

GeoArena: Evaluating Open-World Geographic Reasoning in Large Vision-Language Models

Pengyue Jia^{1,2*}, Yingyi Zhang^{1*}, Xiangyu Zhao^{1†}, Sharon Li²

¹Department of Data Science, City University of Hong Kong

²Department of Computer Sciences, University of Wisconsin-Madison

{jia.pengyue,yzhang6375-c}@my.cityu.edu.hk, xianzhao@cityu.edu.hk, sharonli@cs.wisc.edu

Abstract

Geographic reasoning is a fundamental cognitive capability that requires models to infer plausible locations by synthesizing visual evidence with spatial world knowledge. Despite recent advances in large vision-language models (LVLMs), existing evaluation paradigms remain largely outcome-centric, relying on static datasets and predefined labels that are conceptually misaligned with open-world geographic inference. Such outcome-centric evaluations often focus exclusively on label matching, leaving the underlying linguistic reasoning chains as unexamined black boxes. In this work, we introduce **GeoArena**, a dynamic, human-preference-based evaluation framework for benchmarking open-world geographic reasoning. GeoArena reframes evaluation as a pairwise reasoning alignment task on in-the-wild images, where human judges compare model-generated explanations based on reasoning quality, evidence synthesis, and plausibility. We deploy GeoArena as a public platform and benchmark 17 frontier LVLMs using thousands of human judgments, which complements existing benchmarks and supports the development of geographically grounded, human-aligned AI systems. We further provide detailed analyses of model behavior, including reliability of human preferences and factors influencing judgments of geographic reasoning quality. We open-source GeoArena¹ to foster future research.

1 Introduction

Geographic reasoning—the ability to infer, contextualize, and explain where an observation could plausibly be situated in the world—is a fundamental capability for intelligent systems operating in the

physical world (Mai et al., 2022). It underlies applications ranging from navigation (Venable, 2016) and disaster monitoring (Li et al., 2025a) to crime tracking (Bamigbade et al., 2024), and is commonly studied through image-based geolocation as a canonical stress test (Vo et al., 2017), as illustrated in Figure 1. Crucially, geographic reasoning is inherently open-world: real-world images exhibit high variability in visual appearances across the globe (Wilson et al., 2021), which requires the integration of visual evidence with spatial, environmental, and cultural knowledge (Li et al., 2025b). As a result, geographic inference is under-determined by the input and intrinsically involves abductive reasoning and contextual judgment rather than deterministic prediction (Wilson et al., 2021).

Despite this open-world nature, existing evaluation paradigms for geographic reasoning remain largely outcome-centric, relying on static datasets with predefined labels and metrics such as coordinate distance or categorical accuracy (Liu et al., 2024; Wang et al., 2024; Astruc et al., 2024; Li et al., 2025c; Huang et al., 2025a; Jay et al., 2025; Yerramilli et al., 2025), as shown in Table 1. While effective in controlled settings, this approach faces fundamental limitations in real-world scenarios: static benchmarks are increasingly vulnerable to data contamination from web-scale pretraining, outcome-level metrics collapse complex reasoning processes into a single label match, and authoritative ground truth is often unavailable or ambiguous for in-the-wild images. More fundamentally, label-based evaluation assumes a well-defined notion of correctness that does not hold when multiple hypotheses are plausible. In such settings, evaluating how models reason—how they select, justify, and contextualize evidence—provides a complementary and necessary signal that accuracy alone cannot capture.

To address these issues, we introduce **GeoArena**, a dynamic and realistic evaluation framework for

*Equal contribution

†Corresponding author

¹https://github.com/Applied-Machine-Learning-Lab/ACL2026_GeoArena

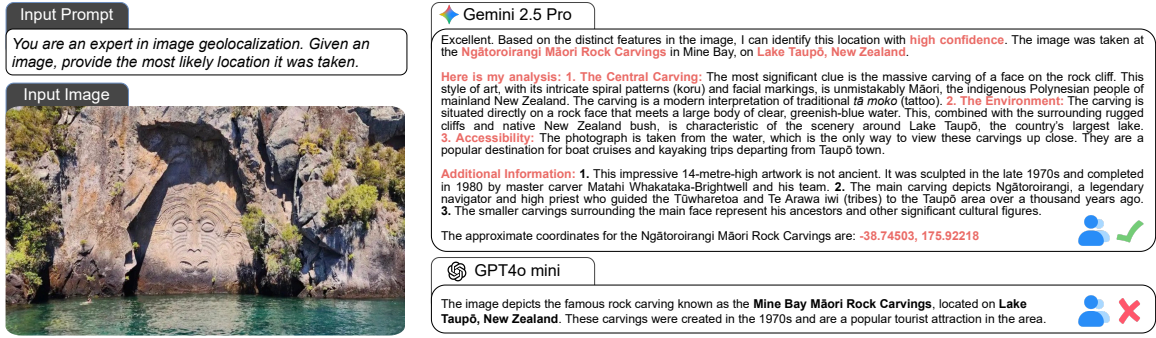


Figure 1: Example of geolocation: identifying the Ngātoroirangi Māori Rock Carvings.

Table 1: Comparison between existing benchmarks and GeoArena.

Benchmarks	Publication	Evaluation	Dynamic Datasets	Reasoning Oriented	Label Free
OSV-5M	CVPR'24	GPS	×	×	×
LLMGeo	CVPR'24	Country	×	×	×
ETHAN	Arxiv'24	GPS	×	×	×
Location-Inference	AAAI'25	GPS, Country, City	×	×	×
GeoChain	ACL'25	GPS, Pass Score	×	✓	×
FairLocator	Arxiv'25	GPS, Street, City, Country, Continent	×	×	×
IMAGEO-Bench	Arxiv'25	GPS, City, State, Country	×	×	×
GeoArena	Ours	User Preference	✓	✓	✓

benchmarking open-world geographic reasoning in LVLMs. GeoArena reframes evaluation as a pairwise reasoning alignment task: given an in-the-wild image and a geographic prompt, two anonymous models generate natural-language explanations of the image’s likely location, and human users vote for the response that better aligns with human geographic expectations in terms of reasoning quality, evidence synthesis, and plausibility. By design, GeoArena addresses the core limitations of existing evaluation paradigms: (1) it utilizes in-the-wild user contributions to mitigate data contamination; (2) it shifts the focus from outcome level to the assessment of geographic reasoning chains by evaluating the logical and linguistic quality of model explanations; (3) it employs a human-preference-driven methodology, reflecting real-world scenarios where precise metadata or ground truth label is unavailable. While human-preference-based evaluation is a standard protocol in other domains (Chiang et al., 2024; Jiang et al., 2024), such a paradigm remains a significant gap in the geographic AI community. GeoArena bridges this gap and can facilitate developing geographic reasoning systems that are both accurate and fundamentally aligned with human-centric logic and real-world utility.

Using GeoArena, we benchmark 17 frontier LVLMs through thousands of human preference judgments collected on in-the-wild images. Our

results reveal clear stratification among models, with frontier systems consistently outperforming smaller variants, and strong open-source families are closing the gap. Importantly, GeoArena produces stable rankings with high agreement between expert and crowd judgments, demonstrating that human preference signals provide a reliable and discriminative evaluation of geographic reasoning quality beyond outcome-level accuracy. Our key contributions can be summarized as follows:

1. We formalize the problem of open-world geographic reasoning evaluation and develop GeoArena, the first dynamic, preference-based framework to address long-standing issues of existing benchmarks.
2. We conduct a comprehensive analysis of the collected user inputs and voting data to demonstrate the reliability and capabilities of GeoArena.
3. We publicly release GeoArena to support research and development in related fields such as LVLm and geographic foundation models.

2 Related Work

Benchmark of Geographic Reasoning. Current research in geographic reasoning focuses on evaluating the cognitive logic of models across diverse spatial tasks. GEOBench-VLM (Danish et al., 2025) provides a framework for evaluating vision-language models on tasks such as scene understanding, object counting, and temporal analysis within geospatial contexts. Another study (Huang et al., 2025b) introduces a benchmark for geometry classification, topological relations, and direction estimation, which assesses core spatial logic using geometries encoded in GeoJSON format. Furthermore, MapEval (Dihan et al., 2024) evaluates map-based reasoning through textual, visual,

and API-based modes, identifying performance gaps in spatial inference tasks like distance and route planning. These geographic reasoning capabilities are frequently evaluated through image geolocalization, which serves as a stress test for synthesizing visual evidence with spatial world knowledge. Common evaluation datasets for geolocalization include IM2GPS (Hays and Efros, 2008) and YFCC (Thomee et al., 2016). On the benchmarking side (Yerramilli et al., 2025), LLM-Geo (Wang et al., 2024) collects data from Google Street View to evaluate various models, while Liu et al. (2024) demonstrates that incorporating Chain-of-Thought (CoT) reasoning improves results on geolocalization tasks. Jay et al. (2025) and Fair-Locator (Huang et al., 2025a) provide generalized evaluation sets and focus on urban geolocalization biases, respectively. In contrast to these static methods, we propose GeoArena, the *first dynamic and user-preference-based benchmark for geographic reasoning*. This approach provides a user-aligned platform for assessing geographic reasoning in real-world applications.

Worldwide Image Geolocalization. Worldwide image geolocalization (Jia et al., 2025a, 2026) is an interdisciplinary task that bridges geography and computer science (Zhang et al., 2026; Xu et al., 2025; Zhu et al., 2023b, 2024a), involving GeoAI (Janowicz et al., 2020; Han et al., 2025; Xu et al., 2016; Zhu et al., 2023a; Kong et al., 2024; Zhu et al., 2024b), spatial data mining (Wang et al., 2020; Zhao and Tang, 2017; Zhao et al., 2022; Zhang et al., 2023b,c; Cheng et al., 2025; Zhao et al., 2017), information retrieval (Jia et al., 2024b, 2025b; Zhang et al., 2023a), and multi-modal modeling (Wang et al., 2023; Zhang et al., 2025). In recent years, thanks to the strong world knowledge and visual understanding capabilities of LVLMs, image geolocalization has made significant progress (Vivanco Cepeda et al., 2023; Li et al., 2024; Haas et al., 2024; Dou et al., 2024; Sarkar et al., 2024; Astruc et al., 2024; Dufour et al., 2025; Li et al., 2025b). Img2Loc (Zhou et al., 2024) is the first to introduce LVLMs into image geolocalization, retrieving similar images’ information and incorporating it as prompts into the LVLm input to utilize the world knowledge acquired during pre-training to predict the image’s location. G3 (Jia et al., 2024a) further improves upon Img2Loc by optimizing both the image retrieval and reasoning processes, enabling the model to obtain more ac-

curate reference information and fully exploit the prediction potential of LVLMs.

3 Problem Definition

We formalize the evaluation of open-world geographic reasoning as a task of measuring the alignment between model-generated explanations and human geographic expectations under uncertainty. Let \mathcal{I} be the space of in-the-wild images and \mathcal{P} be the space of natural language prompts. For a pair $(I, P) \in \mathcal{I} \times \mathcal{P}$, a model M produces a response R in the space of linguistic reasoning \mathcal{R} .

We define the evaluation task as a mapping $\mathcal{A} : \mathcal{R} \times \mathcal{H} \rightarrow \mathbb{R}$, where \mathcal{H} represents the latent space of Human Expectations. The reasoning capability of a model M is evaluated as:

$$E_{\text{reasoning}}(M) = \mathbb{E}_{(I,P) \in \mathcal{I} \times \mathcal{P}}[\mathcal{A}(M(I, P), \mathcal{H})], \quad (1)$$

where \mathcal{A} measures the degree of alignment between the generated reasoning chain and the spatial logic expected by humans. This definition allows the evaluation to function in open-world scenarios where precise metadata is absent and enables a process-oriented assessment of how models synthesize visual evidence with spatial knowledge.

4 GeoArena

GeoArena is an interactive platform designed to evaluate the open-world geographic reasoning capabilities of various LVLMs. In this section, we provide a detailed description of GeoArena, including its system architecture and interface (Section 4.1), data collection process (Section 4.2), the models it evaluates (Section 4.3), and the ranking computation methods (Section 4.4).

4.1 System Architecture and Interface

GeoArena is designed to facilitate the collection of high-fidelity data on geographic reasoning. As illustrated in Figure 2, the system architecture consists of a multi-stage pipeline integrated into a public-facing interface. This design is intended to ensure that every interaction contributes to a controlled and unbiased evaluation, transforming the user interface into a rigorous scientific instrument for observing model behavior.

The **Input Component** serves as the primary gateway for data acquisition at the input level. Users contribute in-the-wild images and specify instructions. To maintain the thematic integrity

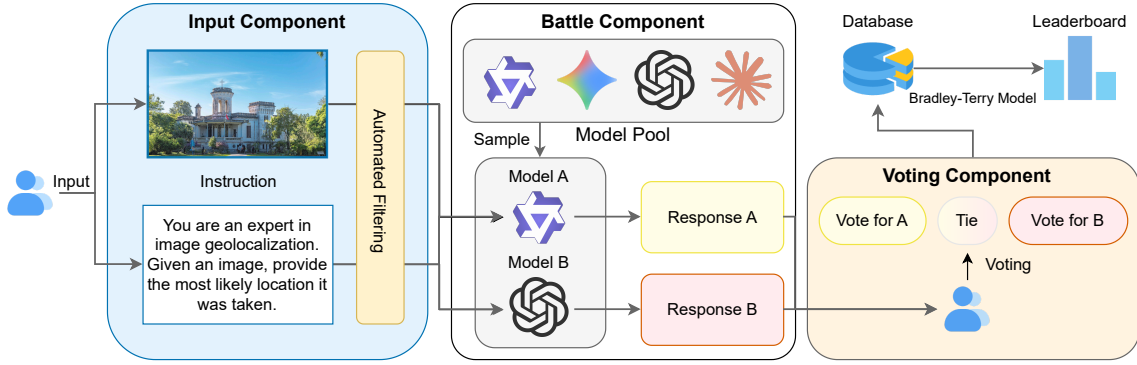


Figure 2: Overview of GeoArena.

of the platform and protect the leaderboard from noise, we implement an automated filtering mechanism as an architectural component powered by LLMs (more details in Appendix A.1). Every user prompt P is processed by a classification function $\phi(P)$ to verify that the query is relevant to geographic reasoning. This layer proactively rejects irrelevant or malicious inputs, ensuring that the collected preferences reflect genuine spatial inference capabilities. For standard interactions, a default instruction is provided based on established protocols in the field (Zhou et al., 2024; Jia et al., 2024a).

The **Battle Component** constitutes the second stage of the pipeline, addressing process neglect at the output level. Once a user input is validated, the system performs sampling to select two models, M_A and M_B , from the participant pool. Each model is required to generate a response R . This stage is engineered to transform the black-box inference of LVLMs into an observable reasoning chain, allowing for a direct assessment of how models synthesize environmental cues and spatial knowledge. The responses are presented in a side-by-side, anonymized format to ensure that the evaluation is based solely on the logical consistency and linguistic quality of the geographic reasoning.

The **Voting Component** serves as the final evaluation level, where human preference data is captured. Users are instructed to judge the responses based on their alignment with human geographic expectations, particularly in terms of the logical groundedness and clarity of the reasoning. The interface provides three mutually exclusive options: “vote for A”, “vote for B”, and “tie”. This preference-driven approach is a deliberate design choice to establish a reliable benchmarking signal in unconstrained scenarios where authoritative labels are absent. By aggregating these pairwise judg-

ments, the platform establishes a leaderboard that reflects real-world utility and the ability of models to communicate spatial logic to human users. The true identities of the models are revealed only after a vote is submitted to prevent brand-related bias and maintain the scientific integrity of the experimental results.

4.2 Data Collection

The data collection protocol of GeoArena is designed to capture the comprehensive reasoning trajectories and preference signals required for a systematic analysis of geographic logic. For every evaluation session, the system records the specific image I , the user prompt P , the anonymized model responses R_A and R_B , and the resulting human preference signal $S \in \{1, 0, 0.5\}$. This multi-modal dataset ensures the traceability of the reasoning process and supports the reproducible computation of model rankings. All collected data are preserved in structured formats that facilitate downstream linguistic analysis and leaderboard updates. To preserve user privacy, we anonymize user inputs and apply filters to remove any potentially sensitive or inappropriate content. Please refer to Appendix A.2 for more details.

4.3 Participating Models

To provide meaningful comparisons of geographic reasoning capabilities, GeoArena includes a diverse selection of both open-source and proprietary models. The selection includes representative Large Vision-Language Models (LVLMs) from multiple providers. For the GPT series (Achiam et al., 2023), it evaluates GPT 4o, GPT 4o mini, GPT 4.1, GPT 4.1 mini, and GPT 4.1 nano. From the Gemini family (Team et al., 2023), it includes Gemini 2.5 pro and Gemini 2.5 flash. The Claude series includes

Claude Opus 4 and Claude Sonnet 4. It also evaluates Llama 4 maverick and Llama 4 scout (Touvron et al., 2023), as well as Gemma 3 models (Team et al., 2025) in sizes of 27B, 12B, and 4B. Additionally, the platform evaluates Qwen 2.5 VL models in sizes of 72B, 32B, and 7B (Bai et al., 2025). As shown in Appendix A.3, GeoArena currently benchmarks 17 models in total. This coverage allows users and researchers to evaluate model performance across different architectures, training paradigms, and geographic reasoning capabilities.

4.4 Ranking Methods

Online Elo Rating. The Elo rating system is a standardized approach to estimate the relative strength of different models based on pairwise comparisons. It provides an interpretable score that reflects the expected probability of one model outperforming another. Formally, given two models M_i and M_j with ratings γ_i and γ_j , the expected probability that model M_i will outperform model M_j is defined as:

$$P(M_i \succ M_j) = \frac{1}{1 + 10^{(\gamma_j - \gamma_i)/\alpha}} \quad (2)$$

where α is a scaling parameter that controls the spread of the probability function, typically set to 400. After observing the actual outcome S_{ij} , where $S_{ij} = 1$ if M_i wins, $S_{ij} = 0.5$ for a tie, and $S_{ij} = 0$ if M_i loses, the rating of model M_i is updated as: $\gamma'_i = \gamma_i + K \cdot (S_{ij} - P(M_i \succ M_j))$, where K is a learning rate that determines the adaptation speed. Although the Elo system is computationally efficient, it is sensitive to the order of matches—an effect that is undesirable for benchmarking static LVLMS. To ensure a stable and order-invariant ranking, we follow prior work (Chiang et al., 2024) and apply the Bradley-Terry model (Bradley and Terry, 1952) to estimate the final scores for the geographic reasoning task.

Bradley-Terry Model. The Bradley-Terry (BT) model provides a principled method to estimate the relative strength of competing models through the maximum likelihood of all observed outcomes. In this framework, each model M_i is assigned a latent strength parameter γ_i . The probability $P(M_i \succ M_j)$ remains consistent with the Elo formulation. The BT model estimates the parameters γ_i by maximizing the likelihood of all recorded pairwise comparisons, accounting for repeated trials through a weighting term W_{ij} . The likelihood

Table 2: GeoArena Leaderboard.

Model	ELO Rating	95% CI lower	95% CI upper
Gemini 2.5 pro	1319.7	974.8	1443.8
Gemini 2.5 flash	1206.5	1062.2	1330.6
Qwen 2.5 VL 72B	1094.5	982.6	1181.9
Gemma 3 12B	1086.5	1002.6	1186.4
Gemma 3 27B	1065.5	959.3	1159.8
GPT 4.1 mini	1059.8	970.0	1161.4
Llama 4 maverick	1046.6	944.6	1115.3
Qwen 2.5 VL 32B	1044.8	964.9	1119.0
GPT 4.1	1044.8	964.9	1119.0
Claude Opus 4	1042.3	933.8	1130.0
Gemma 3 4B	1027.3	936.3	1102.0
Claude Sonnet 4	1019.9	921.3	1113.8
GPT 4o	1000.0	1000.0	1000.0
Llama 4 scout	984.2	876.0	1077.1
Qwen 2.5 VL 7B	950.9	868.4	1056.2
GPT 4.1 nano	917.9	819.1	1015.5
GPT 4o mini	871.6	715.2	1114.7

function is defined as:

$$\mathcal{L}(\Gamma) = \sum_{i,j \in N, i \neq j} W_{ij} \log \left(\frac{1}{1 + 10^{(\gamma_j - \gamma_i)/\alpha}} \right) \quad (3)$$

To compute the final ratings, we apply a linear transformation to align the scores with the standard Elo scale. After fitting the BT model via logistic regression, the estimated parameters $\hat{\gamma}_i$ are transformed as: $\text{rating}_i = \text{scale} \cdot \hat{\gamma}_i + \gamma_{\text{base}}$, where scale is set to 400 and γ_{base} is set to 1000. This transformation preserves the relative ranking while ensuring the scores are consistent with established benchmarking conventions.

Confidence Interval. To ensure that the ranking results are not dependent on a specific sample of comparisons, we estimate confidence intervals (CIs) for the ratings. We adopt a bootstrap procedure (Chiang et al., 2024) that repeatedly resamples the battle outcomes and re-computes the estimates. This approach allows us to quantify the variability in model rankings and provides statistically grounded intervals. The inclusion of confidence intervals is essential to distinguish between meaningful performance differences and those that arise from sampling noise. As a result, the reported rankings offer stronger evidence of the relative strengths of different LVLMS in geographic reasoning.

5 Benchmarks and Results Analysis

5.1 Arena Leaderboard

Table 2 presents the GeoArena leaderboard, representing the geographic reasoning capability of 17 frontier models. The reported 95% confidence intervals are estimated via bootstrap resampling over

100 rounds to account for rating variability across different voting subsets. To maintain the scientific integrity of the experimental results, the input data is processed through the automated quality control layer and manual filtering protocols established in Section 4.1. Several observations can be identified from the results. First, Gemini models demonstrate the highest performance, with Gemini 2.5 pro and Gemini 2.5 flash outperforming all other evaluated systems. This suggests that large-scale, production-level pre-training provides a clear advantage in synthesizing visual cues for geographic reasoning. Second, open-source families such as Qwen 2.5 and Gemma 3 achieve competitive rankings. For instance, Qwen 2.5 VL 72B outperforms Gemma 3 12B and performs comparably to the GPT 4.1 series, indicating that open-source systems are rapidly reducing the gap with proprietary frontier models. Third, several models, including Llama 4 maverick, GPT 4.1, and Claude Opus 4, cluster within the Elo 1040 to 1050 range with significant overlap in confidence intervals, suggesting that their performance differences are not statistically significant. Fourth, smaller variants such as GPT 4.1 nano and GPT 4o mini exhibit a clear decrease in performance. This confirms the difficulty of geographic reasoning, where limited model capacity hinders generalization across diverse global environments. Finally, the wide rating distribution validates that GeoArena is an effective experimental apparatus for distinguishing between frontier systems and lightweight baselines. This differentiation is essential for advancing research on human-aligned geographic reasoning in LVLMs.

5.2 Battle Data Analysis

To provide a comprehensive view of geographic reasoning capabilities, we conduct a pairwise analysis of model interactions, reporting both win rates and battle counts. Figure 3 illustrates the pairwise comparison across models, where the left panel displays head-to-head win rates and the right panel displays the corresponding battle counts. Models are ordered by their average win rate, which makes the hierarchy of geographic reasoning capabilities across the leaderboard interpretable.

Several findings are identified from this analysis. **(1) Frontier systems demonstrate consistent dominance in geographic reasoning.** Gemini 2.5 pro, Gemini 2.5 flash, and GPT 4.1 occupy the top rows, maintaining win rates close to or above 0.7 against nearly all competitors. This persistent

Table 3: Agreement Analysis between Expert and Crowd.

Expert \ Crowd	Left Win	Tie	Right Win	Agreement Rate
Left Win	30	3	3	83.3%
Tie	5	21	6	65.6%
Right Win	2	3	27	84.4%
Agreement Rate	81.1%	77.8%	75.0%	78.0%

advantage suggests that both model capacity and advanced alignment procedures contribute to the robustness of their spatial world knowledge and the logical consistency of their reasoning chains. **(2) Mid-scale models exhibit transitional behavior.** Models such as Gemma 3 12B, Qwen 2.5 VL 72B, and GPT 4.1 mini occupy the middle tier. They achieve favorable outcomes against smaller variants but exhibit substantial performance gaps when challenged by the frontier tier. This suggests a stratification that correlates with the ability to synthesize visual evidence into coherent geographic explanations. **(3) Lower-capacity systems show systematic reasoning deficits.** Models including Gemma 3 4B, Qwen 2.5 VL 7B, GPT 4.1 nano, and GPT 4o mini cluster near the bottom of the heatmap, with win rates typically below 0.3 against larger peers. These deficits are consistent across model families, reflecting limited parameter budgets that restrict the internal representation of global geographic contexts. **(4) Predictable scaling patterns in geographic logic.** Within model families, the quality of geographic reasoning scales predictably with parameter count. For example, the Qwen 2.5 VL series shows clear improvements when moving from 7B to 72B parameters. These trends suggest that scaling determines the consistency and depth of the reasoning process when interpreting environment-specific cues.

5.3 Reliability Analysis of Voting

To validate the reliability of the preference data, we randomly sample 100 instances from the dataset for expert review. In this procedure, an expert is presented with an image, a geographic reasoning prompt, and two anonymized model responses. The expert is instructed to evaluate which response demonstrates superior geographic logic and evidence synthesis. The expert is permitted to use external resources, such as search engines and map services, to verify the factual accuracy of the spatial claims. Each expert evaluation requires approximately 3 to 5 minutes to ensure a rigorous assessment of the reasoning process. Table 3 il-

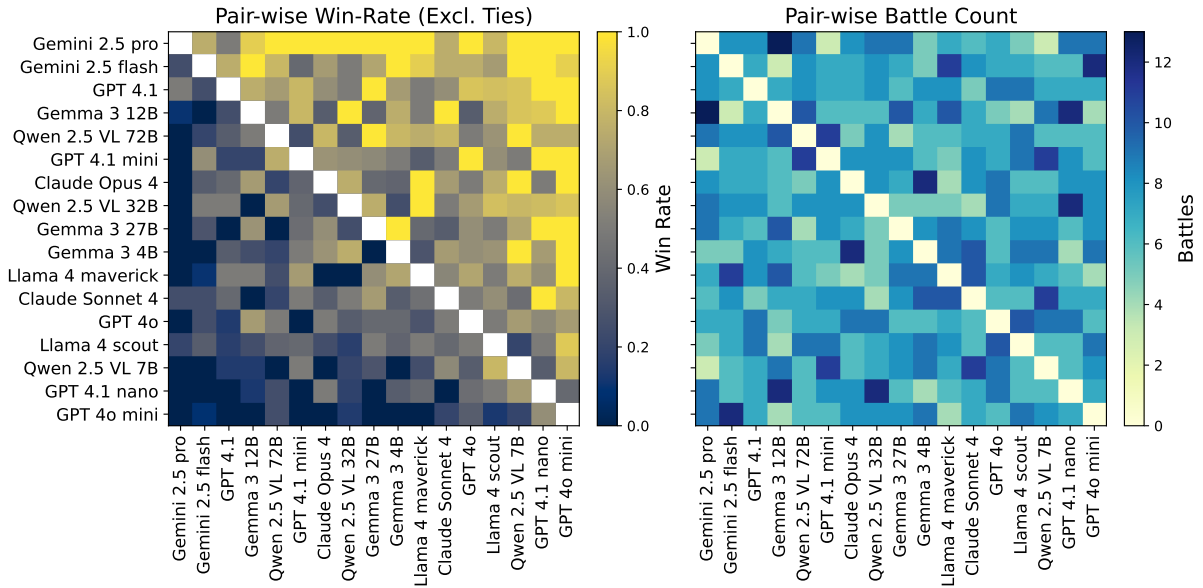


Figure 3: Pair-wise Performance Comparison of Models (Win-Rate and Battle Count).

Table 4: Alignment accuracy of LLMs with human judgments on sampled response pairs.

Model	Gemini 2.5 pro	Qwen 2.5 VL 72B
Accuracy	0.6579	0.4667

illustrates the distribution of preferences between expert and crowd annotations for the sampled instances. We identify a consistently high agreement rate between expert and crowd judgments, ranging from 75% to 85%, with an average agreement of **78%**. According to established studies (Chiang et al., 2024), this level of consensus represents strong agreement. These results demonstrate that the preference signals captured by GeoArena are reliable and accurately reflect geographic reasoning quality.

5.4 Alignment Study

To further examine whether LVLMs can act as reliable evaluators for geographic reasoning, we conduct an alignment study that compares model preferences with human judgments. Specifically, we randomly sample 100 response pairs from the dataset. For each pair, an LVLm is required to determine which response demonstrates superior geographic logic, evidence synthesis, and clarity of expression. The model is restricted to a single output label: win, tie, or loss. The prompt template for this task is provided in Appendix A.4. We then calculate the concordance between the model’s automated judgment and the human-provided prefer-

ence labels.

Table 4 reports the alignment results for Gemini 2.5 pro and Qwen 2.5 VL 72B, representing the top proprietary and open-source models on the GeoArena leaderboard. The results show that Gemini 2.5 pro achieves a higher agreement rate (65.79%) with human evaluations compared to Qwen 2.5 VL 72B (46.67%). This suggests that Gemini 2.5 pro demonstrates stronger alignment with human geographic expectations when assessing reasoning chains.

These findings indicate that while current LVLMs can approximate human preferences to a certain extent, significant gaps remain. This observation supports the requirement for developing more faithful and robust automated evaluators for geographic reasoning and other multimodal tasks.

5.5 Case Study

To illustrate our framework, we present a case study using an image of the Ngātoroirangi Māori Rock Carvings at Mine Bay on Lake Taupō, New Zealand. As shown in Figure 1, different models exhibit varying levels of reasoning depth and factual accuracy. Gemini 2.5 pro produces a comprehensive analysis, identifying salient visual features such as the Māori face carving, surrounding cliffs, and water-based accessibility, while also providing historical and cultural context (e.g., the carving’s creation in 1980 by Matahi Whakataka-Brightwell). In contrast, GPT 4o mini generates only a brief description, lacking explicit reasoning and omit-

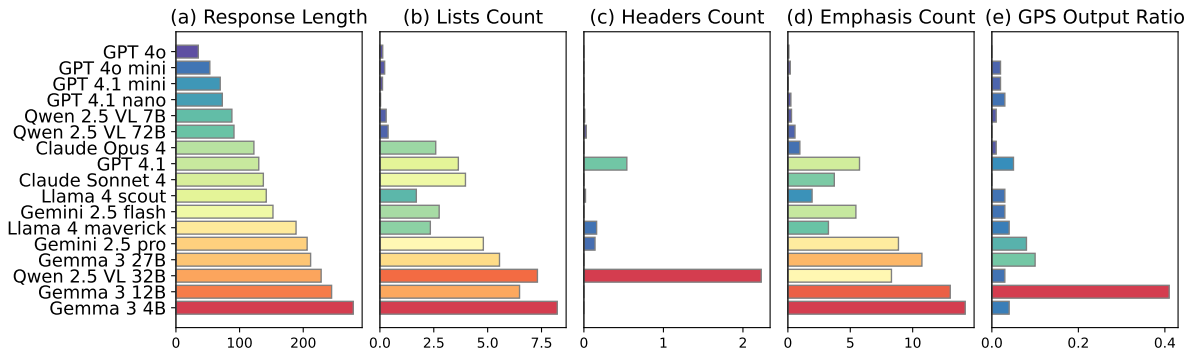


Figure 4: Distribution of Style Features in Model Outputs.

ting cultural details. This comparison underscores the importance of reasoning quality and contextual grounding in geolocation tasks, showing that structured analyses align more closely with human preferences and task requirements. We also give hard cases analysis and more case studies in Appendix A.5 and Appendix A.6.

5.6 Preference Analysis

To identify which linguistic characteristics of geographic reasoning chains drive human preference, we extend the standard Bradley-Terry regression framework by incorporating style-related features as confounding variables following previous work (Chiang et al., 2024; Tianle Li, 2024; Dubois et al., 2024). By including these features in the regression, we can separate the effect of linguistic style from the intrinsic reasoning ability of the model. The style coefficients (β) are estimated via logistic regression, where normalized style features are included alongside model indicators in an extended design matrix. The resulting coefficients quantify the degree to which specific stylistic traits influence human judgments.

In this study, we consider five different features: response length (measured by word count), list count (reflecting structured reasoning steps), header count, emphasis count (including bold and italic items), and GPS output ratio (the proportion of responses providing precise spatial claims). Figure 4 illustrates the distribution of these features across different models, showing variation in the linguistic expression of geographic logic. Our analysis identifies findings consistent with prior studies in other domains (Chiang et al., 2024; Steyvers et al., 2024; Tianle Li, 2024). Specifically, response length exhibits a strong positive correlation with human preference ($\beta_{\text{length}} = 0.526$), suggesting that longer explanations are perceived as provid-

ing more exhaustive geographic evidence. In addition, both list count ($\beta_{\text{list}} = 0.095$) and GPS output ratio ($\beta_{\text{GPS}} = 0.06$) are positively correlated with preference. A higher number of lists often reflects more explicit reasoning steps, while the presence of GPS coordinates provides a concrete anchor for the spatial reasoning chain. However, header count ($\beta_{\text{header}} = -0.153$) and emphasis count ($\beta_{\text{emphasis}} = -0.117$) do not show positive associations with human preference. It is possible that excessive structural markers or textual emphasis are perceived as superficial formatting rather than substantive geographic evidence, and thus do not contribute to the perceived informativeness of the reasoning process. The style-adjusted leaderboard is provided in Appendix A.7.

6 Conclusion

In this work, we introduce **GeoArena**, a formal evaluation regime for benchmarking open-world geographic reasoning in LVLMs. By leveraging in-the-wild user contributions and pairwise human judgments, GeoArena addresses the critical failures of existing static benchmarks, including data contamination, reasoning process neglect, and ground-truth scarcity. Our framework decouples the assessment of logical reasoning from the hard requirement of exact coordinate labels, allowing for a more nuanced understanding of how models synthesize visual evidence with spatial knowledge. Through the implementation of a robust Bradley-Terry model, we establish a reliable leaderboard that reflects human geographic expectations. GeoArena is open-sourced to foster future research. It is anticipated that GeoArena will support the related field in developing geographic reasoning systems that are both logically consistent and aligned with real-world utility.

Limitations

While GeoArena establishes a formal regime for evaluating geographic reasoning, several limitations remain. First, although the platform is open to global participation, the geographic distribution of user-submitted images and votes may exhibit inherent bias based on the current user base. It is anticipated that as the platform attracts a broader demographic over time, this bias will naturally decrease. Moreover, to prioritize user privacy and encourage participation, the platform does not currently track unique user identifiers. This makes it difficult to quantify the impact of individual user bias on the overall leaderboard. Exploring privacy-preserving tracking methods, such as hashed identifiers, represents a potential solution for future development. Despite these challenges, GeoArena provides a unique and essential signal for human-aligned geographic reasoning that traditional static metrics cannot capture.

Ethical considerations

The development and deployment of GeoArena as a formal evaluation regime for geographic reasoning adhere to rigorous ethical standards. We prioritize user privacy and data protection throughout the data acquisition pipeline. The platform does not collect or store personally identifiable information, and users are not required to submit metadata or coordinates tied to private locations. All contributed images and preference votes undergo an anonymization process and are managed in compliance with ethical data management practices. Our human evaluation methodology is restricted to pairwise preference voting and does not involve the collection of sensitive demographic or personal data. No financial compensation or targeted recruitment was involved in the data collection process. These considerations ensure that GeoArena functions as a scientific instrument for the geographic AI community. We use Large Language Models to polish the writing in this work.

Acknowledgments

Pengyue Jia, Yingyi Zhang, and Xiangyu Zhao are supported by Hong Kong Research Grants Council (Research Impact Fund No.R1015-23, Collaborative Research Fund No.C1043-24GF, General Research Fund No. 11218325), and the Institute of Digital Medicine of City University of Hong Kong (No.9229503).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, and 1 others. 2024. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21967–21977.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Opeyemi Bamigbade, John Sheppard, and Mark Scanlon. 2024. Computer vision for multimedia geolocation in human trafficking investigation: A systematic literature review. *arXiv preprint arXiv:2402.15448*.
- Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Jiawei Cheng, Jingyuan Wang, Yichuan Zhang, Jiahao Ji, Yuanshao Zhu, Zhibo Zhang, and Xiangyu Zhao. 2025. Poi-enhancer: An llm-based semantic enhancement framework for poi representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 11509–11517.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.
- Muhammad Danish, Muhammad Akhtar Munir, Syed Roshan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. 2025. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7132–7142.
- Mahir Labib Dihan, Md Tanvir Hassan, Md Tanvir Parvez, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models. *arXiv preprint arXiv:2501.00316*.
- Zhiyang Dou, Zipeng Wang, Xumeng Han, Guorong Li, Zhipei Huang, and Zhenjun Han. 2024. Gaga: Towards interactive global geolocation assistant. *arXiv preprint arXiv:2412.08907*.

- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.
- Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. 2025. Around the world in 80 timesteps: A generative approach to global visual geolocation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23016–23026.
- 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.
- Xiao Han, Chen Zhu, Hengshu Zhu, and Xiangyu Zhao. 2025. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 814–825.
- James Hays and Alexei A Efros. 2008. Im2gps: estimating geographic information from a single image. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.
- Jingyuan Huang, Jen-tse Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. 2025a. VLms as geoguessr masters: Exceptional performance, hidden biases, and privacy risks. *arXiv preprint arXiv:2502.11163*.
- Yu Chin Huang, Yuhan Ji, and Song Gao. 2025b. Evaluating geospatial reasoning capabilities in large language models: A benchmark on geometry classification, topological relations and direction estimation. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications*, pages 64–71.
- Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. 2020. Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond.
- Neel Jay, Hieu Minh Nguyen, Trung Dung Hoang, and Jacob Haimen. 2025. Evaluating precise geolocation inference capabilities of vision language models. *arXiv preprint arXiv:2502.14412*.
- Pengyue Jia, Yiding Liu, Xiaopeng Li, Xiangyu Zhao, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, and Dawei Yin. 2024a. G3: an effective and adaptive framework for worldwide geolocation using large multi-modality models. *Advances in Neural Information Processing Systems*, 37:53198–53221.
- Pengyue Jia, Yiding Liu, Xiangyu Zhao, Xiaopeng Li, Changying Hao, Shuaiqiang Wang, and Dawei Yin. 2024b. Mill: Mutual verification with large language models for zero-shot query expansion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2498–2518.
- Pengyue Jia, Seongheon Park, Song Gao, Xiangyu Zhao, and Yixuan Li. 2025a. Georanker: Distance-aware ranking for worldwide image geolocalization. *arXiv preprint arXiv:2505.13731*.
- Pengyue Jia, Derong Xu, Xiaopeng Li, Zhaocheng Du, Xiangyang Li, Yichao Wang, Yuhao Wang, Qidong Liu, Maolin Wang, Huifeng Guo, and 1 others. 2025b. Bridging relevance and reasoning: Rationale distillation in retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4242–4256.
- Pengyue Jia, Derong Xu, Yingyi Zhang, Xiaopeng Li, Wenlin Zhang, Yi Wen, Yuanshao Zhu, and Xiangyu Zhao. 2026. Georouter: Dynamic paradigm routing for worldwide image geolocalization. *arXiv preprint arXiv:2603.24376*.
- Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhui Chen. 2024. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908.
- Xiangjie Kong, Juntao Wang, Zehao Hu, Yuwei He, Xiangyu Zhao, and Guojiang Shen. 2024. Mobile trajectory anomaly detection: Taxonomy, methodology, challenges, and directions. *IEEE Internet of Things Journal*, 11(11):19210–19231.
- Hao Li, Fabian Deuser, Wenping Yin, Steffen Knoblauch, Wufan Zhao, Filip Biljecki, Yong Xue, and Wei Huang. 2025a. Towards generative location awareness for disaster response: A probabilistic cross-view geolocalization approach. *arXiv preprint arXiv:2512.20056*.
- Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Forty-first International Conference on Machine Learning*.
- Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. 2025b. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. *arXiv preprint arXiv:2506.14674*.
- Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. 2025c. From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models. *arXiv preprint arXiv:2508.01608*.
- Yi Liu, Junchen Ding, Gelei Deng, Yuekang Li, Tianwei Zhang, Weisong Sun, Yaowen Zheng, Jingquan Ge, and Yang Liu. 2024. Image-based geolocation using large vision-language models. *arXiv preprint arXiv:2408.09474*.

- Gengchen Mai, Krzysztof Janowicz, Yingjie Hu, Song Gao, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. 2022. A review of location encoding for geoai: methods and applications. *International Journal of Geographical Information Science*, 36(4):639–673.
- Anindya Sarkar, Srikumar Sastry, Aleksis Pirinen, Chongjie Zhang, Nathan Jacobs, and Yevgeniy Vorobeychik. 2024. Goma-geo: Goal modality agnostic active geo-localization. *Advances in Neural Information Processing Systems*, 37:104934–104964.
- Mark Steyvers, Heliodoro Tejada, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas Mayer, and Padhraic Smyth. 2024. The calibration gap between model and human confidence in large language models. *arXiv preprint arXiv:2401.13835*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.
- Wei-Lin Chiang Tianle Li, Anastasios Angelopoulos. 2024. [Does style matter? disentangling style and substance in chatbot arena](#).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Donald Thomas Venable. 2016. *Improving Real World Performance for Vision Navigation in a Flight Environment*. Ph.D. thesis, Air Force Institute of Technology.
- Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701.
- Nam Vo, Nathan Jacobs, and James Hays. 2017. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 2621–2630.
- Senzhang Wang, Jiannong Cao, and S Yu Philip. 2020. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700.
- Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482.
- Zhiqiang Wang, Dejia Xu, Rana Muhammad Shahroz Khan, Yanbin Lin, Zhiwen Fan, and Xingquan Zhu. 2024. Llmgeo: Benchmarking large language models on image geolocation in-the-wild. *arXiv preprint arXiv:2405.20363*.
- Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. 2021. Visual and object geolocation: A comprehensive survey. *arXiv preprint arXiv:2112.15202*.
- Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, Tong Xu, and 1 others. 2025. From single to multi-granularity: Toward long-term memory association and selection of conversational agents. *arXiv preprint arXiv:2505.19549*.
- Tong Xu, Hengshu Zhu, Xiangyu Zhao, Qi Liu, Hao Zhong, Enhong Chen, and Hui Xiong. 2016. Taxi driving behavior analysis in latent vehicle-to-vehicle networks: A social influence perspective. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294.
- Sahiti Yerramilli, Nilay Pande, Rynaa Grover, and Jayant Sravan Tamarapalli. 2025. Geochain: Multimodal chain-of-thought for geographic reasoning. *arXiv preprint arXiv:2506.00785*.
- Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. 2025. Notellm-2: Multimodal large representation models for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2815–2826.
- Yingyi Zhang, Junyi Li, Wenlin Zhang, Penyue Jia, Xianneng Li, Yichao Wang, Derong Xu, Yi Wen, Huifeng Guo, Yong Liu, and 1 others. 2026. Evoking user memory: Personalizing llm via recollection-familiarity adaptive retrieval. *arXiv preprint arXiv:2603.09250*.
- Zijian Zhang, Ze Huang, Zhiwei Hu, Xiangyu Zhao, Wanyu Wang, Zitao Liu, Junbo Zhang, S Joe Qin, and Hongwei Zhao. 2023a. Mlpst: Mlp is all you need for spatio-temporal prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3381–3390.
- Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi

- Wang, and Zitao Liu. 2023b. Promptst: Prompt-enhanced spatio-temporal multi-attribute prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3195–3205.
- Zijian Zhang, Xiangyu Zhao, Hao Miao, Chunxu Zhang, Hongwei Zhao, and Junbo Zhang. 2023c. Autostl: Automated spatio-temporal multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 4902–4910.
- Xiangyu Zhao, Wenqi Fan, Hui Liu, and Jiliang Tang. 2022. Multi-type urban crime prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4388–4396.
- Xiangyu Zhao and Jiliang Tang. 2017. Exploring transfer learning for crime prediction. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 1158–1159. IEEE.
- Xiangyu Zhao, Tong Xu, Yanjie Fu, Enhong Chen, and Hao Guo. 2017. Incorporating spatio-temporal smoothness for air quality inference. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1177–1182. IEEE.
- Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. 2024. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval*, pages 2749–2754.
- Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Yu. 2023a. Synmob: Creating high-fidelity synthetic gps trajectory dataset for urban mobility analysis. *Advances in Neural Information Processing Systems*, 36:22961–22977.
- Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Yu. 2023b. Difftraj: Generating gps trajectory with diffusion probabilistic model. *Advances in Neural Information Processing Systems*, 36:65168–65188.
- Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. 2024a. Controltraj: Controllable trajectory generation with topology-constrained diffusion model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4676–4687.
- Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Xuetao Wei, and Yuxuan Liang. 2024b. Unitraj: Universal human trajectory modeling from billion-scale worldwide traces. *arXiv preprint arXiv:2411.03859*, 2.

A Appendix

A.1 Verifying the Feasibility of Automatic Prompt Filtering

To maintain a reliable leaderboard, it is essential to ensure that user inputs are relevant to the image’s geographic reasoning. To examine whether LLMs can replace manual filtering, we conduct an experiment to determine whether LLMs can identify when a user prompt requests geolocating an image.

For this study, we construct a binary classification task. We randomly select 100 prompts from our voting data and assign them the label True, indicating that they ask about image geolocation. In parallel, we sample 100 prompts from the Chatbot Arena dataset², which contains general-purpose prompts, and labeled them as False. Each model is given a simple instruction that defines image geolocation, specifies the expected JSON output, and directs the model to respond only with a True or False label. The instruction is given as follows:

You are a prompt classifier. Analyze the provided user prompt and determine if it is asking about image geolocalization. Image geolocalization refers to determining or estimating the geographic location (e.g., city, country, landmark) where an image was taken based on its visual content. Return ONLY a JSON object with one key: “is_geo”. The value must be “true” if the prompt is inquiring about geolocating an image (e.g., “Where was this photo taken?” or instructions for an expert in image geolocalization), or “false” otherwise. If uncertain, default to “false”. Output format (no extra words): “is_geo”: “true”|“false”
User prompt: user_prompt

We evaluate three models: Gemini 2.0 flash, GPT 3.5 turbo, and GPT 4.1 mini. All three models achieve 100% accuracy on this task. The high accuracy is mainly due to two factors. First, most users ask questions through the default prompt provided by GeoArena, which reflects a stable phrasing pattern. Second, prompts that request geographic reasoning usually contain explicit references to places,

²https://huggingface.co/datasets/lmsys/chatbot_arena_conversations

images, or location inference, which makes them easy for the models to detect. These observations show that modern language models can serve as reliable automatic filters for user inputs. Such a mechanism would allow the leaderboard to remain focused on geographic reasoning queries while reducing the need for manual inspection.

A.2 Dataset Characteristics and Composition

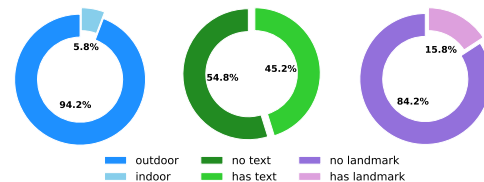


Figure 5: Composition of Image Features in Dataset

To further explore the characteristics of the dataset, we employ GPT 4o to annotate the collected images, focusing on three key aspects:

1. Scene Type: whether the image depicts an indoor or outdoor setting.
2. Text Presence: whether the image contains prominent, recognizable text.
3. Landmark Presence: whether the image features a landmark, such as a historical site or natural icon.

The corresponding results are presented in Figure 5. The figure comprises three doughnut charts, each illustrating the distribution of one of the annotated attributes across the dataset:

1. The first doughnut chart indicates that 94.2% of images are classified as outdoor scenes, with only 5.8% representing indoor environments. This pronounced skew toward outdoor imagery aligns with the global scope of GeoArena, where user-submitted images are likely dominated by exterior scenes captured in diverse geographic contexts.
2. Text Presence: The second doughnut chart reveals a more balanced distribution, with 54.8% of images lacking recognizable text ("no text") and 45.2% containing text ("has text"). This near-equitable split underscores the dataset’s richness, incorporating both text-free natural scenes and images with textual elements such

as signs or labels. This variability is particularly valuable for assessing LVLM capabilities in multi-modal reasoning, where text recognition can enhance location prediction accuracy.

3. **Landmark Presence:** The third doughnut chart shows that 84.2% of images do not contain landmarks ("no landmark"), while 15.8% do ("has landmark"). The low prevalence of landmarks reflects the dataset's emphasis on general geographic scenes rather than iconic or tourist-heavy locations, offering a broad representation of natural and urban environments worldwide. This distribution highlights the dataset's potential to test LVLM generalization across less distinctive locales, a challenging yet realistic scenario for global geolocation. Overall, these distributions reveal the dataset's heterogeneity, making it a robust resource for benchmarking LVLM performance under real-world conditions.

A.3 Models in GeoArena

GeoArena benchmarks 17 models in total, please refer to Table 5 for details.

A.4 LVLM Alignment Evaluation Prompt

The prompt template used for LVLM alignment evaluation is as follows:

You are an expert evaluator in image geolocation tasks. I will give you two model responses to the same geolocation prompt.

Here is the prompt:

- Prompt: {sample['prompt']}

- Image: {sample['image']}

Response A: {sample['response A']}

Response B: {sample['response B']}

Your task is to decide which response is better based on:

1. Accuracy of the predicted location
2. Strength of reasoning and evidence
3. Clarity and specificity

Output only one word:

- "win" if Response A is better
- "loss" if Response B is better
- "tie" if both are equally good

We first prompt the LVLM to act as an expert evaluator for the geographic reasoning task. For each sample, the model is provided with the prompt and the associated images, along with the responses



Figure 6: Images Where Strong Models Excel but Weaker Models Fail

from two candidate models. It is then asked to determine which response is better, considering three dimensions: accuracy, reasoning, and clarity and specificity.

A.5 Difficult Queries Analysis

To better understand the distinction between top-tier and lower-tier models, we examined a set of representative examples where strong models predicted correctly, but weaker models failed, as illustrated in the Figure 6. Specifically, we selected the top-5 models on the leaderboard as well as the bottom-5 models. We then extracted examples where the top-5 models outperformed the bottom-5 models. These images generally do not contain prominent or unique landmarks that can be easily recognized. Instead, they often depict environments that require reasoning over subtle contextual cues—such as vegetation types, architectural style, road structures, or geographic textures—to infer the location. This suggests that strong models may better leverage fine-grained reasoning and background knowledge beyond surface-level features. The presence of such challenging samples highlights a promising direction for future model development: expanding training and evaluation on these harder cases. By doing so, upcoming systems may gain robustness and achieve more consistent performance across diverse scenarios, particularly where visual clues are subtle and non-obvious.


A.6 Additional cases from GeoArena

In this section, we present additional case studies from GeoArena to illustrate the diversity and complexity of the geolocation tasks it encompasses. Figure 7 showcases an image of the iconic Percé Rock in Quebec, Canada, highlighting the model's ability to recognize unique geological formations. Figure 8 features the Olympic Park in Beijing, China, demonstrating the model's profi-

Input Prompt

You are an expert in image geolocation. Given an image, provide the most likely location it was taken.

Input Image



Gemini 2.5 Flash

Based on the distinctive rock formation with an arch and a separate sea stack, the most likely location is **Percé Rock (Rocher Percé), located in Percé, Quebec, Canada.**

Llama 4 Macerick

To determine the location of the image, we need to analyze its features. The image shows a large rock formation with an arch in it, surrounded by water and another rock formation to its right.

- Identify Distinctive Features:** The key feature here is the rock formation with an arch. Such formations are often associated with coastal areas where erosion has shaped the landscape.
- Consider Geographical Context:** The presence of a large body of water and the specific type of rock formation suggests a coastal location. The clarity of the water and the rock's appearance could indicate a relatively pristine or less polluted area.
- Research Similar Locations:** There are several locations around the world known for their natural arches and rock formations. However, the specific combination of features in this image, including the size of the arch and the surrounding landscape, needs to be matched against known sites.
- Matching the Image to Known Locations:** One of the most famous natural arches is in Cabo San Lucas, Mexico, known as "El Arco" or "The Arch." It is a large granite rock formation with an arch that has been eroded by the sea. The image closely resembles descriptions and pictures of El Arco.
- Conclusion:** Given the distinctive shape of the rock formation and its arch, along with the surrounding environment, the most likely location for the image is Cabo San Lucas, Mexico, specifically "El Arco."


The image is likely taken at El Arco in Cabo San Lucas, Mexico.

Figure 7: Additional Case Study: Identifying the Percé Rock.

Input Prompt

You are an expert in image geolocation. Given an image, provide the most likely location it was taken.

Input Image



Claude Sonnet 4

Based on the distinctive architectural and design elements visible in this image, this appears to be taken at **Olympic Forest Park (Olympic Green) in Beijing, China.** The key identifying features include:

- Distinctive lamp posts:** The unique sculptural street lamps with leaf-like designs are characteristic of the Olympic Green area.
- Large open plaza:** The expansive paved public space is typical of the Olympic Forest Park's grand pedestrian areas.
- Modern urban planning:** The organized layout and design aesthetic match Beijing's Olympic district development.
- Architectural style:** The overall design language reflects the modernization efforts made for the 2008 Beijing Olympics.

This area was specifically developed as part of Beijing's Olympic infrastructure and remains a popular public space and tourist destination. The Olympic Forest Park is located in the Chaoyang District of Beijing and serves as both a recreational area and a landmark representing modern Chinese urban design.

Qwen2.5 VL 72b


This image appears to have been taken in **Harbin, China.** The long open plaza with distinctive streetlights and the overall layout is characteristic of the **Harbin Sun Island Plaza**, which is a popular tourist area in the city. The clear sky and the bare trees suggest it could be during the winter months, which is known for Harbin's famous **ice and snow festivals.**

Figure 8: Additional Case Study: Identifying the Olympic Park, Beijing.

Input Prompt

You are an expert in image geolocation. Given an image, provide the most likely location it was taken.

Input Image



Llama 4 Macerick

Based on the image provided, the landscape depicted appears to be a tropical or subtropical region, characterized by a golf course, dense vegetation, and a coastline with clear turquoise water.

- Climate and Vegetation:** The presence of dense lush greenery and palm trees indicates a tropical or subtropical climate. Regions known for such landscapes include parts of the Caribbean, Southeast Asia, and the Pacific Islands.
- Golf Course and Infrastructure:** The well-maintained golf course and the presence of what appears to be resort infrastructure (e.g., buildings along the coast) suggest a popular tourist destination. Many tropical and subtropical regions have developed golf courses and resorts to attract tourism.
- Coastline and Water Characteristics:** The clear turquoise water and the coral reef visible in the shallow waters are characteristic of tropical marine environments. This feature is common in areas like the Caribbean and the Pacific Islands.
- Specific Location Clues:** Without specific landmarks or text, we rely on the overall aesthetic and features. However, the layout of the golf course, the architecture, and the specific shade of the water can sometimes give clues. For instance, the presence of a fringing reef is typical of many Pacific Island locations.

Given these observations, a plausible location for this image could be in the Pacific Islands or the Caribbean. However, without more specific details or recognizable landmarks, pinpointing the exact location is challenging. That said, one of the locations that closely match the description and the visual characteristics is **Fiji**. Fiji is known for its clear waters, coral reefs, lush vegetation, and numerous golf courses and resorts, making it a popular destination for both golfing and beach tourism.

In conclusion, while the exact location cannot be determined with absolute certainty without additional information, Fiji is a strong candidate based on the visual cues provided in the image.

Gemma 3 12b

Okay, let's analyze this image and pinpoint its likely location. Here's my assessment based on visual cues and my expertise in image geolocation:

Most Likely Location: Kempinski Grand Baie, Mauritius

Reasoning: 1. **Vegetation and Landscape:** The lush, tropical vegetation (palm trees, dense foliage) immediately points to a tropical location. 2. **Golf Course:** The presence of a well-manicured golf course is a significant clue. This is not just anyplace; it's a high-end resort area. 3. **Turquoise Water and Barrier Reef:** The incredibly clear, turquoise water and the visible barrier reef are crucial. The color and clarity of the water suggest a specific type of coral reef environment. 4. **Coastal Architecture:** The buildings along the coast feature a distinctive architectural style often seen in luxury resorts. 5. **Match with known images of the Kempinski Grand Baie:** The layout of the golf course and coastline, re-vegetation, and building styles strongly match aerial views and images of the Kempinski Grand Baie resort in Mauritius. The resort has a unique course with distinctive features.

Further refinements: Based on the elevation and angle of the photograph, the photo was likely taken from a drone.

Confidence Level: 95%

Disclaimer: While I am highly confident, certain features can change slightly over time, and other locations could theoretically exhibit similar characteristics. However, the combination of factors makes Kempinski Grand Baie the most probable location.

If you have other images you'd like me to analyze, please send them over!

Figure 9: Additional Case Study: Identifying the Golf Course in Fiji.

Table 5: Large-scale models benchmarked in *GeoArena*. Prices are USD / million tokens (input/output) and USD / thousand (image).

Model	Company	Params	Openness	API Price (input / output / image)
GPT 4o	OpenAI	Unknown	Proprietary	\$2.50 / \$10.00 / \$3.61
GPT 4o mini	OpenAI	Unknown	Proprietary	\$0.15 / \$0.60 / \$0.22
GPT 4.1	OpenAI	Unknown	Proprietary	\$2.00 / \$8.00 / -
GPT 4.1 mini	OpenAI	Unknown	Proprietary	\$0.40 / \$1.60 / -
GPT 4.1 nano	OpenAI	Unknown	Proprietary	\$0.10 / \$0.40 / -
Gemini 2.5 flash	Google DeepMind	Unknown	Proprietary	\$0.15 / \$0.60 / \$0.62
Gemini 2.5 pro	Google DeepMind	Unknown	Proprietary	\$1.25 / \$10.00 / \$5.16
Claude Sonnet 4	Anthropic	Unknown	Proprietary	\$3.00 / \$15.00 / \$4.80
Claude Opus 4	Anthropic	Unknown	Proprietary	\$15.00 / \$75.00 / \$24.00
Llama 4 maverick	Meta	17B/402B	Open-source	\$0.15 / \$0.60 / \$0.67
Llama 4 scout	Meta	17B/109B	Open-source	\$0.08 / \$0.30 / -
Gemma 3 27B	Google	27B	Open-source	\$0.10 / \$0.20 / \$0.03
Gemma 3 12B	Google	12B	Open-source	\$0.15 / \$0.10 / -
Gemma 3 4B	Google	4B	Open-source	\$0.02 / \$0.04 / -
Qwen 2.5 VL 72B	Alibaba	72B	Open-source	\$0.25 / \$0.75 / -
Qwen 2.5 VL 32B	Alibaba	32B	Open-source	\$0.90 / \$0.90 / -
Qwen 2.5 VL 7B	Alibaba	7B	Open-source	\$0.20 / \$0.20 / -

ciency in identifying modern architectural landmarks. Lastly, Figure 9 depicts a golf course in Fiji, emphasizing the model’s capability to infer locations based on environmental and recreational context. These examples underscore *GeoArena*’s effectiveness in challenging models to perform accurate geographic reasoning across a wide range of scenarios.

A.7 Style-Adjusted Elo Ratings

To disentangle response style from geographic reasoning quality, we compute style-adjusted Elo ratings that control for stylistic features. Table 6 presents the adjusted leaderboard.

Notably, while top-tier models (Gemini family) remain dominant after adjustment, the Gemma-3 family exhibits significant ranking drops (e.g., Gemma 3 12B falls from 4th to 9th), suggesting their original performance was partially inflated by verbose responses.

A.8 User Consent

To ensure responsible data usage and protect user privacy, *GeoArena* requires all participants to provide consent before submitting any images or preference votes. When users interact with the platform, they are presented with a clear consent statement indicating that uploaded images and voting records may be used for research purposes and may be released in anonymized form. Users are also informed that participation is voluntary and that

Table 6: Style-adjusted Elo ratings controlling for style features.

Rank	Model	Adjusted Elo
1	Gemini 2.5 pro	1171.21
2	Gemini 2.5 flash	1093.10
3	GPT 4.1	1066.74
4	Qwen2.5 VL 72B	1045.31
5	Qwen2.5 VL 32B	1031.88
6	GPT 4.1 mini	1014.19
7	GPT 4o	1000.00
8	Claude Opus 4	952.71
9	Gemma 3 12B	932.11
10	Claude Sonnet 4	919.71
11	Llama 4 maverick	914.94
12	Gemma 3 27B	910.33
13	Llama 4 scout	878.12
14	Gemma 3 4B	876.96
15	Qwen2.5 VL 7B	868.93
16	GPT 4.1 nano	853.09
17	GPT 4o mini	780.37

they should avoid uploading sensitive or personally identifiable content. These measures confirm that the data included in *GeoArena* is collected with explicit user permission and used strictly within an academic context.