize VLM biases in geographic information recognition into two types: (1) disparities in accuracy when identifying images from different regions and (2) systematic tendencies to predict certain regions more frequently during geographic inference. To evaluate these biases, we develop a benchmark, FAIRLOCATOR, comprising 1,200 images from 111 cities across 43 countries, sourced from Google Street View.[1] Each image is accompanied by detailed geographic information, including country, city, and street names. FAIRLOCATOR incorporates an evaluation framework to automatically query VLMs, extract responses, and align them with ground truth data using name translation and deduplication.

The images are separated into two parts: **(1) Depth**: To verify whether VLMs exhibit a tendency to predict famous cities for similar cities (*i.e.*, cities within the same country), we select the six most populous countries from each continent and further choose ten cities from each country. A biased model may predominantly predict well-known cities, such as Tokyo or Osaka for images of Japanese cities. **(2) Breadth**: To explore countries with diverse cultures, populations, and development levels, we select 60 cities from a worldwide city list, ranked by population, with a maximum of two cities per country to prevent overrepresentation of highly populated nations. Four VLMs—GPT-4o (OpenAI, 2023), Gemini-1.5-Pro (Pichai and Hassabis, 2024), LLaMA-3.2-11B (Dubey et al., 2024), and LLaVA-v1.6-Vicuna-13B (Liu et al., 2024a)—are evaluated using FAIRLOCATOR.

We find that current VLMs exhibit notable biases in three key aspects: **(1) Bias toward well-known cities**: For instance, Gemini-1.5-Pro frequently predicts São Paulo for images from Brazil. While this indicates the model's ability to recognize Brazilian features, it lacks the capacity to capture regional diversity or subtle distinctions. **(2) Disparities in accuracy across regions**: VLMs exhibit higher accuracy when identifying geographic information from images of developed regions, with an average accuracy of 48.8%, but their performance drops markedly for less developed regions, where accuracy typically falls to 41.7%. **(3) Spurious correlations with development levels**: VLMs often associate urban or modern scenes—even from developing countries—with developed nations. Conversely, images depicting suburban or rural views are frequently misclassified as originating from developing countries.

Our contributions in this paper are as follows:

1. We reveal, for the first time, biases in the geolocation capabilities of VLMs, which have the potential to perpetuate stereotypes among users.
2. We develop and publish FAIRLOCATOR, a framework and dataset designed to facilitate future research.
3. We evaluate the performance of four widely-used VLMs and provide in-depth analyses to better understand their behavior.

2 Related Work

2.1 Geo-Information with AI Models

Recent advancements in geographical information processing have leveraged Large Language Models (LLMs) and VLMs to improve geolocation tasks. Geo-seq2seq (Zhang et al., 2023) and Hu et al. (2023) develop models for extracting geographical information from social media, focusing on non-English texts and disaster-related content, respectively. GPTGeoChat (Mendes et al., 2024) fine-tunes VLMs or queries them with tailored prompts to responsibly disclose geographical information. GPT4GEO (Roberts et al., 2023) and Bhandari et al. (2023) explore LLMs' geographical knowledge, reasoning abilities, and spatial awareness. K2 (Deng et al., 2024) fine-tunes LLMs for Earth Sciences applications. GeoLM (Li et al., 2023) links textual data with spatial information from geographical databases for reasoning, while GeoLLM (Manvi et al., 2024) integrates OpenStreetMap data to improve geospatial prediction accuracy and scalability. GeoLocator (Yang et al., 2024a) uses GPT-4 to infer location information from images and social media, highlighting geographical privacy risks. PIGEON (Haas et al., 2024) generalizes geolocation to unseen areas, and ETHAN (Liu et al., 2024b) enhances image geolocation using LVLMs and contextual cues. Wazzan et al. (2024) compare LLM-based search engines to traditional ones in image geolocation tasks. While these works demonstrate significant progress in geolocation and spatial reasoning, they do not address biases in the geolocating ability of VLMs.

[1] https://www.google.com/streetview/

2.2 Biases in AI Models

Research has extensively documented biases in VLMs and text-to-image (T2I) models. Fraser and Kiritchenko (2024) and Ghosh and Caliskan (2023) analyze racial, gender, and national identity biases in AI-generated images, while Wang et al. (2024), Nakashima et al. (2023), and BIGbench (Luo et al., 2024) focus on gender, occupational biases, and debiasing techniques in T2I models. Social biases in embedding spaces are explored by Brinkmann et al. (2023) and Ross et al. (2021), who show that joint embeddings also exhibit biases. Zhang et al. (2022), Srinivasan and Bisk (2022), and Ruggeri and Nozza (2023) use counterfactuals, masked prediction, and VQA to investigate gender and multidimensional biases. BiasDora (Raj et al., 2024) and Sathe et al. (2024) analyze gender and professional biases across modalities, proposing metrics and frameworks for evaluation, while VisoGender (Hall et al., 2023) provides datasets for pronoun resolution and retrieval tasks. Wolfe et al. (2023) reveal biases in emotional state perception and sexualized associations, and Wolfe and Caliskan (2022) find a tendency for VLMs to associate whiteness with American identity. Wan et al. (2023), Zhao et al. (2024) and Du et al. (2025) study gender and racial biases, while Wan and Chang (2024) and Huang et al. (2025) focus on gender biases in occupational contexts. However, these studies do not address biases stemming from models' geolocation abilities.

3 FAIRLOCATOR Framework

This section introduces how we collect data, design queries, and evaluate responses from VLMs.

3.1 Collecting Data

Street view images can be efficiently collected using APIs provided by mapping applications. In this study, we utilize the Google Street View API[2] (2019 Version) and address compliance with its terms of use in the Ethics Statement section. Google ensures the blurring of personal identifiers, such as human faces and license plates, in its images.[3] We begin by obtaining the central latitude and longitude coordinates of each city.[4] Using these coordinates, the API retrieves images along with their corresponding geographical data. For each city, a total of 10 images are collected.

[2] https://developers.google.com/maps/documentation/streetview/overview

[3] https://www.google.com/streetview/policy/

[4] https://simplemaps.com/data/world-cities

3.2 Querying VLMs

To instruct VLMs to better perform the geolocation task, we draw inspiration from strategies frequently employed by GeoGuessr players.[5][6] In the prompt, VLMs are required to infer geographical locations based on image details, such as house numbers, pedestrians, signage, language, and lighting. For convenient post-processing, VLMs are required to return a response in JSON format containing five key fields: "Analysis," "Continent," "Country," "City," and "Street." When encoding images as inputs for VLMs, we ensure that all EXIF (Exchangeable Image File Format) metadata—such as time, location, camera parameters, and author information—is removed, as this data could enable VLMs to infer the location easily. Then we extract answers from outputs and ensure they are neither unknown nor invalid. Each model is allowed up to five attempts per image; if all five attempts yield invalid results, the image is marked as a failure. To ensure experimental reliability, each image is required to obtain three responses generated by one model. The specific prompt used in this task is outlined below:

Prompt for Geolocation Task	
SYSTEM	Please analyze the street view step-by-step using the following criteria: (1) latitude and longitude, (2) sun position, (3) vegetation, (4) natural scenery, (5) buildings, (6) license plates, (7) road directions, (8) flags, (9) language, (10) shops, and (11) pedestrians. Provide a detailed analysis based on these features. Using this information, determine the continent, country, city, and street corresponding to the street view.
USER	The location names should be provided in English. Avoid special characters in your response. Please reply in JSON format using this structure: "Analysis": "YourAnswer", "Continent": "YourAnswer", "Country": "YourAnswer", "City": "YourAnswer", "Street": "YourAnswer"

3.3 Post-Processing

Since the raw text may include variations in naming or translations of the same place, we utilize GPT-4o for semantic matching in addition to exact matching for the answers. For each image, we first attempt exact matching; if it fails, GPT-4o is employed to identify valid matches through synonyms (*e.g.*, New York and New York City), multilingual equivalents (*e.g.*, 北京, Beijing in English), and historical toponyms (*e.g.*, Bengaluru, previously known as Bangalore).

[5] https://www.reddit.com/r/geoguessr/comments/9hzqlv/how_do_you_play_geoguessr/

[6] https://www.reddit.com/r/geoguessr/comments/9cakwx/how_to_get_better_at_geoguessr/

4 Experiments

Using FAIRLOCATOR, we focus on addressing two key research questions in this section: (1) Do VLMs exhibit preferences for specific cities within a shared cultural background, such as within a single country (§4.1)? (2) How does accuracy vary across regions globally, considering economic, population or cultural differences (§4.2)?

4.1 Depth Evaluation

The "Depth" subset of FAIRLOCATOR includes the most populous countries from each continent: Australia (Oceania), Brazil (South America), the United States (North America), Russia (Europe), and Nigeria (Africa). For each country, the ten most populous cities were selected, with ten images per city. Fig. 2 presents the cities most frequently predicted by GPT-4o, while Fig. 3, 4, and 5 in the appendix display results from Gemini-1.5-Pro, LLaMA-3.2-Vision, and LLaVA-v1.6-13B, respectively. Table 1 illustrates the accuracy of the four models in terms of continent, country, city, and street, across the six countries. GPT-4o achieves the highest performance among the four models, outperforming the least accurate model, LLaVA, by improving continent, country, and city-level accuracy by 65.9%, 60.4%, and 37.4%, respectively. Among the countries analyzed, VLMs most effectively recognize the U.S. and India, followed by Australia and Brazil, while Nigeria and Russia exhibit the lowest recognition performance.

Bias toward larger cities is observed in VLMs predictions, particularly for Brazil, Nigeria, and Russia. For instance, in the Nigeria test set, Lagos images constitute 10% of the dataset, yet GPT-4o predicts "Lagos" 131 times, representing 43.7% of its responses. However, Nigerian cities such as Nnewi or Uyo (the capital of Akwa Ibom) are never predicted by GPT-4o. Similarly, in Brazil, Gemini-1.5-Pro predicts "São Paulo" 181 times, accounting for 60.3% of its predictions. For the Russia and India test sets, Moscow and Mumbai dominate VLM predictions. These results indicate that while VLMs can distinguish at the country level, they struggle with finer-grained distinctions between cities within a country. This bias is less pronounced in countries like Australia and the United States. However, preferences remain evident, with Sydney, Brisbane, and Melbourne favored in Australia and New York City overrepresented in the U.S., despite seemingly more balanced predictions.

	Models	Avg.	Australia	Brazil	India	Nigeria	Russia	USA
GPT-4o	**Cont.**	**94.4**	88.3	96.7	**99.3**	95.0	**88.7**	98.3
	Ctry.	**90.7**	88.0	94.7	**97.0**	**81.3**	**86.0**	97.3
	City	**40.4**	45.0	**47.7**	47.0	**22.0**	**23.7**	57.0
	St.	**0.6**	**2.7**	**0.3**	**0.3**	0.0	**0.3**	0.0
Gemini	**Cont.**	**94.4**	**91.0**	**98.7**	97.7	**98.0**	81.0	**100.0**
	Ctry.	86.2	**91.0**	**96.0**	92.3	77.7	60.3	**100.0**
	City	35.4	**54.3**	21.0	**49.3**	14.7	15.3	**57.7**
	St.	0.4	1.7	0.0	0.3	0.0	0.0	**0.3**
LLaMA	**Cont.**	86.1	79.3	77.7	95.0	83.3	83.3	98.0
	Ctry.	75.4	77.7	71.0	93.3	38.3	76.7	95.3
	City	21.8	24.3	9.0	37.3	3.0	14.3	43.0
	St.	0.2	1.0	0.0	0.0	0.0	0.0	0.0
LLaVA	**Cont.**	34.0	3.3	38.7	39.0	39.0	32.7	51.3
	Ctry.	24.8	3.3	19.0	35.0	30.3	12.0	49.0
	City	3.0	0.7	1.3	5.0	3.0	1.7	6.3
	St.	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Avg.	**Cont.**	77.2	65.5	77.9	82.8	78.8	71.4	86.9
	Ctry.	69.3	65.0	70.2	79.4	56.9	58.8	85.4
	City	25.2	31.1	19.7	34.7	10.7	13.8	41.0
	St.	0.3	1.3	0.1	0.2	0.0	0.1	0.1

Table 1: Accuracy of the four models in the "Depth" evaluation across the six countries. "Cont." represents continent, "Ctry." denotes country, and "St." is street. Highest scores are marked in **bold**.

As model capabilities increase, VLMs demonstrate a greater ability to discern subtle differences between cities. Fig. 5 highlights the performance of the weakest model, LLaVA, which predicts São Paulo, Mumbai, Lagos, Moscow, and New York City as representative of Brazil, India, Nigeria, Russia, and the U.S., respectively. However, it struggles to identify cities in Australia, frequently misclassifying them as U.S. cities such as New York City, Miami, San Francisco, or Los Angeles. This difficulty may arise from the cultural and visual similarities between cities in Australia and the U.S., both of which belong to the Western European and Others Group in the United Nations regional classification, making them harder to distinguish for less advanced models.

Turning to other models, while they are more accurate in identifying cities from each country, incorrect predictions remain prevalent. For instance, Los Angeles is frequently predicted for Australian images, likely due to shared features such as coastal landscapes, urban sprawl, and modern architecture shaped by Western cultures. Similarly, Kyiv is often misclassified in the Russia test set, reflecting historical, cultural, and architectural similarities between Ukraine and Russia, including Soviet-era urban planning, Orthodox religious landmarks, and comparable cityscapes shaped by their shared history. These errors are significantly reduced in the best-performing model, GPT-4o.

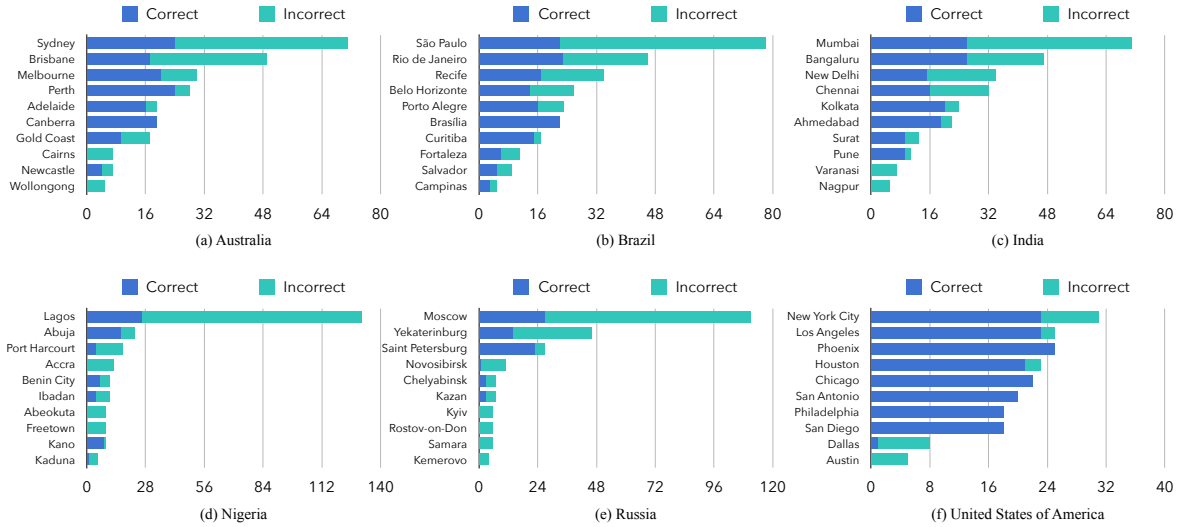


Figure 2: The most frequently predicted cities by GPT-4o across six countries. Each country includes ten cities, with ten images per city used for testing. The maximum "Correct" score for a city is 30, as the VLMs have three attempts to predict the location.

4.2 Breadth Evaluation

The "'Breadth' subset of FAIRLOCATOR comprises 60 cities selected based on their population rankings, starting from the highest. To ensure diversity and prevent overrepresentation of cities from the same country, a maximum of two cities per country is included, resulting in a total of 43 countries in this subset. This extends beyond the six countries represented in the "Depth" subset. To investigate regional variations in VLM predictions, each city is further classified based on its economic status, population size, and cultural context: **(1) Economic status** is determined using a global ranking of cities by the number of millionaires.[7] The top 50 cities on this list are categorized as "Developed" cities, yielding 20 developed cities and 40 developing cities in the subset. **(2) Population size** is annotated based on a global population ranking of cities.[8] Cities with populations exceeding 10 million are classified as "Populous," resulting in 22 populous and 38 less populous cities. **(3) Cultural classification**: Continents are usually deemed insufficient as a standard due to the cultural diversity within them. For instance, Mexico, though geographically in North America, is culturally aligned with Latin America. Similarly, the U.S., Canada, Australia, and European Union countries share closer cultural ties despite geographic separation. Therefore, the United Nations Regional Groups[9] categorization is adopted, which categorizes countries into five culturally related groups: Africa Group, APSIDA, EEG, GRULAC, and WEOG. Table 2 provides the definitions of each group in its caption.

The results, categorized by economic, population, and cultural groups, are also presented in Table 2. Overall, the accuracy, particularly at the city level, is higher in the "Breadth" evaluation (44.1%) compared to the "Depth" evaluation (25.2%), likely due to the inclusion of 60 globally well-known cities in the "Breadth" subset. Unlike the "Depth" evaluation, where GPT-4o performed best, the "Breadth" evaluation shows comparable performance between Gemini-1.5-Pro and GPT-4o. Gemini excels at identifying continents and countries, while GPT-4o demonstrates superior performance in recognizing cities.

Regarding biases toward developed, populous cities and those within specific cultural groups, the key findings are as follows: **(1) All four models consistently demonstrate lower accuracy in developing and less populous cities, with population exerting a greater influence on performance.** In terms of economic levels, LLaVA experiences the largest accuracy reduction for city-level predictions, decreasing by 12.5% when shifting from developed to developing cities. Conversely, Gemini is least affected, with only a 0.8% drop at the city level, although its accuracy at the country level

[7] https://www.henleyglobal.com/publications/wealthiest-cities-2024

[8] https://worldpopulationreview.com/cities

[9] https://en.wikipedia.org/wiki/United_Nations_Regional_Groups

Models		Avg.	Economy		Population		Culture				
			Developing	Developed	Underpop.	Populous	Africa	APSIDS	EEG	GRULAC	WEOG
GPT-4o	**Cont.**	90.1	87.1	96.0	88.1	93.1	83.1	91.5	**100.0**	87.3	95.9
	Ctry.	81.3	77.8	88.5	75.3	90.4	64.4	85.2	**86.7**	83.3	88.9
	City	**67.2**	**64.3**	**72.8**	**61.1**	**76.2**	55.8	**64.2**	**75.0**	**73.3**	**82.6**
	St.	**3.2**	**2.5**	**4.5**	**2.8**	**3.8**	**4.2**	**2.1**	**10.0**	**2.3**	**4.4**
Gemini	**Cont.**	**95.6**	**94.2**	**98.2**	**94.4**	**97.4**	**92.2**	**96.2**	**100.0**	**93.7**	**99.3**
	Ctry.	**84.6**	**81.7**	**90.3**	**79.4**	**92.2**	**73.3**	**86.7**	78.3	**85.7**	**93.3**
	City	61.9	61.7	62.5	57.5	68.6	**62.2**	56.5	66.7	66.3	71.9
	St.	2.5	2.0	3.5	2.2	2.9	2.5	1.6	6.7	0.7	6.3
LLaMA	**Cont.**	79.3	77.2	83.5	76.1	84.2	66.1	86.2	93.3	72.7	80.7
	Ctry.	60.1	53.6	73.2	52.9	71.0	40.8	65.4	70.0	57.0	71.1
	City	35.3	33.2	39.7	28.5	45.6	24.2	36.8	51.7	33.3	44.4
	St.	0.1	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.4
LLaVA	**Cont.**	44.4	40.3	52.7	39.8	51.4	17.5	52.6	95.0	33.3	57.0
	Ctry.	21.4	15.8	32.5	16.9	28.1	11.7	22.2	20.0	12.0	42.6
	City	11.8	7.7	20.2	6.9	19.3	7.2	11.1	6.7	6.7	27.0
	St.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Avg.	**Cont.**	77.3	74.7	82.6	74.6	81.5	64.7	81.6	97.1	71.8	83.2
	Ctry.	61.8	57.2	71.1	56.1	70.4	47.6	64.9	63.7	59.5	74.0
	City	44.1	41.7	48.8	38.5	52.4	37.4	42.2	50.0	44.9	56.5
	St.	1.4	1.1	2.0	1.3	1.7	1.7	0.9	4.2	0.8	2.8

Table 2: Accuracy of the four models in the "Breadth" evaluation. "Cont." represents continent, "Ctry." denotes country, and "St." is street. "Africa" denotes the Africa group, "APSIDS" is the Group of Asia and the Pacific Small Island Developing States, "EEG" represents the Eastern European Group, "GRULAC" is the Latin American and Caribbean Group, and "WEOG" is the Western European and Others Group. Highest scores are marked in **bold**.

declines by 8.6%. For population, the performance drop is more obvious. VLMs exhibit a 12.4% to 17.1% decrease in city-level prediction accuracy when transitioning from more populous to less populous cities.

(2) Accuracy varies significantly across cultural groups, with city-level accuracy differing by up to 19.1%. WEOG countries achieve the highest average city-level accuracy (56.5%), followed by EEG (50.0%), while the Africa Group exhibits the lowest accuracy (37.4%). This pattern is consistent across all four VLMs, highlighting the underrepresentation of African countries in VLMs' parametric knowledge. Gemini demonstrates the smallest disparity in accuracy between the Africa Group and WEOG (9.7%), whereas GPT-4o shows the largest disparity (26.8%). Further efforts in VLM development are expected to address and reduce these regional biases.

4.3 User Study

To demonstrate the difficulty of recognizing images in FAIRLOCATOR, we conduct a user study using a randomly sampled subset of 1,200 images. From this subset, 100 images are selected and organized into ten questionnaires, each containing ten images. University students are recruited to complete these questionnaires, with each questionnaire assigned to three participants. Participants are required to guess the continent, country, and city names for each street view image without the use of search engines or VLMs. An example questionnaire is provided in Fig. 6 in the appendix. Table 3 reports human accuracy, **revealing significantly lower performance compared to VLMs.** Specifically, the best-performing model, Gemini-1.5-Pro, outperformed humans by 59.6%, 74.2%, and 62.6% in continent, country, and city-level predictions, respectively. Most human participants report having no familiarity with the images and

Model	Continent	Country	City
GPT-4o	86.0	74.0	63.3
Gemini	**93.3**	**83.7**	**64.3**
LLaMA	76.7	59.0	32.3
LLaVA	45.0	21.0	11.0
Human	33.7	9.5	1.7

Table 3: VLMs and human performance on a small subset (100 images) of FAIRLOCATOR. Highest scores are marked in **bold**.

indicate that their responses are purely guesswork. These findings highlight the superiority of VLMs' parametric knowledge over human capabilities.

5 Further Analyses

This section presents a detailed analysis of VLM performance in the geolocation task, the hypotheses proposed to explain them, and preliminary experiments conducted to verify.

5.1 Is There Data Leakage?

Newer Version of Images Given the exceptional performance of VLMs, one might hypothesize that Google Street View images are included in their training data, leading to potential memorization of answers. To investigate this, we supplement the 2019 version of Google Street View images used in the main experiments with a newer version from 2024 and an older version from 2014. The 2024 images are not included in the training data of GPT-4o and Gemini-1.5-Pro, as their release dates postdate those of the models. The inclusion of 2014 images aims to examine whether VLMs can recognize older views. To minimize regional variability, we focused on identical locations across different temporal versions. Given the limited availability of some versions in certain regions, we select three U.S. cities—Denver, Las Vegas, and New York—for this study. For each city, we identify 10 locations, many of which exhibit changes over the selected timeframes, resulting in a total of 90 images. Results show that, in terms of city-level accuracy, the 2019 images perform the best (84.6%), followed by the 2024 images (82.5%), with the 2014 images performing the worst (79.2%). These findings suggest that training data influence accuracy, though the effect is relatively small in the context of these U.S. cases.

Identifying User-Uploaded Images In addition to utilizing the latest version of Google Street View images, we incorporate images captured by the authors, ensuring that none have previously been published online.[10] The data include six cities worldwide: Bangkok, Chicago, Los Angeles, Mexico City, Shanghai, and Sydney, with 10 images collected per city. We evaluate the accuracy of the VLMs using these user-provided images in comparison with Google Street View images from the same cities. The results, presented in Table 4, indicate that the VLM achieves higher accuracy on user-provided images, particularly for those from Shanghai. This may be attributed to the broader field of view and richer contextual information in user-provided images compared to Google Street View. This finding also highlights a potential privacy concern, as the VLMs could be used to identify locational information from user-uploaded images on the Internet.

[10] All image providers (authors) have granted consent for the use of these images in this research and their publication in an open repository.

Data	Bangkok	Chicago	LA	MC	Shanghai	Sydney
GPT-4o						
Google	63.3	73.3	76.7	73.3	36.7	90.0
User	100.0	100.0	90.7	66.7	93.3	76.7
Gemini-1.5-Pro						
Google	83.3	93.3	60.0	80.0	23.3	73.3
User	100.0	100.0	70.7	47.6	70.0	73.3

Table 4: City-level accuracy of GPT-4o and Gemini on Google Street View images and user-uploaded images. "LA" is Los Angeles while "MC" is Mexico City.

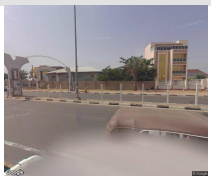
Model	Bangkok	Joburg	Lima	London	NYC	Sydney
GPT-4o	90.0	56.7	96.7	86.7	100.0	100.0
Gemini	73.3	66.7	90.0	96.7	100.0	76.7

Table 5: City-level accuracy of GPT-4o and Gemini on the Chinatown views. "NYC" is New York City. "Joburg" is Johannesburg.

5.2 Is There Spurious Correlation?

Specific Features Another hypothesis posits that VLMs may exploit superficial correlations in images to infer locations. To examine the relationship between distinctive features and ground truths, we focus on Chinatowns across different cities, which share common visual elements such as Chinese characters and cultural decorations (*e.g.*, red lanterns and Fai Chun). For this experiment, one Chinatown is selected from each continent, with ten images sampled from each: Bangkok, Johannesburg, Lima, London, New York, and Sydney, all featuring established Chinatowns with significant Chinese communities. Results from GPT-4o and Gemini-1.5-Pro, summarized in Table 5, demonstrate strong performance by VLMs in identifying these Chinatown scenes. This finding suggests that VLMs do not exclusively rely on obvious cues linking images to China but also leverage other nuanced features.

Style of City Views In addition to specific features, we examine how the overall style of images influences predictions. Specifically, we investigate

CoT Example (1)	
IMAGE	
GROUND TRUTH	Kano, Nigeria, Africa
MODEL	Gemini-1.5-Pro
ANALYSIS	The image appears to be taken in a tropical or subtropical region due to the vegetation and sun angle. The architecture suggests a possibly African context. The road appears to be well-maintained. No license plates or signage are clearly visible to provide definitive clues.
PREDICTION	Kano, Nigeria, Africa

CoT Example (2)	
IMAGE	
GROUND TRUTH	Los Angeles, USA, North America
MODEL	LLaMA-3.2-11B-Vision
ANALYSIS	The image shows a woman standing on a dirt path in a wooded area, holding a cell phone and wearing a pink t-shirt, shorts, and a white visor. She is looking out at a view of mountains and valleys in the distance. * The woman is wearing a pink t-shirt, shorts, and a white visor. * She is holding a cell phone in her right hand. * She is standing on a dirt path in a wooded area. * There are trees and bushes surrounding her. * In the background, there is a view of mountains and valleys. * The sky is blue with some clouds. Based on the analysis, the street view appears to be located in a rural or natural area, possibly in a national park or nature reserve. The presence of mountains and valleys in the background suggests that the location is in a hilly or mountainous region.
PREDICTION	Santa Barbara, USA, North America

whether VLMs exhibit biases, such as favoring developed cities for urban, modern street scenes and developing cities for rural, undeveloped environments. For instance, as shown in Fig. 1(c), GPT-4o predicts urban scenes from Cape Town, South Africa, as San Diego, USA, and Nice, France. Conversely, for more rural images, Gemini-1.5-Pro misidentifies Moscow, Russia, as Kharkiv, Ukraine, and Madrid, Spain, as Seville, Spain. Similarly, LLaMA demonstrates comparable errors: a clean, organized street scene from Brasília, Brazil, is predicted as Sydney, Australia, and a high-rise cityscape from Krasnoyarsk, Russia, is identified as New York, USA. These findings reveal potential regional biases in VLMs when interpreting urban environments.

5.3 Can CoT Help?

To evaluate the performance of VLMs, we analyze their outputs using Chain-of-Thought (CoT) (Kojima et al., 2022; Wei et al., 2022) prompts. We present two example queries: one for Gemini and another for LLaMA. The case study suggests that while CoT reasoning can appear logical, it is not consistently tied to the final answer. In CoT Example (1), Gemini correctly identifies Africa's surroundings but notes the absence of visible license plates or signs that could aid in further country or city analysis. Despite this lack of evidence, the model still predicts the correct answer. Conversely, in CoT Example (2), LLaMA identifies features typical of California but incorrectly predicts Santa Barbara instead of the correct answer, Los Angeles. Across multiple examples, the elements cited in the CoT reasoning process often partially align with the final answer. However, these elements are typically broad and fail to accurately pinpoint specific locations. Relying solely on the reasoning process makes it challenging to determine the exact geographical location of an image. We hypothesize that the model's responses are not derived from genuine reasoning based on image information but are instead influenced by its prior knowledge of geographical locations.

6 Conclusion

This study identifies three types of biases in VLM in geolocation tasks using FAIRLOCATOR, a framework comprising 1,200 images sourced globally from Google Street View. The framework includes two subsets: the "Depth" subset, covering six countries and 60 cities, and the "Breadth" subset, spanning 43 countries and 60 cities. Key findings from the evaluation of four VLMs are as follows: (1) VLM predictions exhibit a bias toward larger cities, particularly in Brazil, Nigeria, and Russia. (2) Higher-performing models show improved ability to discern subtle differences between cities. (3) Accuracy consistently decreases in developing and less populous cities, with population size significantly influencing performance. (4) Accuracy varies notably across cultural groups, with city-level accuracy differing by up to 19.1%. Additionally, while VLMs demonstrate the capability to identify geographical locations, this raises privacy concerns, particularly regarding the potential exposure of personal geographical information in regions where models perform more accurately.

Limitations

This study has several limitations. (1) It does not investigate the underlying causes of biases in geographical information recognition. We hypothesize that these biases arise from imbalanced training datasets, where biased data contribute to the VLM's performance disparities. To test this hypothesis, we propose conducting comparative experiments using models trained on different datasets. Specifically, future research could compare the performance of VLMs trained in China and the United States in recognizing cities within China, providing deeper insights into whether dataset imbalance is a primary factor. (2) The evaluation does not include all countries globally. While we acknowledge the importance of every country, budget constraints limited our evaluation to 111 cities across 43 countries. To mitigate this limitation, we selected countries from diverse regions, cultures, and development levels to ensure broad coverage. Future studies can extend the evaluation by leveraging the workflow outlined in this paper.

Ethics Statements

License of Google Street View Images

In this section, we detail how our work adheres to the Google Street View terms of use.[11] The terms impose four key restrictions, addressed as follows: (1) "Creating data from Street View images, such as digitizing or tracing information from the imagery." Our work does not store or release specific Street View images. Instead, we report aggregated statistics derived from the collected images, with a few example images included solely for illustrative purposes in this paper. (2) "Using applications to analyze and extract information from the Street View imagery." We do not employ external applications for analysis. Instead, we rely on algorithmic methods for visual understanding of the Street View images. (3) "Downloading Street View images to use separately from Google services (such as an offline copy)." Our work utilizes images directly via the Street View API and does not distribute the images as a dataset. Instead, we release only the geographic coordinates, requiring future users to access the same images through the Street View API. (4) "Merging or stitching together multiple Street View images into a larger image." We do not merge or stitch Street View images in any form. By adhering to these restrictions, we ensure compliance with Google's terms of use for Street View, consistent with prior research practices (Fan et al., 2023; Gebru et al., 2017; Ki and Lee, 2021).

[11] https://about.google/brand-resource-center/products-and-services/geo-guidelines

Privacy Issues

Our work acknowledges the potential risk of malicious use, specifically the possibility that VLMs could be exploited to infer the locations of individuals through their publicly posted images. We strongly oppose and do not condone any behavior or activities that misuse this technology for such purposes. The intent of our research is to identify and highlight this potential problem within the context of academic and ethical research. By raising awareness, we aim to foster further discussion and develop safeguards to prevent misuse. Our goal is to advance understanding responsibly, without facilitating or endorsing any unethical applications of this technology.

References

Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. 2023. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*.

Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.

Jannik Brinkmann, Paul Swoboda, and Christian Bartelt. 2023. A multidimensional analysis of social biases in vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4914–4923.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Thomas Buckley, James A. Diao, Pranav Rajpurkar, Adam Rodman, and Arjun K. Manrai. 2023. Multimodal foundation models exploit text to make medical image predictions. *arXiv preprint arXiv:2311.05591*.

Song Chen, Xinyu Guo, Yadong Li, Tao Zhang, Mingan Lin, Dongdong Kuang, Youwei Zhang, Lingfeng Ming, Fengyu Zhang, Yuran Wang, et al.

2025. Ocean-ocr: Towards general ocr application via a vision-language model. *arXiv preprint arXiv:2501.15558*.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. 2025. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170.

Yongkang Du, Jen-tse Huang, Jieyu Zhao, and Lu Lin. 2025. Faircode: Evaluating social bias of llms in code generation. *arXiv preprint arXiv:2501.05396*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhuangyuan Fan, Fan Zhang, Becky PY Loo, and Carlo Ratti. 2023. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences*, 120(27):e2220417120.

Kathleen C Fraser and Svetlana Kiritchenko. 2024. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113.

Sourojit Ghosh and Aylin Caliskan. 2023. 'person'== light-skinned, western man, and sexualization of women of color: Stereotypes in stable diffusion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6971–6985.

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

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems*, 36.

Yingjie Hu, Gengchen Mai, Chris Cundy, Kristy Choi, Ni Lao, Wei Liu, Gaurish Lakhanpal, Ryan Zhenqi Zhou, and Kenneth Joseph. 2023. Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 37(11):2289–2318.

Jen-tse Huang, Yuhang Yan, Linqi Liu, Yixin Wan, Wenxuan Wang, Kai-Wei Chang, and Michael R Lyu. 2025. Fact-or-fair: A checklist for behavioral testing of ai models on fairness-related queries. *arXiv preprint arXiv:2501.05396*.

Donghwan Ki and Sugie Lee. 2021. Analyzing the effects of green view index of neighborhood streets on walking time using google street view and deep learning. *Landscape and Urban Planning*, 205:103920.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. 2023. Geolm: Empowering language models for geospatially grounded language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yi Liu, Junchen Ding, Gelei Deng, Yuekang Li, Tianwei Zhang, Weisong Sun, Yaowen Zheng, Jingquan Ge, and Yang Liu. 2024b. Image-based geolocation using large vision-language models. *arXiv preprint arXiv:2408.09474*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Hanjun Luo, Haoyu Huang, Ziye Deng, Xuecheng Liu, Ruizhe Chen, and Zuozhu Liu. 2024. Bigbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal llm. *arXiv preprint arXiv:2407.15240*.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David B Lobell, and Stefano Ermon. 2024. Geollm: Extracting geospatial knowledge from large language models. In *The Twelfth International Conference on Learning Representations*.

Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. 2024. Granular privacy control for geolocation with vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17240–17292.

Yuta Nakashima, Yusuke Hirota, Yankun Wu, and Noa Garcia. 2023. Societal bias in vision-and-language datasets and models. *NIHON GAZO GAKKAISHI (Journal of the Imaging Society of Japan)*, 62(6):599–609.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*.

Sundar Pichai and Demis Hassabis. 2024. Our next-generation model: Gemini 1.5. *Google Blog Feb 15 2024*.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. Biasdora: Exploring hidden biased associations in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10439–10455.

Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. 2023. Gpt4geo: How a language model sees the world's geography. *arXiv preprint arXiv:2306.00020*.

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008.

Gabriele Ruggeri and Debora Nozza. 2023. A multidimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455.

Ashutosh Sathe, Prachi Jain, and Sunayana Sitaram. 2024. A unified framework and dataset for assessing gender bias in vision-language models. *arXiv preprint arXiv:2402.13636*.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85.

Yixin Wan and Kai-Wei Chang. 2024. The male ceo and the female assistant: Probing gender biases in text-to-image models through paired stereotype test. *arXiv preprint arXiv:2402.11089*.

Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R Lyu. 2023. Biasasker: Measuring the bias in conversational ai system. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 515–527.

Wenxuan Wang, Haonan Bai, Jen-tse Huang, Yuxuan Wan, Youliang Yuan, Haoyi Qiu, Nanyun Peng, and Michael R Lyu. 2024. New job, new gender? measuring the social bias in image generation models. In *Proceedings of the 32nd ACM Multimedia Conference*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Robert Wolfe and Aylin Caliskan. 2022. American== white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 800–812.

Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2023. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1174–1185.

Yifan Yang, Siqin Wang, Daoyang Li, Shuju Sun, and Qingyang Wu. 2024a. Geolocator: A location-integrated large multimodal model (lmm) for inferring geo-privacy. *Applied Sciences*, 14(16):7091.

Zhen Yang, Jinhao Chen, Zhengxiao Du, Wenmeng Yu, Weihan Wang, Wenyi Hong, Zhihuan Jiang, Bin Xu, Yuxiao Dong, and Jie Tang. 2024b. Mathglm-vision: Solving mathematical problems with multimodal large language model. *arXiv preprint arXiv:2409.13729*.

Robert B Zajonc. 1968. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1.

Jingyu Zhang, Alexandra DeLucia, Chenyu Zhang, and Mark Dredze. 2023. Geo-seq2seq: Twitter user geolocation on noisy data through sequence to sequence learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4778–4794.

Yi Zhang, Junyang Wang, and Jitao Sang. 2022. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4996–5004.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *arXiv preprint arXiv:2403.00277*.

A City Predictions from Other VLMs

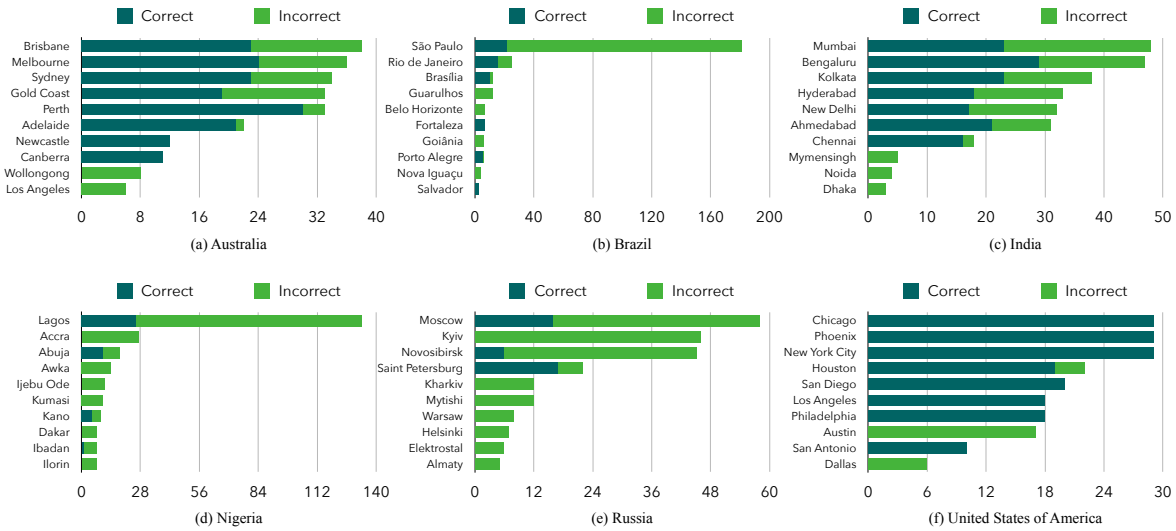


Figure 3: The most frequently predicted cities by Gemini-1.5-Pro across six countries. Each country includes ten cities, with ten images per city used for testing. The maximum "Correct" score for a city is 30, as the VLMs have three attempts to predict the location.

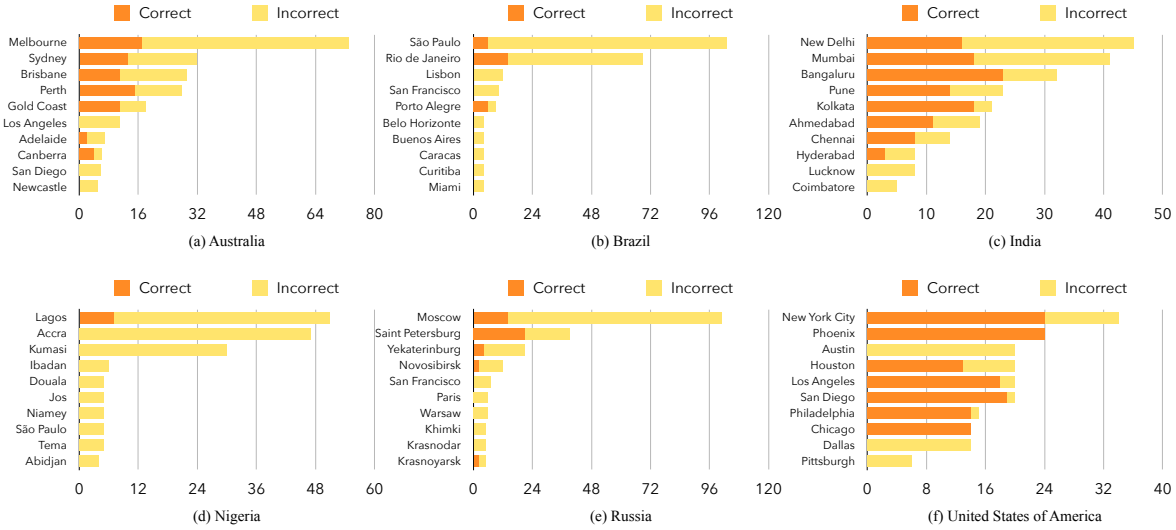


Figure 4: The most frequently predicted cities by LLaMA-3.2-11B-Vision across six countries. Each country includes ten cities, with ten images per city used for testing. The maximum "Correct" score for a city is 30, as the VLMs have three attempts to predict the location.

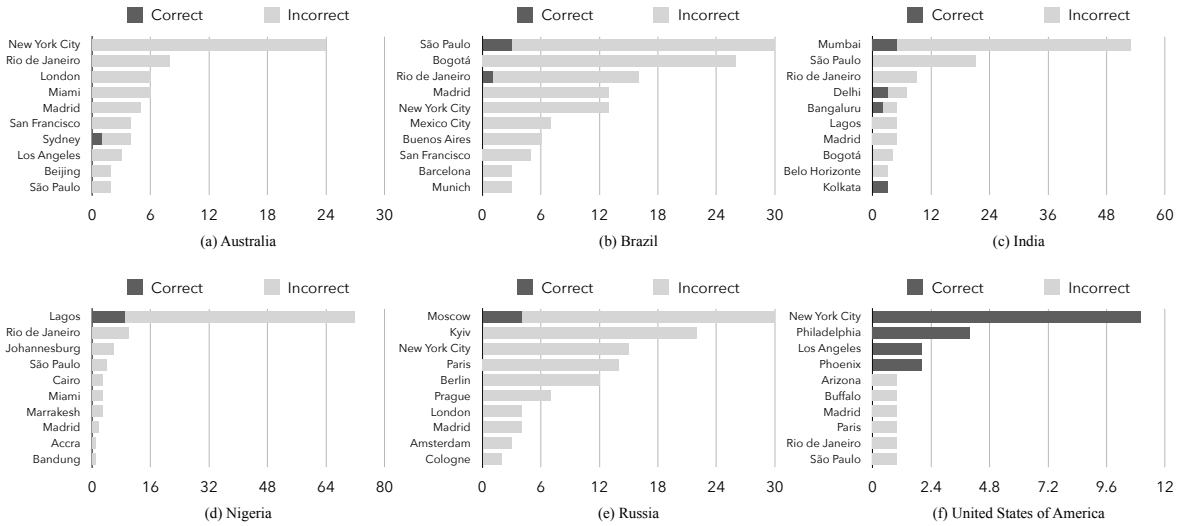
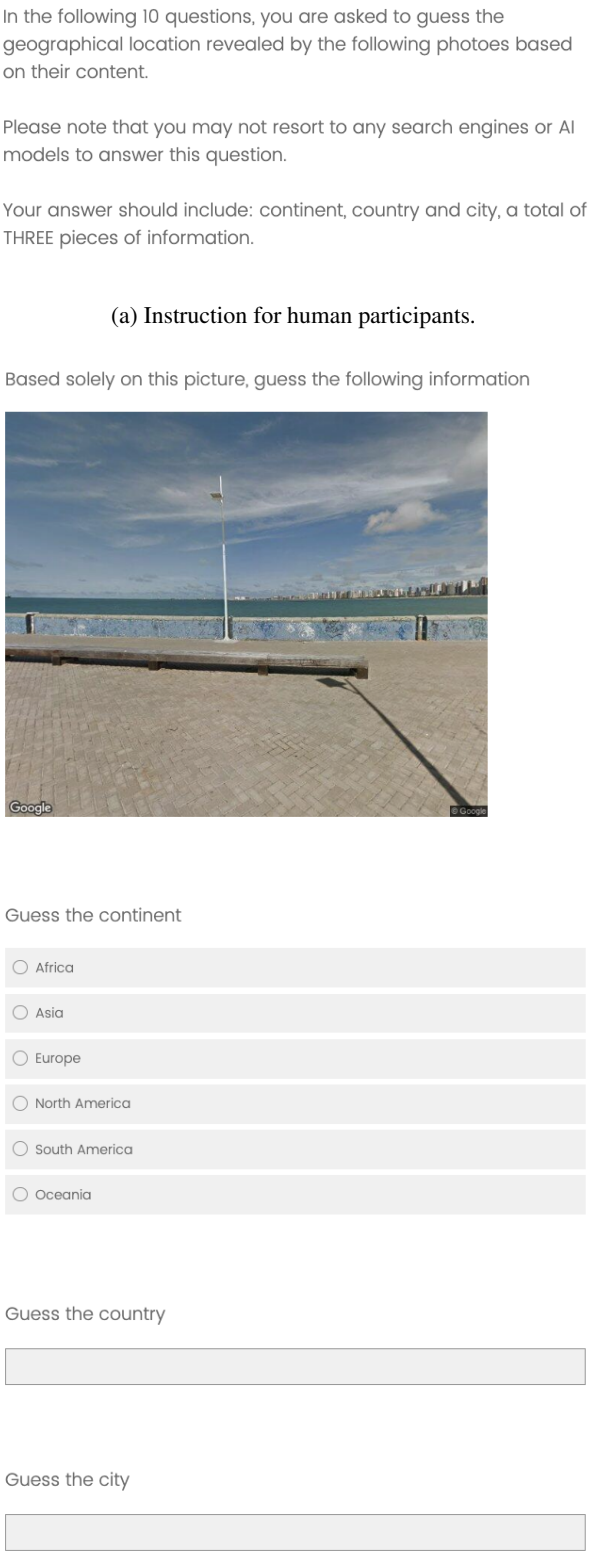


Figure 5: The most frequently predicted cities by LLaVA-V1.6-Vicuna-13B across six countries. Each country includes ten cities, with ten images per city used for testing. The maximum "Correct" score for a city is 30, as the VLMs have three attempts to predict the location.

B User Study Questionnaire



(a) Instruction for human participants.

(b) An example question.

Figure 6: Illustration of our questionnaires.