

From Pixels to Places: A Systematic Benchmark for Evaluating Image Geolocalization Ability in Large Language Models

Anonymous ACL submission

Abstract

Image geolocalization is important for applications in crisis response, digital forensics, and location-based intelligence. While recent advances in large language models (LLMs) offer new opportunities for visual reasoning, their ability to perform image geolocalization remains underexplored. In this study, we introduce a benchmark called **IMAGEO-Bench** that systematically evaluates accuracy, distance error, geospatial bias, and reasoning process. Through experiments on 10 state-of-the-art LLMs, we reveal clear performance disparities, with closed-source models generally showing stronger reasoning. Importantly, we uncover geospatial biases as LLMs perform better in high-resource regions (e.g., North America, Western Europe, and California) while exhibiting degraded performance in underrepresented areas. Regression diagnostics demonstrate that successful geolocalization is dependent on recognizing urban settings, outdoor environments, street-level imagery, and identifiable landmarks. Overall, **IMAGEO-Bench** provides a rigorous lens into the spatial reasoning capabilities of LLMs and offers implications for building geolocation-aware AI systems.

1 Introduction

Image geolocalization, the process of inferring geographic coordinates or addresses from visuals, presents a complex challenge at the intersection of computer vision and geographic information systems (GIS) (Lin et al., 2013; Workman et al., 2015; Jin Kim et al., 2017). The accurate identification of location from images has important implications for real-world scenarios, including digital forensics (Kadha et al., 2025), urban analytics (Lin et al., 2013) and crisis management (Li et al., 2025a; Yin et al., 2025). Conventionally, this task has been investigated with computer vision models, using either supervised classification (Pramanick et al., 2022; Seo et al., 2018) or retrieval-based methods

(Vivanco Cepeda et al., 2023; Zhu et al., 2022; Vo and Jacobs, 2017). These models often leverage metadata embedded in image files (Samano et al., 2020) or rely on extensive GIS databases (Hays and Efros, 2014; Brejcha and Čadík, 2017) to recognize landmark or visual cues. However, early methods often fail to generalize to previously unseen environments and perform inadequately in the absence of distinctive landmarks. Their effectiveness is further constrained by data coverage and resolution, especially when the metadata is missing.

Recent advances in large language models (LLMs) have transformed the landscape of image understanding and spatial reasoning (Chen et al., 2024; Wu et al., 2024; Tian et al., 2025). These models can describe intricate details, understand spatial relationships between objects, and infer context from subtle visual cues (Chen et al., 2024; Yang et al., 2025). By leveraging extensive world knowledge and visual-textual alignment, LLMs offer a promising avenue for overcoming the limits of conventional geolocalization, even when confronted with out-of-distribution data (Zhou et al., 2024; Xu et al., 2024). Recent research further demonstrates that targeted strategies, such as supervised fine-tuning (Yi and Shan, 2025) and retrieval-augmented inference (Li et al., 2025b), can substantially enhance LLMs’ geolocalization capabilities.

In parallel with LLM advances, there has been rapid progress in developing benchmarks to evaluate their visual and linguistic capabilities (Lacoste et al., 2023; Hua et al., 2024; Fan et al., 2024). A representative example is GEO-Bench, which evaluates LLMs on classification and segmentation tasks in geoscience (Lacoste et al., 2023). For image geolocalization, benchmarks like Img2Loc (Zhou et al., 2024) and LLMGeo (Wang et al., 2024) have assessed LLMs’ ability to localize images. However, these efforts often rely on established datasets and lag in incorporating the recent model advances. Moreover, key dimensions like

LLMs’ spatial reasoning abilities and potential geospatial biases remain underexplored. To address these gaps, we introduce **IMAGEO-Bench**, a benchmark that systematically evaluates LLMs’ image geolocalization abilities.

Our key contributions are listed as follows. (i) **Comprehensive benchmark datasets**: We compile three datasets with diverse geographic regions, urban/rural scenes, and location indicators, constructed via stratified sampling and augmented with privately collected images to ensure broad coverage. (ii) **Multi-dimensional evaluation framework**: We propose an evaluation framework that looks beyond standard accuracy to assess spatial precision and computational cost, providing a nuanced understanding of model performance across various dimensions. (iii) **Geographic bias quantification**: We quantify geographic biases in LLMs by revealing performance disparities across different regions at both global and local scales. (iv) **Reasoning process analysis**: We employ statistical regression to analyze the reasoning processes behind LLM geolocalization and to identify the most influential visual features.

2 Related Work

2.1 Image Geolocalization

Image geolocalization in computer vision has focused on two main approaches: classification-based and retrieval-based models (Weyand et al., 2016; Seo et al., 2018; Muller-Budack et al., 2018; Pramanick et al., 2022; Clark et al., 2023). Classification methods (e.g., PlaNet (Weyand et al., 2016)) treat geolocation as a discrete prediction task by assigning an image to a predefined spatial grid. Retrieval-based methods (e.g., IM2GPS (Hays and Efros, 2008)) localize an image by matching it to a geotagged reference set based on visual or semantic similarity. Metric learning models like NetVLAD (Arandjelović et al., 2016), SAFA (Shi et al., 2020), and VIGOR (Zhong et al., 2021) further boost retrieval performance using cross-view embeddings and attention mechanisms.

However, these methods face several limitations: (i) they rely on large labeled datasets and generalize inadequately to unseen or underrepresented regions; (ii) their predictions are often opaque, lacking interpretability or insight into the reasoning process; and (iii) they struggle to incorporate multimodal semantic cues that humans intuitively use for geolocation. These limitations call for an ap-

proach that can leverage world knowledge and multimodal cues to generalize more effectively.

2.2 LLMs for Spatial Reasoning

Recent progress in vision-language pre-training has given rise to multimodal LLMs with strong visual reasoning capabilities. Foundation models (e.g., CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021)) have established powerful image-text alignment tools enabling zero-shot classification and retrieval. A few recent studies have evaluated modern LLMs’ abilities to localize images (Wang et al., 2024; Zhou et al., 2024). For example, LLM-Geo (Wang et al., 2024) benchmarks GPT-4V and LLaVA on challenging in-the-wild images. It reveals limited accuracy of investigated models and highlights room for improvement. Img2Loc (Zhou et al., 2024) introduces a retrieval-augmented generation (RAG) pipeline that combines LLMs with large-scale image search to refine coordinate estimates. A more recent study presents a benchmark built on Google Street images (Jay et al., 2025), showing that VLMs can achieve sub-300km median errors but suffer from notable failure modes and reasoning transparency.

While these studies represent important progress, they are constrained by task-specific designs, limited geographic scope, and the absence of standardized evaluation metrics. These underscore the need for a systematic framework to assess LLMs’ geolocation capabilities across diverse regions and real-world conditions. To address these gaps, we introduce **IMAGEO-Bench**, a benchmark for systematically evaluating LLMs’ performance on image geolocalization. The framework of this benchmark is illustrated in Figure 1.

3 Benchmark Data Preparation

We construct three benchmark datasets from distinct sources with their geographic distribution illustrated in Figure 1(a). We opt to build new benchmark datasets for two main reasons. First, this helps mitigate data contamination, as widely used datasets like IM2GPS3k and YFCC4k (Hays and Efros, 2008) may have been incorporated in LLM training corpora. Second, our datasets are designed to span diverse visual contexts (e.g., indoor vs. outdoor) and geospatial scales (e.g., global vs. U.S. specific), allowing for a more rigorous assessment of LLM capabilities. Sample images of these benchmark datasets are presented in **Appendix A**.

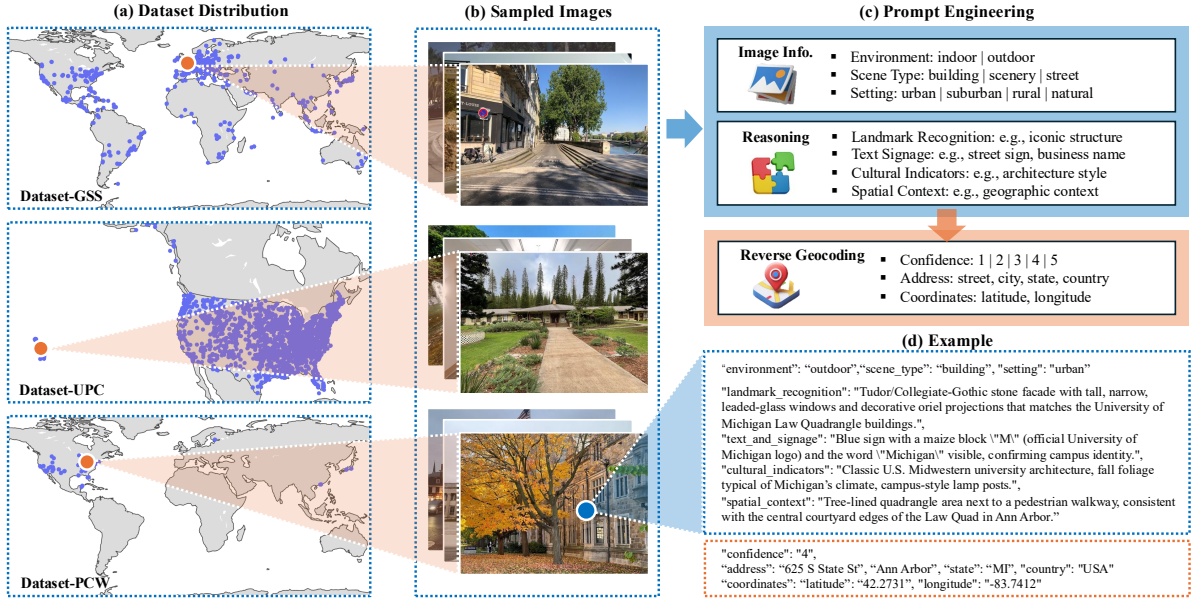


Figure 1: The illustrative framework to implement this study. (a) Data distribution for each benchmark dataset. (b) Sampled images from each dataset. (c) Prompt design. (d) An example of output by the o3 model.

3.1 Data-GSS: Global Streetscape Set

Dataset-GSS is derived from the NUS Global Streetscapes dataset (Hou et al., 2024), a collection of approximately 10 million streetscape images collected from Mapillary and KartaView across 688 cities worldwide. Each image is geo-tagged with precise latitude and longitude coordinates and has city-level labels (Hou et al., 2024). We build Dataset-GSS by refining this manually labeled subset through a multi-stage filtering process. First, we filter the images with all contextual labels, which results in 8,568 images. Second, we only retain the images with the annotations—*quality: good*, *reflection: no*, and *glare: no*—to ensure image quality, narrowing the dataset to 6,178 images. Last, we manually filter to exclude images with insufficient geographic information (e.g., no signs or clear visual cues). After filtering, we obtain Dataset-GSS, which comprises 6,152 high-quality street-level images from 396 cities across 123 countries. The data distribution is presented in **Appendix B**.

3.2 Data-UPC: U.S. POIs Crowdsourced Set

Dataset-UPC is constructed from crowdsourced images from Google Maps Points of Interest (POIs) to capture a diverse range of regions and location types. We first sample POIs from the Google Maps dataset released by UC San Diego (Yan et al., 2023), which contains 4,963,111 unique POIs across the United States collected up to September 2021. To ensure geographic coverage in Dataset-UPC, we employ a multi-stage stratified sampling strategy.

We begin with compiling a list of 17 distinct POI categories, including “restaurant,” “hotel,” “park,” and “scenic spot,” to guide the sampling. For each state, we randomly select three POIs from each category, resulting in a pool of 4,325 POIs. We then refine this pool through a two-step filtering process to ensure data usability and integrity. First, we discard entries with non-functional URLs. Second, we remove advertisement images and those containing identifiable human faces. After this curation, Dataset-UPC comprises 2,929 images spanning all 50 states and the District of Columbia.

3.3 Data-PCW: Privately Collected Wild Set

Dataset-PCW is constructed as a held-out validation set to prevent potential data leakage. It consists of 220 original photo-address pairs collected by the authors. Since none of the images overlap with existing datasets or known LLM training sources, Dataset-PCW is well suited to evaluate LLMs’ geolocation performance on unseen data.

4 Benchmark Development

IMAGEO-Bench defines a unified end-to-end task for image geolocation. It includes a structured output format, reasoning template, evaluation metrics, and post-hoc feature-importance analysis tools.

4.1 Task Formalization

Given an input image x , the task requires models to predict a structured geographic output tuple: $G = (C, A, (\phi, \lambda))$, where (ϕ, λ) are the predicted

latitude and longitude in decimal degrees, A is a structured textual address, and $C \in \{1, 2, 3, 4, 5\}$ denotes the model’s self-reported confidence score. The objective is to minimize the great-circle distance error $d((\phi, \lambda), (\phi^*, \lambda^*))$ between the predicted and ground-truth coordinates. During inference, models are strictly restricted from accessing external APIs or GPS metadata.

4.2 Output Schema and Reasoning Template

To evaluate LLMs’ geolocation capabilities, we require models to explicitly reason about the given image by extracting and interpreting salient visual features, as illustrated in Figure 1(c). These features are collected based on prior research identifying layout, signage, architecture, road, infrastructure, and buildings as critical cues for image geolocation (Li et al., 2025b; Jay et al., 2025; Wu and Huang, 2022; Tian et al., 2017). Since image geolocation often relies on recognizing features tied to specific places, prompting LLMs to surface these elements can facilitate a step-by-step reasoning process similar to that of human geographers.

Specifically, the structured model output comprises three key components: (i) **Image information**: High-level descriptors characterizing the environment (indoor, outdoor), scene type (building, scenery, street, mixed), and setting (urban, suburban, rural, natural). (ii) **Reasoning**: A detailed justification for location inference, organized into four distinct categories of visual evidence: Landmark recognition—Unique natural or human-made structures serving as distinctive geographic markers. Text and signage—Textual clues from visible signs, license plates, storefronts, or other textual indicators. Cultural indicators—Region-specific architectural styles, cultural symbols, or prevalent societal patterns. (iv) Spatial context—Environmental layout, vegetation patterns, road networks, infrastructure types, or building density. (iii) **Geolocalization**: The final structured geographic prediction tuple $(C, A, (\phi, \lambda))$, as formally defined.

4.3 Evaluation Metrics

We establish a set of metrics to measure LLMs’ performance: (i) **Identification rate**: It measures the model’s ability to produce a valid output, defined as the percentage of input images for which a complete and identified location in the formatted JSON is returned. (ii) **Geographic accuracy**: It evaluates the semantic correctness of the predicted address A_{pred} against the ground

truth A_{true} at three administrative levels, including (i) Country accuracy ($Acc_{country}$), (ii) State accuracy (Acc_{state}), and (iii) City accuracy (Acc_{city}). This metric quantifies the proportion of predictions that align with the ground-truth labels. (iii) **Distance error**: It calculates the great-circle distance d (kilometers) between the predicted coordinates (ϕ_p, λ_p) and the true coordinates (ϕ_t, λ_t) using the Haversine formula, $d = 2r \arcsin(\sqrt{a})$, where $a = \sin^2(\Delta\phi/2) + \cos(\phi_p) \cos(\phi_t) \sin^2(\Delta\lambda/2)$, $\Delta\phi = \phi_t - \phi_p$, and $\Delta\lambda = \lambda_t - \lambda_p$, and r represents the Earth’s mean radius (6,371 km). (iv) **Computational cost**: It assesses the model’s efficiency by tracking the computational costs, including the averaged cost Avg_{Cost} and the average number of tokens per inference Avg_{Tokens} .

4.4 Post-hoc Feature-Importance Diagnostics

IMAGEO-Bench includes an interpretability module to assess how specific structured output fields correlate with geolocation performance. This diagnostic suite supports model-agnostic evaluation, identifying which visual attributes are most predictive of successful localization. Specifically, we construct a diagnostic set from each model’s structured JSON outputs. Key categorical predictors—such as environment type, scene category, setting, and landmark presence—are extracted from the predefined schema and one-hot encoded relative to baseline categories. We further define two metrics to represent the geolocation outcome, including (i) a binary indicator representing city-level prediction correctness, and (ii) a continuous log-transformed distance error. Let $\mathbf{x} \in \mathbb{R}^d$ denote this encoded feature vector, and let Y represent the geolocation outcome (e.g., city-level correctness or log-distance error). To quantify feature importance, we fit the following two regression models.

Logistic regression. For the city-level correctness indicator $Y \in \{0, 1\}$, we estimate:

$$\mathbb{P}(Y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b), \quad (1)$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function. We report $\exp(w_j)$ as the odds ratio of feature j relative to its reference category.

Ridge regression. For the log-distance error, we use ridge regression to fit:

$$\log(d) = \mathbf{w}^\top \mathbf{x} + b + \varepsilon, \quad (2)$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, by minimizing the penalized squared error $\sum_i (y_i - \hat{y}_i)^2 + \lambda \|\mathbf{w}\|_2^2$ with $\lambda = 1.0$. All predictors are one-hot encoded binary

variables, so additional normalization is unnecessary. Because all features share the same scale, the ridge coefficients w are directly comparable. We aggregate $|w_j|$ within each semantic field to obtain field-level importance.

The importance of the field level is calculated by aggregating the absolute regression coefficients associated with each predictor group. Coefficients are interpreted relative to their baseline categories, with careful consideration given to potential collinearity effects. These diagnostic results provide valuable information on the semantic and visual cues that most strongly influence model geolocation performance.

5 Experimental Setup

Prompt Design. We employ a structured prompt to ensure consistent and parsable outputs. The main prompt directs a model to use a step-by-step, chain-of-thought reasoning process. To further guide the model’s inference, we include a one-shot example within the prompt. We mandate that the response be a valid JSON object partitioned into three sections: (i) Image information for high-level scene classification, (ii) reasoning for a detailed analysis of visual evidence, and (iii) image geolocation for the final coordinate estimation. Specific prompt design is presented in **Appendix C**.

Model Selection. We select a set of candidate LLMs with vision capabilities from several prominent providers. The models under evaluation include OpenAI’s o3, gpt-4.1, and gpt-4.1-mini; Anthropic’s claude-3.5-haiku; Google’s gemini-1.5-pro, gemini-2.5-flash, and gemini-2.5-pro; and various Llama models, specifically llama-3.2-11b, llama-3.2-90b, and llama-4-maverick-17b.

Experiment Settings. In our experimental setup, we query each model with a temperature setting of 0.7, which encourages a balance between factual recall and inferential creativity. We log all outputs for each image, including the model’s raw JSON response and error states, to a .jsonl file for subsequent analysis. At the conclusion of each experimental run, we also compute and record a summary of the total operational cost and token consumption. Representative model outputs of image geolocation are provided in **Appendix I**.

6 Results

6.1 Benchmark Performance

To benchmark LLMs’ performance, we generate two sets of scatter plots using the introduced met-

rics. Table 1 presents a quantitative summary, including accuracy at various geospatial scales, median distance error, and computational costs. For visual analysis, Figure 2 displays scatter plots for latitude prediction on Dataset-GSS. The latitude result on Dataset-UPC, longitude result on Datasets-GSS & -UPC, and benchmark performance on Dataset-PCW are provided in **Appendix D**, **Appendix E**, and **Appendix F**, respectively.

Performance across Dataset-GSS and Dataset-UPC shows clear differences (Table 1). This discrepancy is likely due to the inherent characteristics of the compiled datasets. Dataset-GSS primarily consists of publicly available global street-level images captured in outdoor environments, which provide clear geographic indicators. In contrast, Dataset-UPC comprises crowdsourced images contributed by Google Maps users, many of which are taken indoors, lack explicit geographic cues or distinctive architectural details. These characteristics significantly complicate accurate identification of geolocation. As a result, most LLMs achieve higher accuracy with Dataset-GSS across all levels.

Closed-source models consistently outperform their open-source counterparts. For example, gemini-2.5-pro achieves remarkable results in Dataset-GSS, with a median error of 4.2 km and country-level accuracy of 92.4%. Other top performers like gpt-4.1 and o3 maintain high performance with median errors of 10.5 km and 8.2 km, respectively, in Dataset-GSS. Conversely, open-source models such as llama-4-17b and llama-3.2-11b exhibit substantially higher median errors (e.g., 127.0 km for llama-4-17b in Dataset-GSS and 534.3 km in Dataset-UPC).

We also observe a clear scaling trend. For example, gpt-4.1 surpasses its smaller counterpart, gpt-4.1-mini, with a city-level accuracy of 39.6% versus 34.7% and a much lower median distance error (10.5 km versus 34.8 km). Similarly, llama-3.2-90b improves upon llama-3.2-11b with a city-level accuracy of 26.3% versus 24.0%, and a median distance error of 15.7 km versus 66.6 km. However, performance differences are not solely size-dependent. Smaller models like gemini-2.5-flash (city accuracy 41.7%, median error 6.0 km) outperform many larger models, suggesting that other factors (e.g., model architecture, multimodal fusion techniques, and training quality) can contribute to improved performance. This reinforces that scaling is helpful but not sufficient on its own.

Confidence scores vary significantly across the

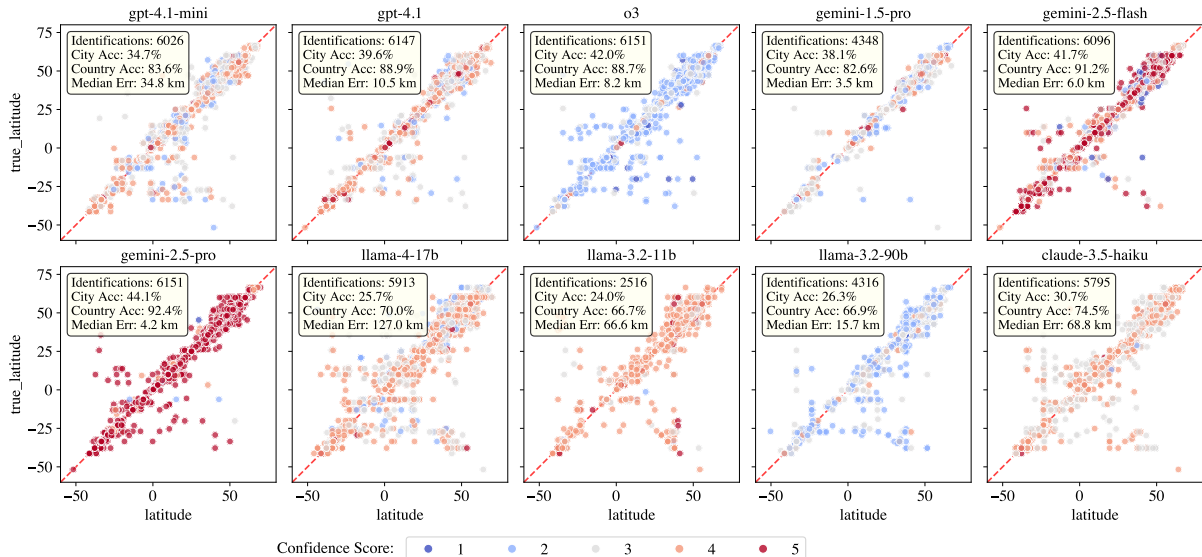


Figure 2: Benchmark performance based on latitude prediction for Dataset-GSS. Perfect predictions lie on the red dashed diagonal line. The proximity of points to this line indicates the model’s accuracy, with the color of each point representing the model’s self-reported confidence score on a scale of 1 to 5.

Model	Dataset-GSS				Dataset-UPC							
	City Acc (%)	Country Acc (%)	Mean Dist (km)	Median Dist (km)	Avg Tokens	Avg Cost (\$)	City Acc (%)	State Acc (%)	Mean Dist (km)	Median Dist (km)	Avg Tokens	Avg Cost (\$)
gpt-4.1-mini	34.7%	83.6%	381.1	34.8	3245	0.00164	12.5%	32.9%	768.8	376.7	1496	0.00092
gpt-4.1	39.6%	88.9%	296.1	10.5	1842	0.00536	20.7%	44.2%	685.8	236.1	1465	0.00455
o3	42.0%	88.7%	288.1	8.2	2912	0.01460	23.9%	45.8%	662.7	214.3	2281	0.01150
gemini-1.5-pro	38.1%	82.6%	161.4	3.5	1178	0.00243	16.3%	36.6%	406.5	109.5	1165	0.00237
gemini-2.5-flash	41.7%	91.2%	227.0	6.0	1379	0.00141	19.9%	44.6%	587.7	212.2	1326	0.00128
gemini-2.5-pro	44.1%	92.4%	216.3	4.2	1367	0.00556	26.4%	49.5%	544.4	150.4	1328	0.00517
llama-4-17b	25.7%	70.0%	840.3	127.0	2961	0.00093	9.1%	24.8%	1217.5	534.3	1767	0.00061
llama-3.2-11b	24.0%	66.7%	797.4	66.6	823	0.00015	8.2%	18.9%	955.5	353.2	822	0.00015
llama-3.2-90b	26.3%	66.9%	382.9	15.7	7211	0.00865	11.6%	24.0%	706.8	163.0	5206	0.00625
claude-3.5-haiku	30.7%	74.5%	568.2	68.8	2478	0.00281	8.7%	30.0%	697.1	258.7	1421	0.00190

Table 1: LLM performance benchmark based on Dataset-GSS and Dataset-UPC.

models, among which gemini-2.5-pro exhibits the highest confidence, while o3 reports the lowest. Even within a single provider, confidence levels differ: for instance, gpt-4.1 is substantially more confident than gpt-4.1-mini, and gemini-2.5-pro scores higher than gemini-1.5-pro. For models like gpt-4.1, o3, and claude-3.5-haiku, higher confidence scores generally align with better prediction accuracy within the same model. However, higher confidence does not always imply greater performance across models. For instance, gemini-2.5-pro shows a higher confidence distribution than o3, yet their performance in Dataset-GSS is comparable (median errors of 4.2 km versus 8.2 km). These findings indicate that confidence alone is not a reliable proxy for model accuracy and should be interpreted in conjunction with performance metrics.

6.2 Geospatial Analysis

We conduct geospatial analyses on Dataset-GSS and Dataset-UPC to evaluate model performance across geographic regions and scales. For Dataset-

GSS (Figure 3), models such as o3 and gemini-2.5-pro achieve near-perfect accuracy (approaching 1.0) in regions like North America, Russia, and Australia. In contrast, llama-3.2-11b consistently underperforms, with accuracy often below 0.6 and approaching zero in several Latin European countries. Models like llama-3.2-90b and claude-3.5-haiku show intermediate accuracy (0.6–0.9). It is noting that the United States, Australia, and West European Countries maintain high accuracy across all models, suggesting strong model generalization in these well-represented regions.

For Dataset-UPC (Figure 4), we again observe distinct geographic disparities across U.S. states. Specifically, gemini-2.5-flash, gemini-2.5-pro, and o3 consistently perform well, particularly in California, Hawaii, and Florida. In contrast, llama-3.2-11b falls below 0.5 accuracy in nearly all states. gpt-4.1 shows moderate performance, with accuracy typically between 0.6 and 0.8 in coastal regions. California and Florida emerge as consistent high-performing states across models, likely due

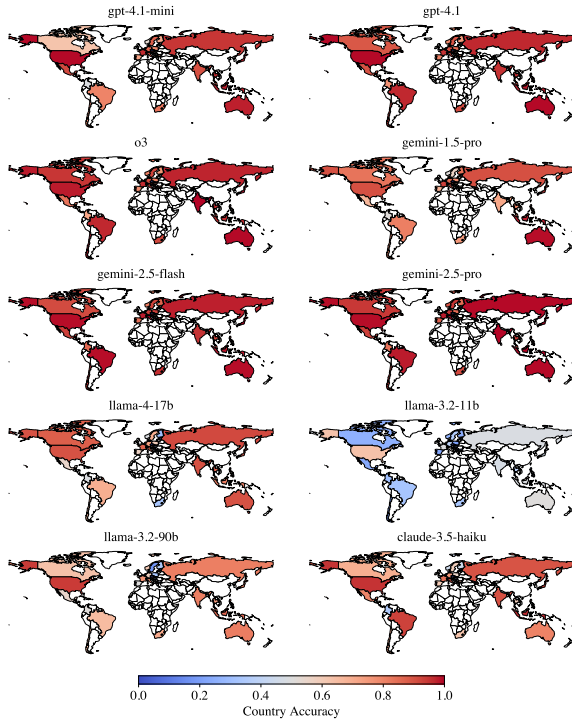


Figure 3: Country-level averaged accuracy across models on Dataset-GSS (a threshold of 20 sampled images).

to their dense urban environments and distinctive landmarks that can facilitate image geolocation.

Together, the results illustrate potential geospatial biases in model behavior. High performance in regions like the United States, Australia, and parts of Western Europe points to the models’ strong alignment with densely sampled, visually distinctive, and data-rich environments. In contrast, accuracy declines sharply in countries and states with fewer training representations or more ambiguous visual features. These patterns reflect a potential limitation that LLMs are not yet equipped to generalize robustly across diverse global geographies.

6.3 Reasoning Analysis

Feature Attribution via Reasoning. We apply logistic regression diagnostics to examine the features for the LLM reasoning process. The summarized result for Dataset-GSS is illustrated in Figure 5. Image elements are categorized into four semantic groups: environment, scene type, setting, and landmark. Across the evaluated LLMs, we observe the following key trends.

First, nearly all models strongly favor urban and outdoor environments while consistently penalizing rural, natural, and indoor scenes. One possible explanation is that dense urban contexts, featuring detailed façades, prominent signage, and infrastructure elements (e.g., roads and street furniture),

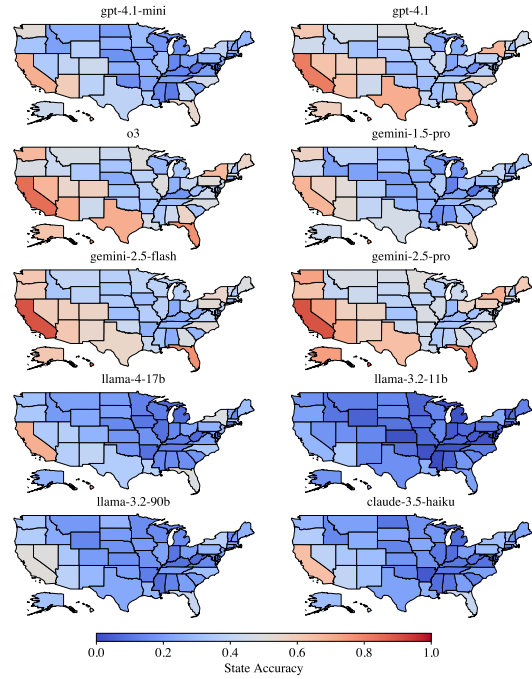


Figure 4: U.S. state-level averaged accuracy across models on Dataset-UPC (a threshold of 20 sampled images).

provide reliable visual cues for accurate geolocation. Conversely, natural scenes, which often lack distinctive visual anchors, pose significant localization challenges, as reflected by consistently negative model associations.

Second, regarding scene type, street-level imagery provide stronger localization signals than building façades or general scenery. This is likely because street scenes often contain rich contextual cues (e.g., textual signage, road markings, and distinctive urban infrastructure) that are both visually salient and semantically interpretable by LLMs. In contrast, scenic images without clear architectural landmarks can negatively impact prediction.

Third, the presence of distinctive landmarks consistently improves the geolocation performance. Landmark cues are typically assigned strong positive weights, indicating their critical role in guiding correct predictions. However, the claude-3.5-haiku model is a notable outlier: it assigns marginal or even negative weights to the landmark cue. A post-hoc audit reveals that it labels “landmark = yes” on 99.67% of images, compared to 41.77% for all other models. This over-identification of landmarks suggests that the model frequently misclassifies ordinary structures as distinctive, thereby diluting the predictive value of this feature. Additional analyses and visualizations, including distance error evaluations on Datasets-UPC & PCW, are provided in **Appendix G**.

604 Limitations

605 This study opens several avenues for future re-
606 search. First, while IMAGEO-Bench spans multi-
607 ple datasets and geographic scales, the benchmark
608 cannot fully capture the diversity of real-world vi-
609 sual environments. Certain regions, scene types,
610 and visual conditions—particularly rural, natural,
611 indoor, or low-resource settings—can remain un-
612 derrepresented in the reported findings. A more
613 comprehensive evaluation incorporating broader
614 geographic coverage, seasonal variation, and cul-
615 turally diverse built environments could provide a
616 stronger test of model robustness under real-world
617 variability. Such expansions are especially impor-
618 tant for assessing deployment readiness in global
619 or safety-critical applications.

620 Second, the benchmark primarily emphasizes
621 visually grounded reasoning, which may under-
622 state the role of language-based priors and latent
623 world knowledge in LLM geolocation. Models
624 may rely on implicit textual associations, regional
625 stereotypes, or training-induced biases when visual
626 anchors are weak or ambiguous. Future work could
627 more explicitly probe the interaction between vi-
628 sual evidence and language-based cues, such as ma-
629 nipulating textual context, masking salient visual
630 features, or introducing counterfactual prompts.

631 Third, strategies such as targeted data augmen-
632 tation, region-adaptive prompting, or bias-aware
633 calibration could be explored to mitigate the pro-
634 nounced geographic performance disparities ob-
635 served across regions, moving toward more equi-
636 table and globally robust geolocation systems.

637 Ethical Considerations

638 Image geolocation poses ethical risks related
639 to privacy, surveillance, and misuse, as inferring
640 locations from images can enable intrusive track-
641 ing or unintended disclosure of sensitive places.
642 Although this study relies on publicly available or
643 author-collected images and restricts access to ex-
644 ternal metadata, the demonstrated capabilities of
645 LLMs underscore the need for responsible deploy-
646 ment. In addition, the observed geographic per-
647 formance disparities raise concerns about fairness,
648 as errors may disproportionately affect underrep-
649 resented regions. These considerations highlight the
650 importance of bias auditing, uncertainty-aware re-
651 porting, and clear usage guidelines when applying
652 geolocation systems in real-world settings.

References

- 654 Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomáš 654
655 Pajdla, and Josef Sivic. 2016. *NetVLAD: CNN ar- 655*
656 *chitecture for weakly supervised place recognition.* 656
657 *In Proc. IEEE Conf. on Computer Vision and Pattern 657*
658 *Recognition (CVPR)*, pages 5297–5307. 658
- Jan Brejcha and Martin Čadík. 2017. State-of-the-art in 659
660 visual geo-localization. *Pattern Analysis and Appli- 660*
661 *cations*, 20(3):613–637. 661
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie 662
663 Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind 663
664 Neelakantan, Pranav Shyam, Girish Sastry, Amanda 664
665 Askell, and 1 others. 2020. Language models are 665
666 few-shot learners. *Advances in neural information 666*
667 *processing systems*, 33:1877–1901. 667
- Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, 668
669 Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. 669
670 SpatialVlm: Endowing vision-language models with 670
671 spatial reasoning capabilities. *In Proceedings of the 671*
672 *IEEE/CVF Conference on Computer Vision and Pat- 672*
673 *tern Recognition*, pages 14455–14465. 673
- Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, 674
675 Vicente Vivanco Cepeda, and Mubarak Shah. 2023. 675
676 Where we are and what we’re looking at: Query 676
677 based worldwide image geo-localization using hier- 677
678 archies and scenes. *In Proceedings of the IEEE/CVF 678*
679 *Conference on Computer Vision and Pattern Recog- 679*
680 *nition*, pages 23182–23190. 680
- Lizhou Fan, Wenyue Hua, Xiang Li, Kaijie Zhu, 681
682 Mingyu Jin, Lingyao Li, Haoyang Ling, Jinkui 682
683 Chi, Jindong Wang, Xin Ma, and 1 others. 2024. 683
684 Nphardeval4v: A dynamic reasoning benchmark of 684
685 multimodal large language models. *arXiv preprint 685*
686 *arXiv:2403.01777*. 686
- James Hays and Alexei A Efros. 2008. *Im2gps: esti- 687*
688 *imating geographic information from a single image.* 688
689 *In 2008 IEEE Conference on Computer Vision and 689*
690 *Pattern Recognition*, pages 1–8. IEEE. 690
- James Hays and Alexei A Efros. 2014. Large-scale 691
692 image geolocation. *In Multimodal location esti- 692*
693 *mation of videos and images*, pages 41–62. Springer. 693
- Yujun Hou, Matias Quintana, Maxim Khomiakov, Win- 694
695 ston Yap, Jiani Ouyang, Koichi Ito, Zeyu Wang, 695
696 Tianhong Zhao, and Filip Biljecki. 2024. Global 696
697 streetscapes—a comprehensive dataset of 10 million 697
698 street-level images across 688 cities for urban science 698
699 and analytics. *ISPRS Journal of Photogrammetry 699*
700 *and Remote Sensing*, 215:216–238. 700
- Wenyue Hua, Kaijie Zhu, Lingyao Li, Lizhou Fan, 701
702 Shuhang Lin, Mingyu Jin, Haochen Xue, Zelong 702
703 Li, Jindong Wang, and Yongfeng Zhang. 2024. Dis- 703
704 entangling logic: The role of context in large lan- 704
705 guage model reasoning capabilities. *arXiv preprint 705*
706 *arXiv:2406.02787*. 706

707	Neel Jay, Hieu Minh Nguyen, Trung Dung Hoang, and Jacob Haimès. 2025. Evaluating precise geolocation inference capabilities of vision language models. <i>arXiv preprint arXiv:2502.14412</i> .	
708		
709		
710		
711	Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In <i>Proc. Int. Conf. on Machine Learning (ICML)</i> , pages 4904–4916.	
712		
713		
714		
715		
716		
717	Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. 2017. Learned contextual feature reweighting for image geo-localization. In <i>Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition</i> , pages 2136–2145.	
718		
719		
720		
721		
722	VijayaKumar Kadha, Sambit Bakshi, and Santos Kumar Das. 2025. Unravelling digital forgeries: A systematic survey on image manipulation detection and localization. <i>ACM Computing Surveys</i> , 57(12):1–36.	
723		
724		
725		
726	Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, and 1 others. 2023. Geo-bench: Toward foundation models for earth monitoring. <i>Advances in Neural Information Processing Systems</i> , 36:51080–51093.	
727		
728		
729		
730		
731		
732		
733	Hao Li, Fabian Deuser, Wenping Yin, Xuanshu Luo, Paul Walther, Gengchen Mai, Wei Huang, and Martin Werner. 2025a. Cross-view geolocalization and disaster mapping with street-view and vhr satellite imagery: A case study of hurricane ian. <i>ISPRS Journal of Photogrammetry and Remote Sensing</i> , 220:841–854.	
734		
735		
736		
737		
738		
739		
740	Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. 2025b. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. <i>arXiv preprint arXiv:2506.14674</i> .	
741		
742		
743		
744	Tsung-Yi Lin, Serge Belongie, and James Hays. 2013. Cross-view image geolocalization. In <i>Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition</i> , pages 891–898.	
745		
746		
747		
748	Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In <i>Proceedings of the European conference on computer vision (ECCV)</i> , pages 563–579.	
749		
750		
751		
752		
753	Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. 2022. Where in the world is this image? transformer-based geo-localization in the wild. In <i>European Conference on Computer Vision</i> , pages 196–215. Springer.	
754		
755		
756		
757		
758	Alec Radford, Jong Wook Kim, and Christopher <i>et al.</i> Hallacy. 2021. Learning transferable visual models from natural language supervision. In <i>Proc. Int. Conf. on Machine Learning (ICML)</i> , pages 8748–8763.	
759		
760		
761		
	Noe Samano, Mengjie Zhou, and Andrew Calway. 2020. You are here: Geolocation by embedding maps and images. In <i>European Conference on Computer Vision</i> , pages 502–518. Springer.	762 763 764 765
	Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. 2018. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In <i>Proceedings of the European Conference on Computer Vision (ECCV)</i> , pages 536–551.	766 767 768 769 770
	Zhicheng Shi, Yang Li, Siyu Li, and Jiebo Luo. 2020. SAFA: Structure-aware feature aggregation for cross-view image-based geo-localization. In <i>Proc. ACM Int. Conf. on Multimedia (MM)</i> , pages 1633–1641.	771 772 773 774
	Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. 2025. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. <i>arXiv preprint arXiv:2504.03164</i> .	775 776 777 778 779
	Yicong Tian, Chen Chen, and Mubarak Shah. 2017. Cross-view image matching for geo-localization in urban environments. In <i>Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition</i> , pages 3608–3616.	780 781 782 783 784
	Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In <i>Advances in Neural Information Processing Systems (NeurIPS)</i> .	785 786 787 788 789
	Nam Vo and David W. Jacobs. 2017. Revisiting IM2GPS in the deep learning era. <i>arXiv preprint arXiv:1705.04838</i> .	790 791 792
	Zhiqiang Wang, Dejia Xu, Rana Muhammad Shahroz Khan, Yanbin Lin, Zhiwen Fan, and Xingquan Zhu. 2024. Lmgeo: Benchmarking large language models on image geolocation in-the-wild. <i>arXiv preprint arXiv:2405.20363</i> .	793 794 795 796 797
	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in neural information processing systems</i> , 35:24824–24837.	798 799 800 801 802 803
	Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet - photo geolocation with convolutional neural networks. In <i>European Conference on Computer Vision (ECCV)</i> , pages 37–55. Springer.	804 805 806 807
	Scott Workman, Richard Souvenir, and Nathan Jacobs. 2015. Wide-area image geolocalization with aerial reference imagery. In <i>Proceedings of the IEEE International Conference on Computer Vision</i> , pages 3961–3969.	808 809 810 811 812
	Meiliu Wu and Qunying Huang. 2022. Im2city: image geo-localization via multi-modal learning. In <i>Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery</i> , pages 50–61.	813 814 815 816 817

818 Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan
819 Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind’s
820 eye of llms: visualization-of-thought elicits spatial
821 reasoning in large language models. *Advances in*
822 *Neural Information Processing Systems*, 37:90277–
823 90317.

824 Shixiong Xu, Chenghao Zhang, Lubin Fan, Gaofeng
825 Meng, Shiming Xiang, and Jieping Ye. 2024. Ad-
826 dressclip: Empowering vision-language models for
827 city-wide image address localization. In *Euro-*
828 *pean Conference on Computer Vision*, pages 76–92.
829 Springer.

830 An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang,
831 and Julian McAuley. 2023. Personalized showcases:
832 Generating multi-modal explanations for recommen-
833 dations. In *Proceedings of the 46th International*
834 *ACM SIGIR Conference on Research and Develop-*
835 *ment in Information Retrieval*, pages 2251–2255.

836 Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn
837 Han, Li Fei-Fei, and Saining Xie. 2025. Thinking
838 in space: How multimodal large language models
839 see, remember, and recall spaces. In *Proceedings of*
840 *the Computer Vision and Pattern Recognition Con-*
841 *ference*, pages 10632–10643.

842 Qiang Yi and Lianlei Shan. 2025. Geolocsft: Effi-
843 cient visual geolocation via supervised fine-tuning
844 of multimodal foundation models. *arXiv preprint*
845 *arXiv:2506.01277*.

846 Wenping Yin, Yong Xue, Ziqi Liu, Hao Li, and Martin
847 Werner. 2025. Llm-enhanced disaster geolocaliza-
848 tion using implicit geoinformation from multimodal
849 data: A case study of hurricane harvey. *International*
850 *Journal of Applied Earth Observation and Geoinfor-*
851 *mation*, 137:104423.

852 Yanhua Zhong, Yuqiang Wu, Sheng Zheng, Yi Yang,
853 and Zhiwu Ma. 2021. **VIGOR: Cross-view image**
854 **geo-localization beyond one-to-one retrieval**. In *Proc.*
855 *IEEE/CVF Conf. on Computer Vision and Pattern*
856 *Recognition (CVPR)*, pages 8636–8645.

857 Zhongliang Zhou, Jieli Zhang, Zihan Guan, Mengxuan
858 Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai.
859 2024. **Img2loc: Revisiting image geolocation**
860 **using multi-modality foundation models and image-**
861 **based retrieval-augmented generation**. In *Proceed-*
862 *ings of the 47th international acm sigir conference on*
863 *research and development in information retrieval*,
864 pages 2749–2754.

865 Sijie Zhu, Mubarak Shah, and Chen Chen. 2022.
866 **Transgeo: Transformer is all you need for cross-**
867 **view image geo-localization**. In *Proceedings of the*
868 *IEEE/CVF Conference on Computer Vision and Pat-*
869 *tern Recognition*, pages 1162–1171.

A Sample Images 870

871 Sample images from our three datasets in the bench-
872 mark are presented in Figure 7. 872

B Image Data Distribution 873

874 The sample distribution based on annotated fea-
875 tures provided by the NUS team (Hou et al., 2024)
876 for Dataset-GSS is shown in Table 2. Table 3
877 presents the distribution of features extracted from
878 model outputs by gemini-2.5-pro across our bench-
879 mark datasets. Please note that these features are
880 not human-annotated, as our goal is not to evaluate
881 the influence of specific image attributes on LLMs’
882 geolocation performance. 882

Feature	Value	Count
lighting_condition	day	5601
lighting_condition	dusk/dawn	319
lighting_condition	night	232
view_direction	front/back	5451
view_direction	side	701
weather	clear	4638
weather	cloudy	1189
weather	snowy	175
weather	rainy	147
weather	foggy	3

Table 2: Sample distribution of Dataset-GSS.

Feature	Value	GSS	UPC	PCW
environment	outdoor	6144	2400	264
	indoor	2	522	4
	mixed	5	5	2
scene_type	street	4636	120	33
	mixed	1072	685	141
	scenery	414	403	56
	building	29	1719	40
setting	urban	4350	595	149
	suburban	998	1675	81
	rural	755	373	8
	natural	48	283	31
	mixed	0	1	1

Table 3: Sample distribution based on gemini-pro-2.5.

C Prompt Design 883

884 The prompt design shown below (Figure 11) en-
885 ables LLMs to perform image geolocation from
886 a given image. The prompt design shown be-
887 low is guided by two key strategies: (i) chain-of-
888 thought reasoning (Wei et al., 2022) and (ii) few-
889 shot learner guidance (Brown et al., 2020). Each
890 model is instructed to follow a structured reasoning
891 process: first, analyze the visual content (e.g., land-
892 marks, textual signage, and cultural indicators); 892

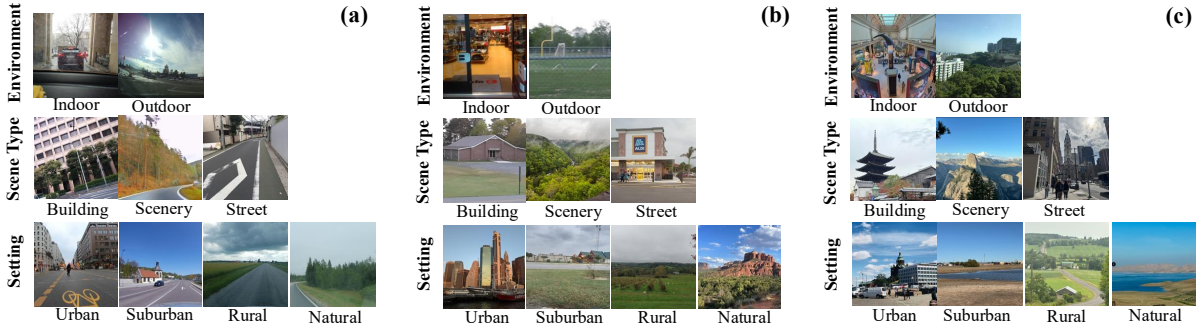


Figure 7: Sample images from the benchmark datasets: (a) Dataset-GSS; (b) Dataset-CUS; (c) Dataset-PCW.

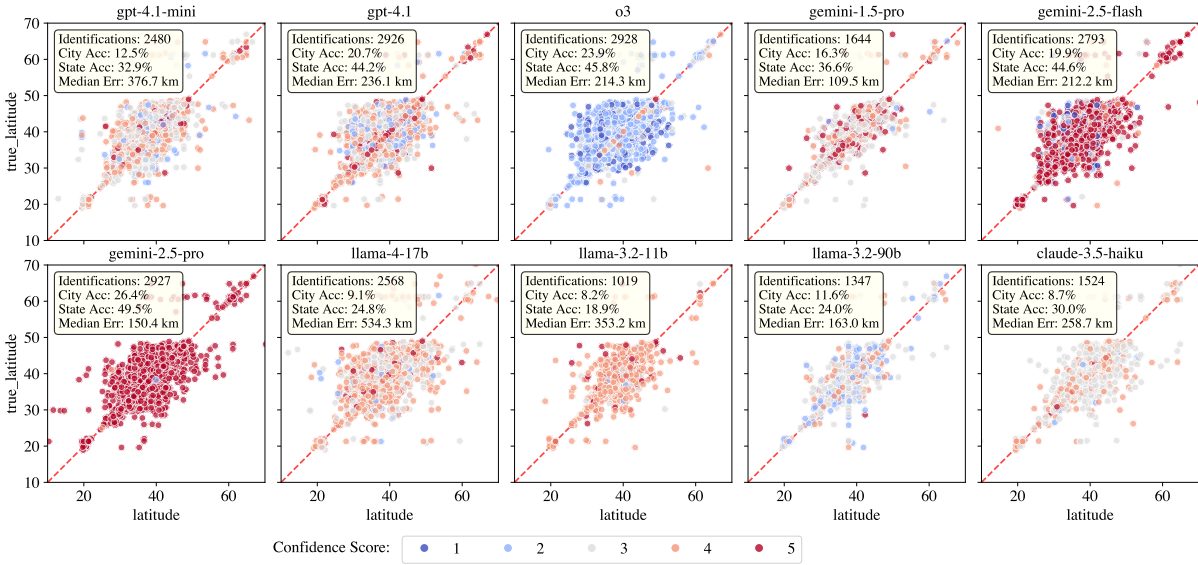


Figure 8: Benchmark performance based on latitude prediction for Dataset-UPC. Perfect predictions lie on the red dashed diagonal line. The proximity of points to this line indicates the model’s accuracy, with the color of each point representing the model’s self-reported confidence score on a scale of 1 to 5.

then, apply chain-of-thought reasoning to integrate these observations; and finally, produce geocoded predictions accompanied by confidence scores.

D Latitude Analysis on Datasets-UPC

Figure 8 presents model performance on latitude prediction for Dataset-UPC, complementing our earlier analysis of latitude on Dataset-GSS.

E Longitude Analysis on Datasets-GSS & -UPC

Figure 9 presents model performance on longitude prediction, complementing our earlier analysis of latitude. Overall trends are consistent: closed-source models generally achieve higher accuracy, while open-source models exhibit greater variance. In Dataset-GSS (Figure 9, most models produce predictions that align well with the ground-truth

longitudes, as evidenced by tight clustering along the diagonal. Closed-source models such as gemini-2.5-pro, o3, and gpt-4.1 demonstrate strong alignment, whereas open-source models show more scattered predictions across the longitude range.

Despite the similarity with latitude analysis (Figure 2, the longitude scatter plots reveal clear differences. For Dataset-UPC (Figure 10, predictions form a distinctive cube-like shape: both predicted (x-axis) and true (y-axis) longitudes fall between approximately -120° and -66° , consistent with the full longitudinal extent of the U.S. continent. However, rather than aligning along the diagonal, predictions are densely concentrated in the middle of this range. This pattern is observed across most models, regardless of architecture or source. This pattern further suggests that, in the absence of clear visual cues, models often default to generic or centrally located longitude values. While such pre-

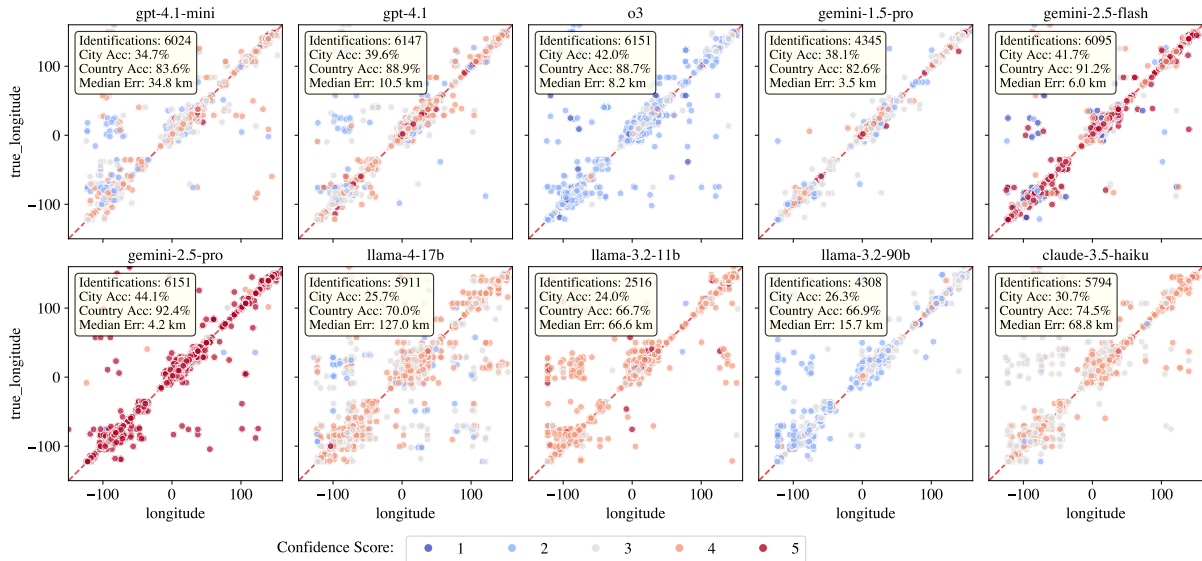


Figure 9: Benchmark performance based on longitude prediction for Dataset-GSS. Perfect predictions lie on the red dashed diagonal line. The proximity of points to this line indicates the model’s accuracy, with the color of each point representing the model’s self-reported confidence score on a scale of 1 to 5.

928 dictions typically fall within plausible U.S. bounds, 929 they are frequently assigned to widely dispersed 930 ground-truth locations. A likely explanation is that, 931 unlike latitude—which can be inferred from en- 932 vironmental features such as vegetation, climate, 933 or sun angle—longitude provides fewer visually 934 distinctive signals across many U.S. scenes. Conse- 935 quently, even high-performing models tend to pro- 936 duce centralized, non-specific longitude estimates 937 when visual evidence is ambiguous.

938 F Benchmark Performance on 939 Dataset-PCW

940 Table 4 shows the performance of models’ geoloca- 941 tion prediction on unseen images from Dataset- 942 PCW. The results reveal similar performance 943 trends across models as observed with the other 944 two datasets (see Table 1), with closed-source mod- 945 els maintaining an edge over open-source alterna- 946 tives. This time, however, o3 achieves the highest 947 city-level accuracy at 60.8%, closely followed by 948 gpt-4.1 at 60.0%, while gemini-2.5-pro displays 949 precision with the lowest median distance error of 950 0.7 km. In contrast, open-source models lag be- 951 hind: llama-4-17b and llama-3.2-11b reach only 952 40.8% and 41.9% city-level accuracy, respectively, 953 with higher median errors of 34.3 km and 10.5 km.

954 Similarly, the computational cost analysis mir- 955 rors earlier patterns. While llama-3.2-11b offers 956 the most cost-effective solution at \$0.00014 per 957 request with only 783 average tokens, this efficiency

958 comes at the expense of reduced accuracy. Con- 959 versely, o3 delivers superior performance but at 960 a premium cost of \$0.01117 per request. Among 961 the models, gemini-1.5-pro and gemini-2.5-flash 962 show competitive performance (54.4% city accu- 963 racy, 1.5 km median error) with relatively efficient 964 cost (1,180 tokens, \$0.00244 per request). These 965 findings show that while larger models often per- 966 form better, the trade-off between model size, cost, 967 and accuracy remains nuanced for practical cases.

Model	City Acc (%)	Country Acc (%)	Median Dist (km)	Avg Tokens	Avg Cost (\$)
gpt-4.1-mini	51.8%	95.3%	2.9	3041	.00153
gpt-4.1	60.0%	97.7%	1.4	1621	.00476
o3	60.8%	96.9%	0.9	2350	.01117
gemini-1.5-pro	54.4%	93.3%	1.5	1180	.00244
gemini-2.5-flash	50.2%	96.3%	1.9	1346	.00133
gemini-2.5-pro	58.7%	97.0%	0.7	1345	.00534
llama-4-17b	40.8%	93.4%	34.3	2823	.00089
llama-3.2-11b	41.9%	89.4%	10.5	783	.00014
llama-3.2-90b	50.0%	89.7%	1.9	6848	.00822
claude-3.5-haiku	42.0%	94.1%	10.0	2350	.00266

Table 4: LLM benchmark performance on Dataset-PCW.

968 G Extended Feature Attribution Results 969

969 We perform a post-hoc diagnostic to quantify how 970 well each model’s own extracted semantic fields 971 predict its success or failure. For each model, we 972 regress its city-level prediction correctness (binary) 973 and its log-transformed distance error (continuous) 974 against these four fields using logistic and ridge 975 regression. Results are reported on all three bench- 976 mark splits.

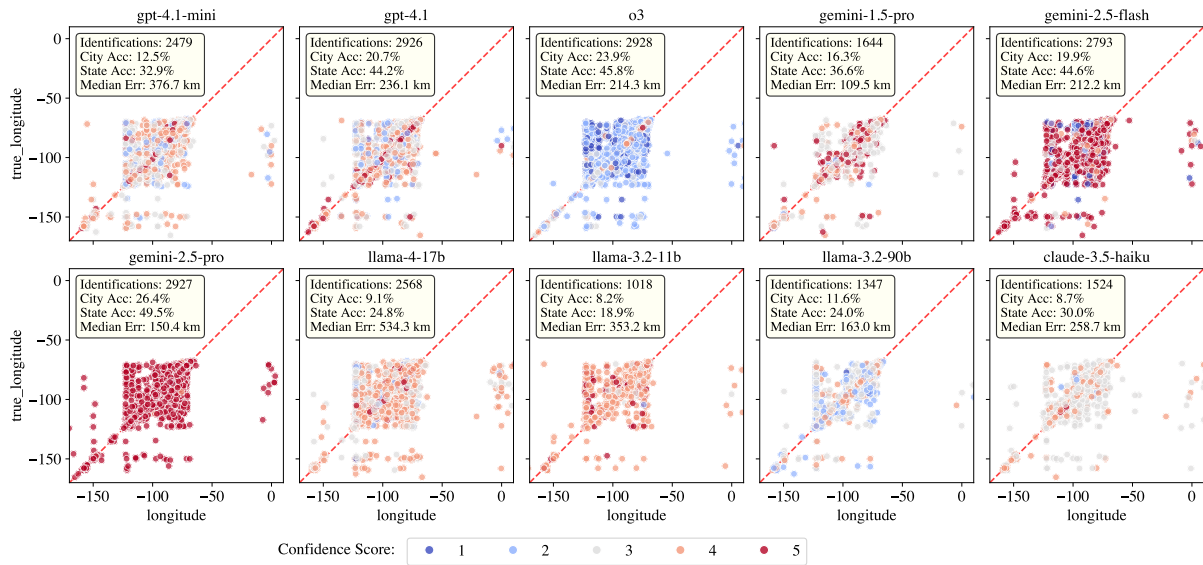


Figure 10: Benchmark performance based on longitude prediction for Dataset-UPC. Perfect predictions lie on the red dashed diagonal line. The proximity of points to this line indicates the model’s accuracy, with the color of each point representing the model’s self-reported confidence score on a scale of 1 to 5.

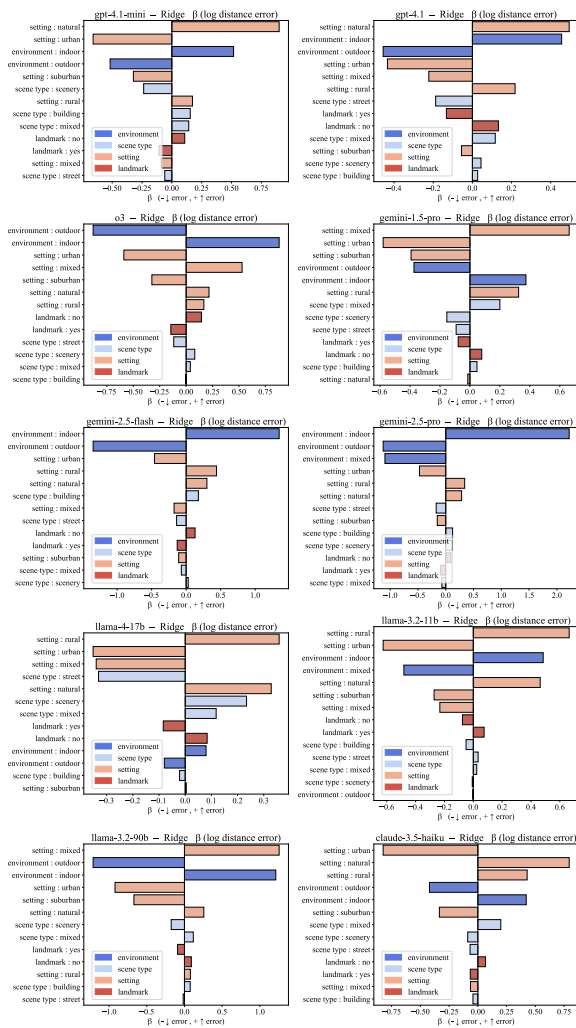


Figure 12: Ridge regression coefficients on Dataset-GSS, using log-distance error as the response. Negative values indicate cues that decrease error; positive values indicate cues that increase error.

This diagnostic analysis reveals how effectively current models ground their geolocation predictions in structured semantic cues, offering insight into their underlying visual-geospatial reasoning capabilities. On Dataset-GSS, which primarily features outdoor urban scenes rich in spatial context, models exhibit stronger alignment between their predicted semantic fields and their final performance. Regression models based on these four elements (environment, scene type, setting, and landmark) explain up to 17% of the variance in log-distance error (R^2), with gemini-2.5-pro consistently leading in both interpretability and precision. Its ability to generate semantically rich, spatially aligned tags translates into the lowest MAE_{\log} and the highest R^2 , suggesting a high degree of internal consistency between scene understanding and long coordinate prediction. In contrast, although llama-4-17b achieves the highest city-level classification accuracy on both Datasets-GSS and -UPC, its large distance errors indicate that these correct predictions may stem from over-reliance on salient, shortcut features—such as recognizable logos or architectural patterns—without engaging in deeper spatial reasoning or refinement.

977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001

Prompt design for image geolocation.

```

SYSTEM_PROMPT = "You are an AI assistant
specialized in
geocoding analysis from images."

IMAGE_LOCATION_PROMPT =
You are given an image. Use a step-by-step
(chain-of-thought) reasoning process to
infer the most likely location.

Structure your response in three sections:
1. Image Information - Describe the general
setting of the image.
2. Reasoning - Provide detailed
justifications using observable visual
cues.
3. image geolocation - Output your
best-guess location with a confidence
score.

Your response must be valid JSON in this exact
format:
{ "image_information": {
  "environment": "indoor|outdoor",
  "scene_type":
    "building|scenery|street|mixed",
  "setting": "urban|suburban|rural|natural"},
  "reasoning": {
    "landmark_recognition": "e.g., iconic
structures,
architectural landmarks, natural
features",
    "text_and_signage": "e.g., street signs,
business
names, license plates, visible text",
    "cultural_indicators": "e.g., architectural
styles,
regional patterns, cultural elements",
    "spatial_context": "e.g., geographic
relationships,
infrastructure patterns"},
  "reverse_geocoding": {
    "confidence": "1|2|3|4|5",
    "address": {
      "street": "street address",
      "city": "city name",
      "state": "state/province",
      "country": "country name"},
    "coordinates": {
      "latitude": "decimal degrees",
      "longitude": "decimal degrees"}}}

Example:
{ "image_information": {
  "environment": "outdoor",
  "scene_type": "building",
  "setting": "urban"},
  "reasoning": {
    "landmark_recognition": "Empire State
Building
with distinctive Art Deco spire clearly
visible",
    "text_and_signage": "NYC taxi markings
and New York license plates",
    "cultural_indicators": "Dense Manhattan
urban
layout with characteristic skyscraper
arrangement",
    "spatial_context": "Midtown Manhattan
street grid
and building density patterns"},
  "reverse_geocoding": {
    "confidence": "5",
    "address": {
      "street": "350 5th Ave",
      "city": "New York",
      "state": "NY",
      "country": "USA"},
    "coordinates": {
      "latitude": "40.7484",
      "longitude": "-73.9857"}}}

Guidance:
- Return only valid JSON, no additional text
- Return a confidence score between 1 (low) and
5 (high)
- For image geolocation, return only one
location that you are mostly confident

```

Figure 11: Prompt used to elicit structured geolocation reasoning and prediction from LLMs.

Model	Accuracy \uparrow	AUC \uparrow	MAE _{log} \downarrow	R ² \uparrow
gpt-4.1-mini	0.691	0.736	0.761	0.102
gpt-4.1	0.680	0.753	0.644	0.133
o3	0.707	0.777	0.633	0.141
gemini-1.5-pro	0.635	0.682	0.585	0.099
gemini-2.5-flash	0.699	0.760	0.575	0.165
gemini-2.5-pro	0.691	0.742	0.547	0.171
llama-4-17b	0.733	0.738	1.008	0.091
llama-3.2-11b	0.667	0.617	1.013	0.085
llama-3.2-90b	0.647	0.667	0.763	0.108
claude-3.5-haiku	0.676	0.747	0.836	0.140

Table 5: Field-driven performance on Dataset-GSS.

Model	Accuracy \uparrow	AUC \uparrow	MAE _{log} \downarrow	R ² \uparrow
gpt-4.1-mini	0.852	0.695	0.926	0.083
gpt-4.1	0.793	0.739	0.908	0.177
o3	0.778	0.737	0.920	0.176
gemini-1.5-pro	0.755	0.681	0.834	0.030
gemini-2.5-flash	0.801	0.685	0.888	0.129
gemini-2.5-pro	0.759	0.687	0.874	0.159
llama-4-17b	0.897	0.686	1.030	0.053
llama-3.2-11b	0.858	0.666	1.015	0.071
llama-3.2-90b	0.759	0.689	1.029	0.087
claude-3.5-haiku	0.842	0.713	0.942	0.051

Table 6: Field-driven performance on Dataset-UPC.

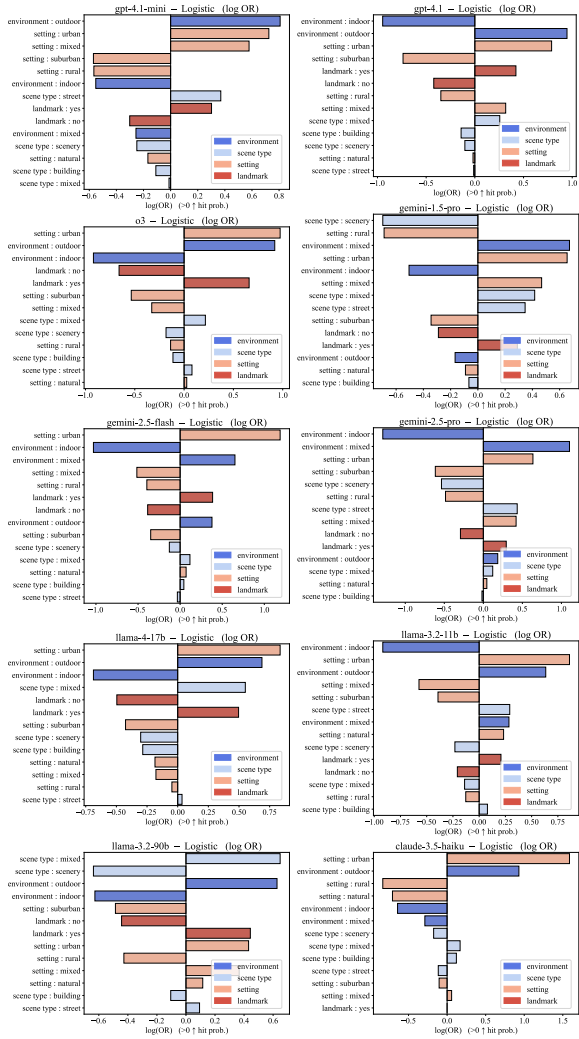


Figure 13: Log-odds coefficients from diagnostic logistic regression on Dataset-UPC. Positive values increase the likelihood of correct city prediction; negative values decrease it.

Model	Accuracy \uparrow	AUC \uparrow	MAE _{log} \downarrow	R ² \uparrow
gpt-4.1-mini	0.741	0.765	0.735	0.200
gpt-4.1	0.727	0.756	0.551	0.282
o3	0.742	0.770	0.541	0.271
gemini-1.5-pro	0.703	0.750	0.477	0.089
gemini-2.5-flash	0.757	0.790	0.630	0.244
gemini-2.5-pro	0.752	0.787	0.559	0.238
llama-4.17b	0.684	0.770	0.951	0.119
llama-3.2-11b	0.676	0.737	0.959	0.104
llama-3.2-90b	0.771	0.824	0.495	0.389
claude-3.5-haiku	0.766	0.833	0.783	0.217

Table 7: Field-driven performance on Dataset-PCW.

The diagnostic further underscores the challenges posed by visually sparse or ambiguous environments. On Dataset-UPC, where scenes (e.g., indoors, signage fragments) often lack distinct geographic structures, field-based regressions degrade sharply across all models. In particular, llama-3.2-90b demonstrates that large open-weight models can internalize meaningful spatial priors when appropriately scaled and tuned, achieving high accuracy and strong R² in PCW. Still, the overall

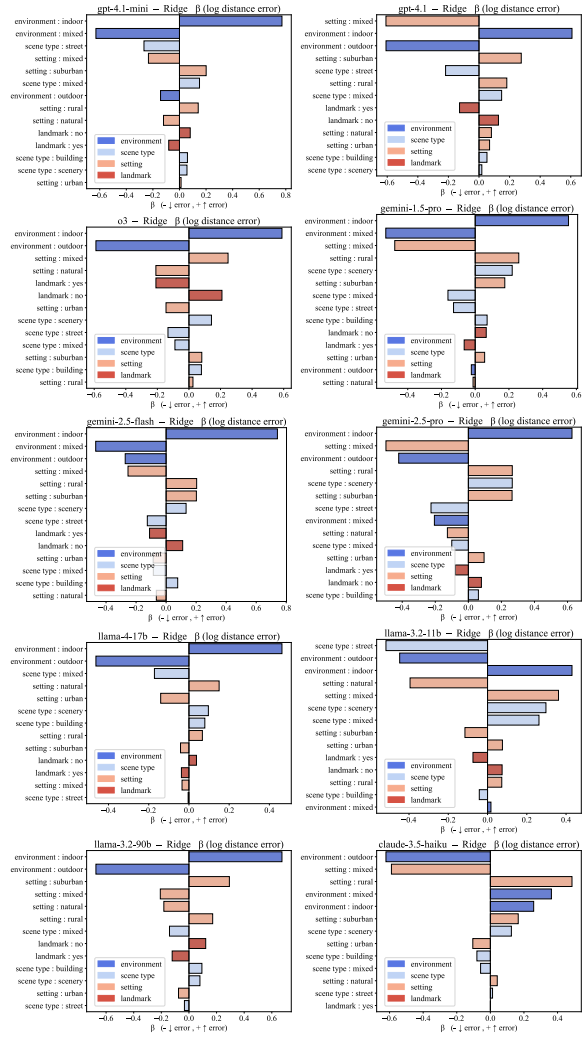


Figure 14: Ridge regression coefficients on Dataset-UPC, using log-distance error as the response. Negative values indicate cues that decrease error; positive values indicate cues that increase error.

explanatory power of these four fields remains limited: Even in the best case, the variance in distance errors over 60% remains unaccounted for. This gap suggests that high-level tags like “rural” or “landmark” only scratch the surface of the complex reasoning chain needed for accurate localization. Finer-grained semantics, such as regional architecture, vegetation type, textual cues, and spatial layout, are likely necessary to capture how LLMs reason about place fully.

Figure 12 displays the ridge-regression coefficients for the four semantic fields on Dataset-GSS, with the log-distance error as the dependent variable. Each bar shows the directional contribution of a cue: larger negative values (left of the axis) correspond to stronger error reduction, whereas positive values indicate error amplification. Among all cues, setting and environment are the most influential. Outdoor scenes receive the most negative weights,



Figure 15: Mixed unigram-bigram word cloud derived from the reasoning fields: (a) gpt-4.1-mini, (b) gpt-4.1, (c) gemini-1.5-pro, (d) gemini-2.5-flash, (e) gemini-2.5-pro, (f) llama-4-17b, (g) llama-3.2-11b, (h) llama-3.2-90b, (i) claude-3.5-haiku.

confirming that external spatial anchors (skylines, road geometry, and signage) can substantially improve geolocation prediction. By contrast, indoor scenes attract large positive weights (notably for gemini-2.5-pro), underscoring how precision degrades once exterior context is lost. The setting field exhibits a similar polarity: urban imagery lowers error, whereas rural or natural settings raise it. This likely reflects biases in several models training data, which may overrepresent urban environments and their characteristic spatial patterns such as grid layouts or dense built form.

Figures 13 extends this analysis to Dataset-UPC. The predictive power of semantic fields declines sharply: R^2 values drop and landmark coefficients collapse toward zero, indicating reduced access to external anchors. In contrast, proprietary models like gemini-2.5-pro and o3 adopt flatter, more conservative profiles, yielding fewer high-confidence errors. Overall, model performance peaks when scenes are urban, outdoor, and landmark-rich, but degrades rapidly under distribution shift.

H Extended Word-Cloud Analysis

Figure 15 presents the word clouds for the nine remaining models. First, most models rely on traffic-related terms such as “sign,” “street,” and “road.” This common vocabulary highlights a

general model preference for OCR-visible and regulation-focused features, especially both gpt-4.1 and gpt-4.1-mini prominently feature these traffic cues. The gemini models similarly emphasize precise geolocation cues but differ notably in style: gemini-1.5-pro uniquely includes the placeholder “none,” explicitly signaling missing visual evidence, whereas gemini-2.5-flash and gemini-2.5-pro adopt a more concise, instruction-oriented vocabulary with words such as “specific” and “taken.” In contrast, the llama models predominantly rely on vague tokens including “text,” “layout,” and “presence,” with llama-3.2-11b occasionally producing debug-like outputs and llama-90b displaying some spatial awareness via the terms “street,” “landmarks.” claude-3.5-haiku distinctly diverges from the other models by reducing its dependence on traffic terms, instead emphasizing scene descriptors such as “architectural,” “trees,” and “design.” Collectively, these lexical patterns align with the quantitative results presented earlier: most models rely on abundant and discrete visual cues.

I Sample Output by LLMs

Tables 8–10 present sampled outputs from o3 and gemini-2.5-pro across our benchmark datasets. Each sample includes the model’s geolocation prediction along with its corresponding reasoning,

1085 structured according to our standardized reasoning
1086 template. The “ground truth” column provides the
1087 correct location for comparison. The tables feature
1088 both correct and incorrect predictions to illustrate
1089 the strengths and failure cases of the models.

1090 We observe consistent model behaviors. When
1091 images contain clear textual or architectural an-
1092 chors, both o3 and gemini-2.5-pro achieve high
1093 accuracy. Notable examples include gemini-2.5-
1094 pro localizing Jakarta via an Indonesian traffic sign
1095 (Table 8, row 1), identifying a Seattle-area free-
1096 way from overhead signage (row 3), and pinpoint-
1097 ing Riga from a Latvian facade (row 5). It also
1098 correctly resolves the “American Clock & Watch
1099 Museum” in Bristol, CT (Table 9, row 2). Like-
1100 wise, o3 infers the 7th arrondissement in Paris from
1101 French café signage (Table 8, row 2). However,
1102 this confidence often carries into near-misses: o3
1103 misattributes Philadelphia’s City Hall to Honolulu
1104 (Table 9, row 1), while gemini-2.5-pro maps a Mex-
1105 ican coastal bar to New York City (Table 10, row 6),
1106 both with confidence 4–5—highlighting a bias to-
1107 ward visually salient urban hubs. In text-sparse
1108 scenes, their strategies diverge: gemini-2.5-pro ex-
1109 plicitly notes missing text (e.g., Table 10, row 2),
1110 while o3 defaults to stylistic heuristics, referenc-
1111 ing “Mediterranean tiles” in Fort Worth (Table 9,
1112 row 6) or “Midwest campus” traits in Duluth (Ta-
1113 ble 10, row 4). Illumination also affects perfor-
1114 mance: gemini-2.5-pro correctly predicts Macki-
1115 naw City at night using bridge geometry and aurora
1116 cues (Table 10, row 2), whereas o3 falters under
1117 low contrast in Lubbock (row 3). Both models
1118 struggle most in visually repetitive rural or coastal
1119 scenes, with o3 drifting from Luxembourg to Ger-
1120 many (Table 8, row 7) and gemini-2.5-pro misplac-
1121 ing a Mexican inlet to Florida (Table 10, row 7),
1122 relying on vague terms like “grassy embankment”
1123 or “sandy shoreline.”

Image	Ground-Truth	LLM Output
	Address: Jakarta, Indonesia Coordinates: -6.2088, 106.8459	Model: gemini-2.5-pro environment: outdoor scene_type: street setting: urban landmark_recognition: The tall, light-colored residential building visible in the background is the Permata Senayan apartment complex in Jakarta. text_and_signage: Multiple road signs are visible in Indonesian. The most prominent yellow sign reads "Hati - Hati Lintasan Bus Transjakarta," which translates to "Be Careful, Transjakarta Bus Crossing." Another sign below it reads "HATI-HATI PORTAL BUSWAY." Transjakarta is the Bus Rapid Transit (BRT) system exclusive to Jakarta, Indonesia. A vehicle license plate starting with the letter 'B' is also partially visible, which is the vehicle registration code for the Jakarta metropolitan area. cultural_indicators: The language on the signs is Indonesian. The high density of traffic, including numerous motorcycles, is characteristic of major cities in Indonesia, particularly Jakarta. spatial_context: The scene depicts a crossing for a dedicated busway. The combination of the specific sign configuration, the location of the Permata Senayan apartment building in the background, the fence on the right (belonging to the Gelora Bung Karno complex), and the orientation of the road points directly to Jalan Tentara Pelajar, near the Palmerah railway station in Central Jakarta. The camera is facing southeast along this road. Confidence: 5 Address: Jalan Tentara Pelajar, Jakarta, JAKARTA, Indonesia Coordinates: -6.2117, 106.7977
	Address: Paris, France Coordinates: 48.8523, 2.3541	Model: o3 environment: outdoor scene_type: street setting: urban landmark_recognition: Haussmann-style cream stone buildings with wrought-iron balconies typical of central Paris streets; red café awnings common on well-known pedestrian market streets such as Rue Cler. text_and_signage: Word "CAFÉ" in French clearly visible on red awning; menu board beside wooden café frontage also in French. cultural_indicators: Parisian sidewalk-café culture, narrow cobbled roadway shared by pedestrians, architectural moldings and window shutters characteristic of late-19th-century Paris buildings. spatial_context: Pedestrian market street lined with small shops, cafés, and outdoor produce stalls matches the layout of the 7th-arrondissement market street Rue Cler located between Avenue de la Motte-Picquet and Rue de Grenelle. Confidence: 3 Address: Rue Cler, Paris, ÎLE-DE-FRANCE, France Coordinates: 48.8569, 2.3062
	Address: Shoreline, USA Coordinates: 47.7633, -122.3239	Model: gemini-2.5-pro environment: outdoor scene_type: street setting: suburban landmark_recognition: The primary landmarks are the highway infrastructure itself, including a multi-lane freeway, an overpass, and large overhead gantry signs. The presence of tall evergreen trees is characteristic of the Pacific Northwest region of the United States. text_and_signage: Multiple highway signs are clearly visible. One sign indicates 'EXIT 176' for 'NE 175th St' and 'Shoreline' is 1/2 mile ahead. Another sign lists upcoming exits and their distances: 'NE 145th St' (2 miles), 'Northgate Way' (3 1/4 miles), and 'N 85th St' (5 miles). These exits, street names, and locations (Shoreline, Northgate) are all situated along Interstate 5 in the northern part of Seattle and its immediate suburbs in Washington State. cultural_indicators: The style of the green highway signs with white text (Highway Gothic font) conforms to the Manual on Uniform Traffic Control Devices (MUTCD) used in the United States. The use of 'miles' for distance measurement is standard in the US. spatial_context: The image is taken from a vehicle traveling northbound on a major freeway. The exit numbers and destinations confirm the location is on Interstate 5 (I-5) North. Specifically, the signs for Exit 176 (NE 175th St) being 1/2 mile away, and being mounted on an overpass, place the camera's location just south of the NE 185th St overpass in Shoreline, Washington, looking north. Confidence: 5 Address: Interstate 5 N, Shoreline, WA, USA Coordinates: 47.7592, -122.3293
	Address: Hildesheim, Germany Coordinates: 52.1528, 9.9359	Model: o3 environment: outdoor scene_type: street setting: suburban landmark_recognition: No globally iconic landmark visible; single-family house with red tile roof and fenced yard on right side. text_and_signage: Large white construction-information board on the left shows an EU flag logo and multi-line text typical of EU-funded public works signs in Germany; road signs are European style (red-bordered triangular warning sign, round red-bordered 30 km/h speed limit, blue shared-path sign). cultural_indicators: Red/grey concrete block sidewalk with integrated cycle lane, metal mesh fencing, and the specific designs of the traffic signs match common German street infrastructure; house architecture (white stucco, steep red clay roof) also typical for central Germany. spatial_context: Two-lane road with modest traffic, greenery and detached housing indicate a suburban area on the edge of a medium/large German city. Confidence: 2 Address: Berliner Straße, Potsdam, BRANDENBURG, Germany Coordinates: 52.4000, 13.0400
	Address: Riga, Latvia Coordinates: 56.9492, 24.1100	Model: gemini-2.5-pro environment: outdoor scene_type: street setting: urban landmark_recognition: The architecture is characteristic of a Northern or Eastern European old town, specifically the Hanseatic style found in the Baltic states. The pink building at the end of the street with a stepped gable is a distinctive local landmark. text_and_signage: A crucial clue is the text 'BERGA FASADES' visible on the gable of the pink building. 'Fasādes' is the Latvian word for 'facades'. Below the windows on the same building, the word 'RESTORĀNS' (Latvian for 'restaurant') can be faintly seen. These linguistic cues definitively point to Latvia. cultural_indicators: The cobblestone street, the style of the buildings, and the general atmosphere are highly indicative of the Old Town (Vecrīga) in Riga, Latvia. The presence of numerous outdoor cafes is also typical for this historic tourist area. spatial_context: The photograph is taken looking down a narrow, pedestrianized street. By cross-referencing the building with 'BERGA FASĀDES' with maps of Riga, the location is identified as Tirgoņu iela, looking towards the intersection with Šķūņu iela. Confidence: 5 Address: Tirgoņu iela, Rīga, RĪGA, Latvia Coordinates: 56.9479, 24.1084
	Address: Cainta, Philippines Coordinates: 14.5765, 121.1261	Model: o3 environment: outdoor scene_type: street setting: urban landmark_recognition: No globally iconic landmark visible; scene shows small shops and residential structures typical of Southeast Asian urban neighborhoods. text_and_signage: English construction warning signs such as 'DANGER DRILLING IN PROGRESS' and 'THIS PROJECT WILL BENEFIT THE ...' plus campaign-style posters with English names (e.g., 'BEN') suggest the Philippines where English is an official language used in public signage. cultural_indicators: Dense overhead utility wires, jeepney-sized blue mini-truck, tropical coconut palm, concrete block houses with metal window grilles, and roadside sari-sari-store style stalls strongly resemble streets in Philippine barangays. spatial_context: Narrow two-way lane lined with stalls and residences, temporary construction barriers, and cluttered electrical infrastructure typical of inner neighborhoods in Metro Manila cities such as Quezon City. Confidence: 3 Address: Quezon City, METRO MANILA, Philippines Coordinates: 14.6500, 121.0500

Table 8: Representative LLM image geolocalization output in Dataset-GSS.

Image	Ground-Truth	LLM Output
	Address: Philadelphia, USA Coordinates: 39.9531, -75.1635	Model: o3 environment: outdoor scene_type: building setting: urban landmark_recognition: Ornate Second-Empire style municipal building with a central clock tower topped by a statue, matching the well-known Philadelphia City Hall and its William Penn statue text_and_signage: No readable signage visible but the clock face and tower design are distinctive enough for identification cultural_indicators: Presence of both English and Chinese text matches Honolulu Chinatown area; tropical plants and palms fit Hawaiian climate spatial_context: Street-level retail space in a mid-rise urban block typical of downtown Honolulu near Queen Emma Street corridor Confidence: 4 Address: 1280 Queen Emma Street, Honolulu, HI, USA Coordinates: 21.3148, -157.8576
	Address: Bristol, CT Coordinates: 41.6796, -72.9408	Model: gemini-2.5-pro environment: outdoor scene_type: building setting: suburban landmark_recognition: The main building is the American Clock & Watch Museum, a specific cultural landmark. An ornate street clock is also visible on the left side of the property. text_and_signage: A large sign in the foreground and a smaller sign on the building itself clearly read 'AMERICAN CLOCK & WATCH MUSEUM'. cultural_indicators: The building exhibits a classic American Colonial architectural style, characterized by its white clapboard siding, symmetrical facade, and multi-pane sash windows, which is very common in the New England region of the United States. spatial_context: The image shows a large historic house set back from a road with a lawn in front, next to a smaller, attached building and a parking area, which is consistent with the layout of a museum in a suburban or small-town setting. Confidence: 5 Address: 100 Maple St, Bristol, CT, USA Coordinates: 41.6766, -72.9461
	Address: Norfolk, NE Coordinates: 42.0158, -97.4102	Model: gemini-2.5-pro environment: indoor scene_type: building setting: urban landmark_recognition: The most significant visual feature is the ceiling and upper walls being completely covered in autographed US dollar bills. This is a distinctive tradition of a limited number of bars in the United States, making it a key identifying feature. text_and_signage: Several crucial pieces of text are visible. A sign on the lower right, partially obscured, reads 'THE MINT BAR'. A second sign, located under a television on the left side, says 'UNION BAR'. The combination of a bar featuring both 'Union Bar' and 'The Mint Bar' names, along with the dollar bill decor, is highly specific. A large white flag with a blue 'W' is prominently displayed, which is the victory flag for the Chicago Cubs baseball team. While this could be found in many bars, it fits the context of a Midwestern location. Other text on dollar bills like 'COLORADO' and 'BAYTOWN' are likely from visitors and less indicative of the primary location. cultural_indicators: The tradition of pinning signed currency to the walls is a cultural practice in some American saloons and dive bars, often accumulating over many decades. The Chicago Cubs 'W' flag is a strong element of American sports culture, particularly in the Midwest. The green, shamrock-shaped string lights suggest an Irish pub theme or St. Patrick's Day decorations, common in American bars. spatial_context: The sheer density of the currency covering almost every available surface suggests this is a very old, long-standing tradition at this establishment. This points towards a historic local bar rather than a new or chain establishment. A search for bars with these specific features (dollar bills, 'Union Bar', 'The Mint Bar') confirms the identity of a well-known location in Nebraska. Confidence: 5 Address: 1401 10th St, Gering, NE, USA Coordinates: 41.8242, -103.6625
	Address: Biloxi, MS Coordinates: 30.4461, -88.9394	Model: o3 environment: outdoor scene_type: building setting: suburban landmark_recognition: No globally iconic landmark, but the facade resembles mid-1990s—2000s U.S. suburban office-park architecture (stucco/EIFS, shallow arch window tops, square columns). text_and_signage: The only visible text is the street number "1721" mounted on the entrance canopy. cultural_indicators: Presence of large live-oak trees, beige stucco, and green-tinted ribbon windows strongly suggest a warm southern-U.S. region, particularly Texas where this design and vegetation are common. spatial_context: Looks like a stand-alone low-rise in an office park surrounded by mature trees rather than dense urban grid. Confidence: 2 Address: 1721 River Run, Fort Worth, TX, USA Coordinates: 32.7362, -97.3620
	Address: Bay Village, OH Coordinates: 41.4794, -81.8815	Model: gemini-2.5-pro environment: outdoor scene_type: mixed setting: suburban landmark_recognition: The primary identifying feature is the unique circular monkey bar structure, known as a 'RingTrek' climber. The scene also includes adjacent tennis courts with green surfaces and a chain-link fence, a swing set with distinctive beige bucket-like seats, and a small play structure with a green hexagonal roof. The specific combination and arrangement of this playground equipment are highly distinctive. text_and_signage: A brown sign is visible on a post, likely a standard park rules sign, but the text is illegible and does not contribute to the identification. cultural_indicators: The style of the playground equipment is common in modern North American public parks. The houses visible in the background are of a typical suburban American design. spatial_context: The layout of a playground directly adjacent to tennis courts, bordered by trees and with residential housing in the near distance, is a common configuration for a community park in the United States. The deciduous trees suggest a temperate climate. Confidence: 5 Address: Continental Pl, Glen Cove, NY, USA Coordinates: 40.8529, -73.6338
	Address: Colby, KS Coordinates: 39.2779, -100.9437	Model: gemini-2.5-pro environment: outdoor scene_type: scenery setting: rural landmark_recognition: The central feature is a small, yellow statue of a monkey-like figure in a raincoat. This is a known, unique roadside attraction called the 'I-80 Rainslicker Monkey' or 'Monkey Tree', which serves as an informal memorial. text_and_signage: In the background, on top of the grassy embankment, there are road signs. A cropped image clearly shows a US Interstate shield for Interstate 80 (I-80) and a US Highway shield for US Route 30 (US-30). These two highways run concurrently in this region. cultural_indicators: The object is a small, informal roadside memorial, a cultural practice found along highways in the United States, often maintained by locals to commemorate an event or person. spatial_context: The image is taken from a low vantage point looking up at a grassy embankment next to a major highway. The presence of both I-80 and US-30 signs, combined with the rolling, grassy, and sparsely populated landscape, is characteristic of southern Wyoming. Confidence: 5 Address: Interstate 80, Buford, WY, USA Coordinates: 41.2251, -105.4242

Table 9: Representative LLM image geolocalization output in Dataset-UPC.

Image	Ground-Truth	LLM Output
	Address: Philadelphia, USA Coordinates: 39.9531, -75.1635	Model: o3 environment: outdoor scene_type: building setting: urban landmark_recognition: Ornate Second-Empire style municipal building with a central clock tower topped by a statue, matching the well-known Philadelphia City Hall and its William Penn statue text_and_signage: No readable signage visible but the clock face and tower design are distinctive enough for identification cultural_indicators: French Second-Empire architecture common to late-19th-century U.S. civic buildings, specifically the gray granite and elaborate sculptural ornamentation characteristic of Philadelphia City Hall spatial_context: Large civic structure occupying a city square in the heart of a dense downtown grid, consistent with Center City Philadelphia where City Hall sits at the crossing of Broad and Market Streets confidence: 5 address: 1400 John F Kennedy Blvd, Philadelphia, PA, USA coordinates: 39.9526, -75.1636
	Address: Lubbock, USA Coordinates: 33.5162, -101.919	Model: o3 environment: outdoor scene_type: scenery setting: suburban landmark_recognition: No globally iconic landmark; scene shows a man-made pond or small lake with bald cypress trees that commonly line park lakes in the southern United States. text_and_signage: No readable text or signage visible in the image. cultural_indicators: Dormant warm-season grass and bald cypress suggest a warm climate that still experiences a cool season; this combination is common in North Texas municipal parks. spatial_context: Flat terrain, scattered single-story buildings/roofs and a roadway in the distance imply a suburban park setting rather than an isolated rural lake or dense urban waterfront. confidence: 2 address: 1405 Wilson Creek Pkwy, McKinney, TX, USA coordinates: 33.1974, -96.6232
	Address: Tampa, USA Coordinates: 28.1284, -82.368	Model: Gemini-2.5-pro environment: outdoor scene_type: street setting: suburban landmark_recognition: The image features storefronts in a modern suburban shopping plaza, not a unique landmark, but characteristic of this type of development. text_and_signage: Multiple signs are clearly visible. The main storefronts are for 'COLD STONE CREAMERY' and 'PARIS BAGUETTE'. Crucially, the street number '17519' is displayed on a pillar at the entrance of the Cold Stone Creamery. A poster in the window also provides a URL: www.ColdStoneCakes.com, which is a US-based domain. cultural_indicators: The architectural style of the buildings is typical of contemporary American suburban retail centers or strip malls. The combination of these specific international food chains (Cold Stone Creamery and Paris Baguette) is common in more affluent suburban areas in the United States. spatial_context: The image shows two businesses located directly next to each other in a shopping center. The combination of Cold Stone Creamery, Paris Baguette, and the specific address number '17519' provides a unique set of data points to query for a specific location. confidence: 5 address: 17519 W Colonial Dr, Winter Garden, FL, USA coordinates: 28.552942, -81.564633
	Address: Ann Arbor, USA Coordinates: 42.2991, -83.7129	Model: o3 environment: outdoor scene_type: scenery, setting: suburban landmark_recognition: No clearly identifiable landmark; landscape of mixed deciduous and coniferous trees with a small glimpse of a red-brick building typical of Midwestern college campuses text_and_signage: No readable street signs, business names, or license plates visible cultural_indicators: Streetlights with simple black curved arms and a red fire-hydrant are common municipal furnishings in many U.S. northern cities; red-brick institutional building hints at a school or campus built mid-20th-century, often found in the Upper Midwest spatial_context: Lightly trafficked two-lane road inside a landscaped area, snow cover, and dense spruce/pine mix are consistent with cities along Lake Superior such as Duluth, Minnesota confidence: 2 address: University Drive, Duluth, MN, USA coordinates: 46.8185, -92.0844
	Address: Mount Vernon, USA Coordinates: 38.7082, -77.087	Model: o3 environment: outdoor scene_type: mixed setting: rural landmark_recognition: Formal, walled kitchen garden (potager) with espaliered fruit trees and neat vegetable beds; small ancillary buildings with steep red-tile roofs and dormer windows suggest 17th-18th-century French estate outbuildings text_and_signage: No readable signage or text visible cultural_indicators: Clay-tile roofing, cream masonry walls, symmetrical dormers and the very structured 'potager' layout are typical of historic French châteaux estates, especially the Potager du Roi built for Louis XIV near Versailles spatial_context: Garden enclosed by low walls, outbuildings just inside, wide sky and distant tree line hint at location on slightly elevated plateau much like the Royal Kitchen Garden site south of the Palace of Versailles confidence: 2 address: 10 Rue du Maréchal Joffre, Versailles, ÎLE-DE-FRANCE, France coordinates: 48.8015, 2.133
	Address: Tulum, Mexico Coordinates: 20.2029, -87.4734	Model: Gemini-2.5-pro environment: outdoor scene_type: building setting: urban landmark_recognition: The location is identified as GITANO Garden of Love in New York City. Distinctive features include a thatched roof over a patio, real tree trunks used as support columns, a prominent disco ball, a gravel floor, and an overall 'jungle chic' or 'Tulum-inspired' aesthetic. text_and_signage: No legible text or signage is clearly visible that aids in identification. A poster on a column is intentionally obscured. cultural_indicators: The decor is highly characteristic of the Gitano brand, which blends a bohemian, jungle-like atmosphere (often associated with Tulum, Mexico) with urban nightlife elements like disco balls and moody lighting. This specific style is a strong cultural indicator for this particular chain of venues. spatial_context: The scene depicts a trendy outdoor bar/restaurant patio, a type of venue common in major urban centers. The specific combination of design elements, including the layout of the patio, the stacked terracotta pots, and the wooden spool table, was matched via comparison with online photos of the identified venue in New York City. confidence: 5 address: 76 Varick Street, New York, NY, USA coordinates: 40.72234, -74.00681

Table 10: Representative LLM image geolocalization output in Dataset-PCW.