TOWARDS DIVERSE GLOBAL IMAGE GEOLOCALIZATION: BASELINES, DATASET, AND BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Diverse Global Image Geolocalization, an extended task that generalizes image geolocalization beyond the street-view domain to ten distinct domains. Inputs mainly originate from continuous domains, such as street-view, remote sensing, terrain and plant, and maps, or from discrete domains, such as landmark, road-network, indoor, shapes, space, and UAV imagery. We provide a complete suite for this task, comprising GeoVerse100K, a training dataset, a ten-domain evaluation benchmark, and a baseline. GeoVerse100K is constructed with a geographic-attribute sampling method to ensure balanced geographic coverage and a reward-based sampling method to better support RL-based training. We further divide the benchmark into 37 subcategories and conduct a comprehensive evaluation of both general and specialized models. To establish a strong baseline, we propose two training strategies tailored to the multi-domain setting: domain-balanced sampling and a difficulty-guided curriculum. We then finetune Qwen3.5-9B using GRPO with a distance-aware reward. Experiments show that current specialized models suffer from severe generalization limitations, whereas our baseline trained on GeoVerse100K achieves balanced performance across all ten domains. The code and dataset will be made publicly available.

1 INTRODUCTION

Image geolocalization (Hays & Efros, 2008) aims to infer the capture location of an image from its visual content, usually in the form of coordinates or addresses. As an important task that connects visual perception with geographic understanding, it has broad value in disaster assessment, intelligent navigation, and urban computing. At the same time, this task also has a competitive and entertaining aspect, which fosters player communities such as GeoGuessr (geo, 2025) and Tuxun (tux, 2025) and further increases public interest. For these reasons, improving the accuracy, robustness, and generalization ability of image geolocalization is of clear practical and scientific importance.

Early studies on image geolocalization (Flatow et al., 2015; Hays & Efros, 2008; Workman et al., 2015) often focus on place recognition (Torii et al., 2013) and image matching within limited geographic regions (Torii et al., 2015; Arandjelovic et al., 2016; Weyand et al., 2016). As research progresses, some works begin to extend the task from local place recognition to global-scale geolocalization (Vo et al., 2017; Seo et al., 2018). More recently, with the rapid progress of multimodal large language models, image geolocalization shifts from traditional classification (Radenović et al., 2019; Izbicki et al., 2020; Kordopatis-Zilos et al., 2021) and retrieval toward (Workman et al., 2015; Haas et al., 2024; Jia et al., 2025b) generative modeling based on visual reasoning (Li et al., 2024; Dou et al., 2026; Liu et al., 2025). Accordingly, model outputs evolve from a single latitude-longitude prediction to richer forms that also include reasoning processes and natural-language location descriptions (Han et al., 2025; Song et al., 2025; Campos et al., 2026; Cheng et al., 2026). In parallel, dataset annotations become increasingly fine-grained, beginning to incorporate natural-language geographic descriptions (Li et al., 2025a; Zhang et al., 2025b; Ma et al., 2025), human reasoning traces (Yi & Shan, 2025; Wang et al., 2025a; Jin et al., 2026; Jia et al., 2026b), and tool-use trajectories (Wang et al., 2025d; Jia et al., 2026a).

However, existing datasets still suffer from two key limitations. First, many datasets focus on a single visual domain, such as street-view. This limits the model's ability to learn geographic cues that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

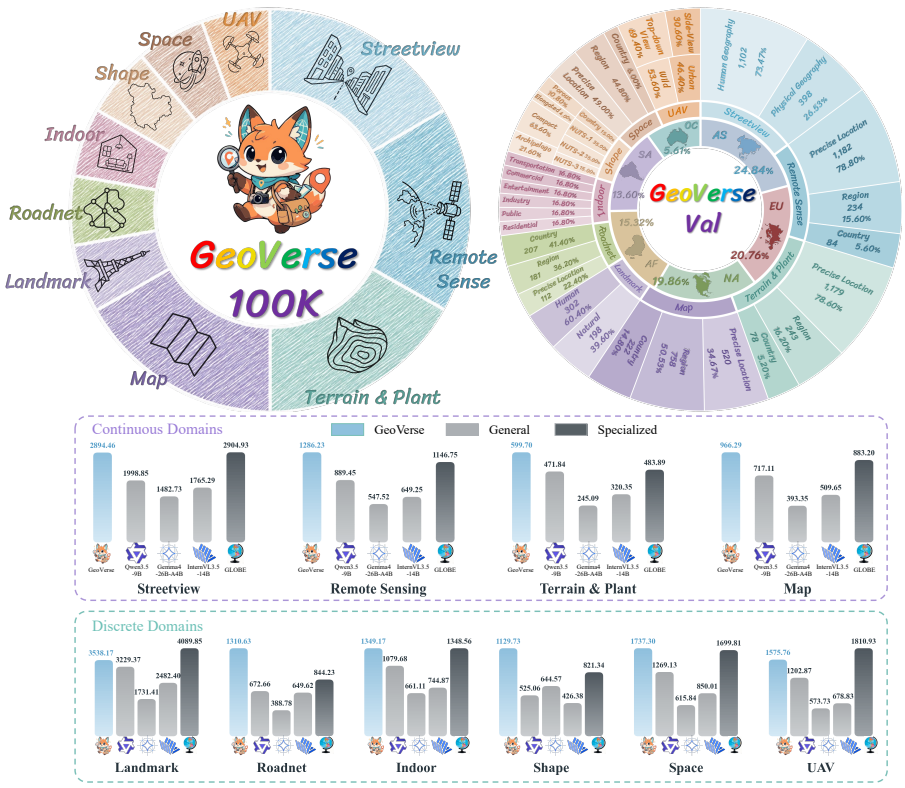


Figure 1: **Composition of GeoVerse100K and comparison of GeoVerse with other models.**

transfer across domains and also makes overfitting to a specific domain more likely. Although researchers introduce datasets for UAV views (Zheng et al., 2020; Ji et al., 2025a; Chen et al., 2025b; Ji et al., 2025b; Liu et al., 2026a), aerial imagery (Workman et al., 2015; Srivastava et al., 2025; Berton et al., 2024), and remote sensing (Zhang et al., 2025a; Li et al., 2025c; Liu et al., 2026b), which broaden the task to more viewpoints and cross-view settings (Ye et al., 2025; Zheng et al., 2026b; Xu et al., 2025a; Kristoffersen et al., 2026), these datasets are often small and cover only limited geographic areas. Second, mainstream street-view geolocalization datasets are largely built from user-uploaded content and therefore show strong spatial bias toward high-resource regions, such as Europe and North America. As a result, they often suffer from severe regional class imbalance. Together, these two problems restrict the generalization, robustness, and practical reliability of geolocalization models in open-world settings.

Meanwhile, a growing number of studies use reinforcement learning to train large models to provide reasoning for their location predictions, particularly through frameworks based on relative reward optimization algorithms such as GRPO (Guo et al., 2025). However, for overly easy samples, the performance differences among candidate outputs are often small, leaving only limited effective signals for driving policy improvement. For overly difficult samples, model outputs are often poor as a whole, making it difficult to form stable relative rankings. Therefore, geolocalization datasets should also account for the distribution of sample difficulty, reward separability, and how learnable the samples are for models at different capability levels.

To bridge this gap, in this paper, we propose an extended task named Diverse Global Image Geolocalization, which extends the original setting to a broader range of domains. We further construct a new dataset and benchmark, GeoVerse100K, specifically designed for the reinforcement-learning paradigm and covering ten different domains for training and comprehensive evaluation on this task. We also establish a baseline for future research on this extended task. To address the problems of poor global coverage and uneven difficulty distribution, we design a geography-aware and a reward-based sampling strategy, making the dataset better aligned with current mainstream training paradigms while maintaining a reasonable spatial distribution.

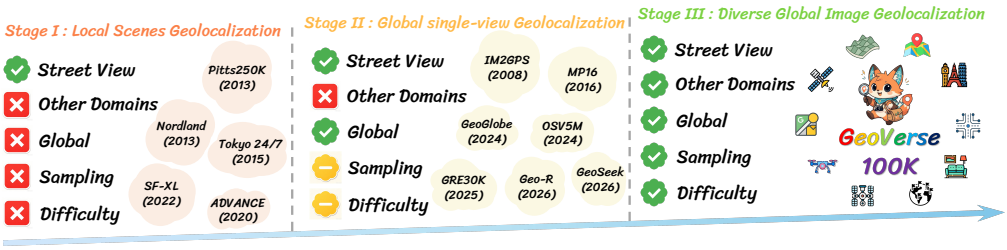


Figure 2: **Evolution of geolocalization datasets.** As the field progresses from local to global and from prediction to reasoning, future geolocalization datasets should include more diverse inputs.

We systematically evaluate general VLMs and specialized geolocalization models. We find that some specialized models underperform general models outside their training domains. Although training in a single domain improves accuracy within that domain, it degrades performance in more than half of the evaluations in other domains, revealing both knowledge transfer and interference across domains. GeoVerse, trained on GeoVerse100K, achieves more balanced performance across all ten domains and the highest aggregate GeoScore for both continuous and discrete domains. These findings highlight the importance of multi-domain training for improving generalization in global image geolocalization. To sum up, our main contributions are summarized as follows:

- We introduce Diverse Global Image Geolocalization and construct GeoVerse100K. We extend the research scope to ten visual domains and develop geographic attribute sampling and reward-aware sampling strategies. These strategies jointly address geographic distribution and reward separability for reinforcement learning, providing a unified data foundation.
- We establish GeoVerse-Val, a benchmark supporting fine-grained analysis. It contains 9,000 images spanning ten visual domains and 37 subcategories. Through comparisons of general and specialized models, we reveal limitations in the cross-domain generalization of existing methods.
- We develop GeoVerse, a unified baseline for diverse geolocalization. We train Qwen3.5-9B with a reinforcement learning method tailored to this task and introduce domain-balanced interleaving and a difficulty-guided curriculum to accommodate differences in data volume and learning difficulty across domains. Experiments show the effectiveness of our sampling and training strategies.

2 Related Work

2.1 Global Image Geolocalization

Global image geolocalization is defined as inferring the capture location of an image from its geographic cues around the world. Traditional methods usually address this task through classification-based, (Weyand et al., 2016; Seo et al., 2018; Müller-Budack et al., 2018; Theiner et al., 2022; Pramanick et al., 2022) retrieval-based (Radenović et al., 2019; Liu et al., 2019; Ge et al., 2020; Haas et al., 2024; Daruna et al., 2026) or the combination of them. (Izbicki et al., 2020; Kordopatis-Zilos et al., 2021; Gadi et al., 2026) With the development of vision-language models, some works further enhance this paradigm with CLIP (Vivanco Cepeda et al., 2023; Li et al., 2023; Haas et al., 2023; Xu et al., 2025b; Zhou et al., 2024; Chen et al., 2025a; Jia et al., 2025a; Lindenberger et al., 2025) and RAG (Jia et al., 2024; 2025b; Ghasemi et al., 2025; Le-Duc et al., 2026). As VLMs and reinforcement learning continue to advance, a series of works begin to train VLMs with RL-based methods (Li et al., 2025a; Jin et al., 2026; Wu et al., 2026; Yu et al., 2026) so that the models can improve performance while also producing reasoning processes. More recent methods also explore the use of external tools (Ji et al., 2026; Jia et al., 2026a; Li et al., 2026b) or multi-model agent workflows (Yang et al., 2026; Shi et al., 2026; Zheng et al., 2026a) to further improve accuracy. However, because most traditional geolocalization datasets focus only on street-view localization, these methods are often restricted to a specific domain, which limits their generalization ability.

2.2 Global Geolocalization Datasets and Benchmarks

As shown in Fig. 2, geolocalization datasets and benchmarks evolve from local scenes to global coverage and from discriminative recognition to generative reasoning. Early datasets such as

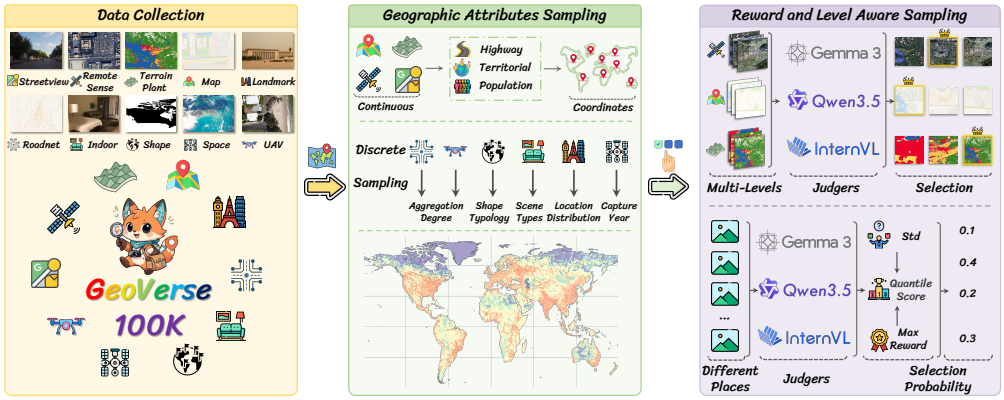


Figure 3: **Construction pipeline of GeoVerse100K.** After data cleaning, the data undergo geographic attribute sampling followed by reward-aware sampling, resulting in a geographically balanced and representative dataset suitable for reinforcement learning with GRPO.

IM2GPS (Hays & Efros, 2008) and YFCC100M (Thomee et al., 2016) are largely built on user-uploaded photos. With the growth of global image platforms and map platforms, later datasets such as OSV5M (Astruc et al., 2024), GeoGlobe (Han et al., 2025), and GeoComp (Song et al., 2025) gradually extend the task to global street-view geolocalization. With the progress of VLMs, datasets such as MP16-Reason (Li et al., 2025a), GRE30K (Wang et al., 2025a), GeoSeek (Jin et al., 2026), and SpotAgenticCoT (Jia et al., 2026a) further provide reasoning annotations including chain-of-thought data and tool-use trajectories for training large models. Although these works significantly expand the boundary of the task, most existing datasets are still built around a single dominant viewpoint and also suffer from geographically imbalanced distributions. Meanwhile, as reinforcement learning becomes a mainstream paradigm for this task, existing datasets still often lack difficulty-aware filtering, making it hard to produce stable and discriminative reward signals during training. At the same time, recent benchmarks increasingly emphasize natural-language location descriptions (Ji et al., 2026; Jin et al., 2026), interpretable reasoning processes (Li et al., 2026a; Talreja et al., 2026), and evaluation for reward modeling (Zheng et al., 2026a; Li et al., 2026b), yet benchmarks that jointly support global coverage and multi-domain inputs remain limited.

3 GeoVerse100k Overview

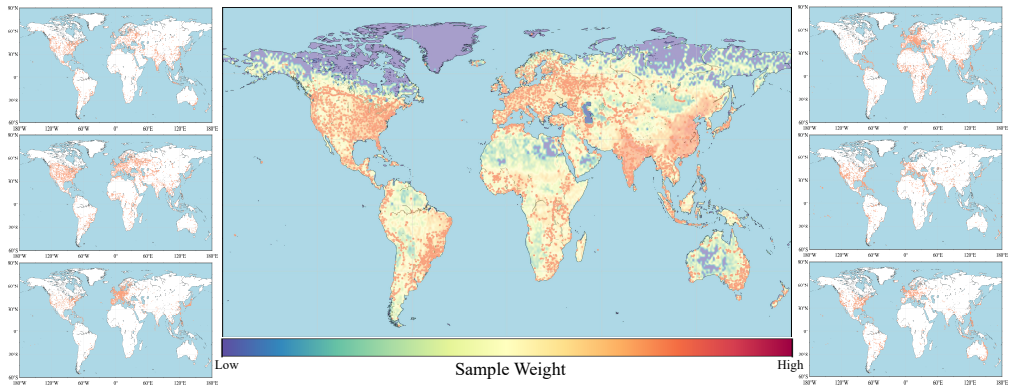
3.1 Diverse Global Image Geolocalization

Diverse Global Image Geolocalization aims to infer the capture location of an input image on a global scale from the visual, semantic, and geographic cues contained in the image. Unlike traditional global image geolocalization, which mainly targets street-view imagery and relies on a single visual domain, this task emphasizes that models must localize images under conditions where multiple sources, viewpoints, and scene types coexist. Thus, the model must not only be capable of identifying location from images but also possess cross-domain generalization ability.

Formally, given an image I, the model needs to learn a mapping from visual observation to geographic coordinates $g = (\phi, \lambda)$ or a discrete geographic location r, where ϕ and λ denote latitude and longitude, respectively. Compared with settings that optimize only on a single data distribution, Diverse Global Image Geolocalization places greater emphasis on the model's ability to adapt to complex real-world input. It can therefore be regarded as a definition of global localization that is closer to open-environment application scenarios.

3.2 Dataset Composition

GeoVerse100k includes about 100,000 images from ten different domains, among which 9,000 images are used for evaluation, and the remainder are used for training. Specifically, GeoVerse100k includes globally distributed continuous domains such as street-view, remote-sensing, plant-and-terrain, and map images. It also includes relatively limited discrete domains such as landmark,



(a) Landmark, Roadnet and Indoor (b) Streetview, Remote Sense, Terrain & Plant and Map (c) Shape, Space and UAV

Figure 4: **Spatial distribution of the ten domains after Geographic Attribute Sampling.** Different colors indicate the sampling weights assigned to continuous domains at each location.

road-network, indoor, shape, space, and UAV images. For continuous-domain data, each domain contains 15,000 training images and 1,500 evaluation images. For discrete-domain data, each domain contains 5,000 training images and 500 evaluation images. All images in GeoVerse100k are sampled from publicly available datasets or websites as shown in Tab. 5. At the same time, we design two sampling strategies to construct a data distribution with suitable location and difficulty distributions at the global scale.

4 Dataset Construction

4.1 Data Collection and Preprocessing

The collected data are first preprocessed to remove corrupted samples and to ensure that each domain covers as broad a geographic distribution as possible. The data sources, sample spaces, and collection procedures of each domain are summarized in Tab. 5. The training data are then progressively down-sampled through two sampling strategies.

4.2 Geographic Attribute Sampling

Geographic Attribute Sampling is designed to address the problem that previous datasets are overly concentrated in specific geographic regions, such as Europe and North America. This issue not only harms model generalization but also introduces geographic bias into the learned model. To mitigate this problem, we design a sampling strategy based on geographic attributes. For continuous domains, we adopt a common hierarchical sampling strategy and use street-view images to construct the shared geographic anchors. Given a budget of $N_{\text{loc}} = 25{,}000$ locations, the quota assigned to country or region i is

$$m_i = N_{\text{loc}} w_i, \qquad w_i = 0.3\frac{A_i}{\sum_j A_j} + 0.4\frac{P_i}{\sum_j P_j} + 0.3\frac{R_i}{\sum_j R_j}, \tag{1}$$

where A_i, P_i, and R_i denote its land area, population, and road length. Within each country, we divide the coordinate space into $1^\circ \times 1^\circ$ cells and sample cell c according to

$$p_{i,c} = \frac{\log(1 + \text{Pop}_{i,c})}{\sum_{c' \in \mathcal{C}_i} \log(1 + \text{Pop}_{i,c'})}, \tag{2}$$

where $\text{Pop}_{i,c}$ is the cell population and $\mathcal{C}_i$ contains eligible cells in country i. The logarithm prevents over-concentration in large cities. These 25,000 coordinates are shared by the street-view, remote-sensing, map, and plant-and-terrain domains. For the discrete domains, each domain is down-sampled to 10,000 samples. Their sampling strategies are tailored to their geographic characteristics. More details are provided in Tab. 6, and the distribution after sampling is shown in Fig. 4.

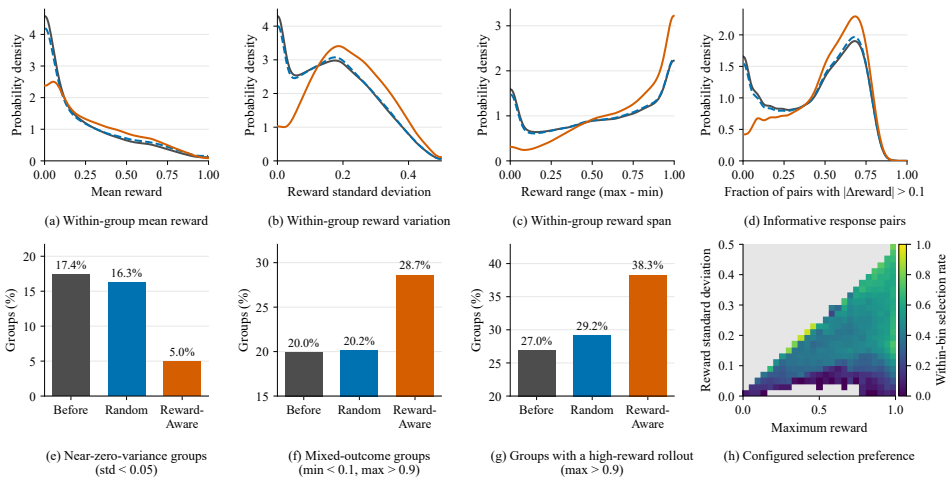


Figure 5: **Comparison between Random and Reward-Aware Sampling.** Near-zero-variance groups have reward standard deviations below 0.05, mixed-outcome groups contain both low-reward and high-reward responses, and informative response pairs have absolute reward differences greater than 0.1. Panel (h) bins the original samples by maximum reward and reward standard deviation, while bins containing fewer than 10 original samples are masked. Gray, blue, and orange represent samples before selection, random sampling, and reward-aware sampling, respectively.

4.3 Reward-Aware Sampling

As geolocalization moves toward reward-driven reinforcement learning, conventional datasets generally lack difficulty-aware filtering. Samples that are either overly difficult or overly easy fail to produce stable and effective reward signals. From the perspective of a widely adopted GRPO policy, desirable training data should both induce sufficiently large relative reward gaps and contain at least some high-quality candidate solutions. To this end, after Geographic Attribute Sampling, we further perform Reward-Aware Sampling. Specifically, for each image x_i in domain d, we use $M = 3$ generation models, each producing $K = 8$ candidate responses. The k-th response from model m, denoted by $y_{i,m,k}$, predicts location $\hat{g}_{i,m,k}$ and receives the distance-aware reward

$$r_{i,m,k} = \mathcal{R}(x_i, y_{i,m,k}) = \begin{cases} \exp\left(-10\,\dfrac{d(\hat{g}_{i,m,k}, g_i^*)}{D_{\max}}\right), & d(\hat{g}_{i,m,k}, g_i^*) < D_{\max}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where g_i^* is the ground-truth location, $d(\cdot,\cdot)$ is the great-circle distance, and $D_{\max}$ is the maximum distance threshold. Across the $M \times K$ rewards, we compute the population standard deviation σ_i and maximum reward $r_i^{\max} = \max_{m,k} r_{i,m,k}$. These statistics serve as proxies for reward separability and the availability of high-quality candidate solutions, respectively.

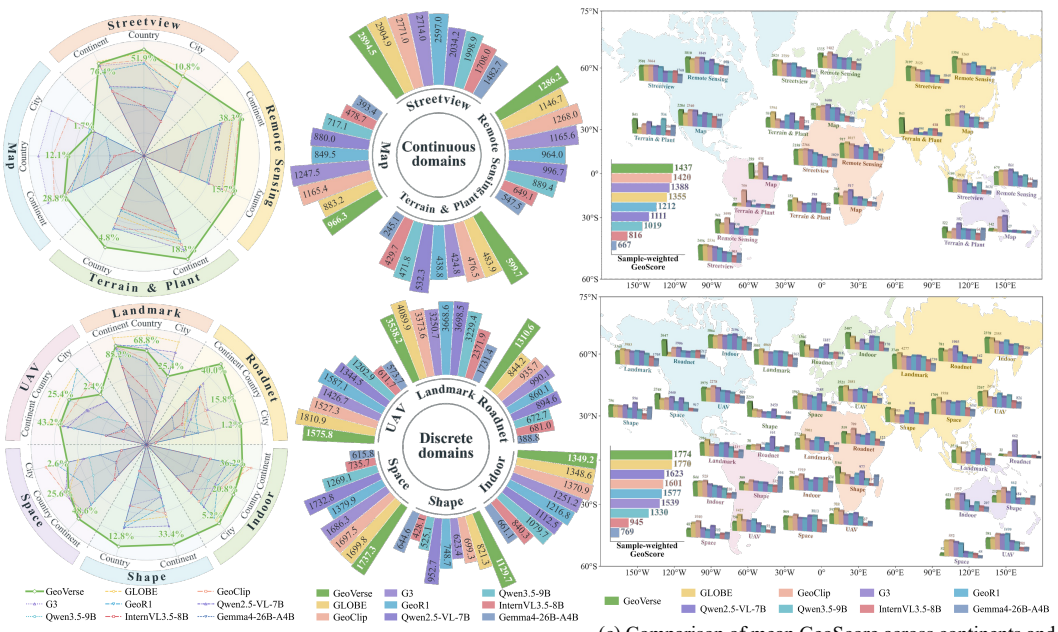
To account for differences in reward distributions across domains, we convert both statistics into within-domain percentile scores, $s_i^{\text{std}} = \text{RankPct}(\sigma_i)$ and $s_i^{\max} = \text{RankPct}(r_i^{\max})$, where $\text{RankPct}(\cdot) \in [0, 1]$. For a domain containing N_d candidate images, we define

$$v_i = \left(s_i^{\text{std}}\right)^{\alpha} \left(s_i^{\max}\right)^{\beta}, \qquad p_i = \frac{v_i}{\sum_{j=1}^{N_d} v_j}. \tag{4}$$

We then perform weighted sampling without replacement using p_i, following the domain quotas specified in Sec. 3. For remote-sensing, map, and plant-and-terrain domains, we evaluate multiple image levels at each location selected through reward-aware sampling of street-view images. For location u, the level-selection weights are

$$q_{u,t} = \frac{\sigma_{u,t} + \epsilon}{\sum_{t'=1}^{n_u} (\sigma_{u,t'} + \epsilon)}, \tag{5}$$

where n_u is the number of available levels, $\sigma_{u,t}$ is the reward standard deviation for level t, and ϵ ensures numerical stability. Sampling with these weights preserves the shared locations while favoring levels with greater candidate reward variation. As shown in Fig. 5, compared with random sampling, reward-aware sampling reduces near-zero-variance and increases mixed-outcome groups.



(a) Comparison of accuracy. (b) Comparison of mean GeoScore. (c) Comparison of mean GeoScore across continents and weighted GeoScore.

Figure 6: **Comparison of GeoVerse with other models.** GeoVerse maintains balanced performance across domains and outperforms the other models in weighted GeoScore.

5 Benchmark and Baseline

Benchmark Construction. GeoVerse-Val is sampled independently. For continuous and discrete domains, we sample 1,500 and 500 images, respectively, resulting in 9,000 images in total. Each sample is annotated with its latitude and longitude, domain label, sub-category label, and continent label, enabling evaluation under different fine-grained settings and across different regions.

Evaluation Metrics and Protocol. We evaluate all methods under a coordinate-based geolocalization setting. Given an input image, the model is required to output a predicted location in textual form, which is then converted into latitude and longitude using OpenCage. For models that directly output coordinates, we do not apply geocoding. The localization error is computed as the great-circle distance between the predicted coordinate and the ground-truth coordinate. We report threshold accuracy at multiple distance thresholds, denoted as $\mathrm{Acc}@\tau$, measuring the percentage of predictions whose errors are smaller than τ kilometers. Additionally, we report GeoScore as a continuous distance-aware metric, defined as

$$\text{GeoScore} = 5000 \times \exp\left(\frac{-10e}{18050}\right), \tag{6}$$

where e is the localization error in kilometers.

Training of Baseline. To provide a comparable baseline for future research on Diverse Global Image Geolocalization, we train Qwen3.5-9B with GRPO using the distance-aware reward in Eq. (3) and a format reward. During training, to accommodate the large number of visual domains in our task, we further introduce a domain-balanced reward curriculum composed of **domain-ratio interleaving** and **difficulty scheduling**. For domain balancing, we mix data from each continuous domain at three times the rate of each discrete domain, matching the 3:1 ratio between their dataset sizes. This prevents samples from a single domain from forming long consecutive blocks in the training sequence. For difficulty scheduling, we use the domain-wise medians of reward mean and standard deviation to divide samples into four groups, which are presented in the order of high-mean / low-variance, low-mean / low-variance, high-mean / high-variance, and low-mean / high-variance. These two strategies preserve a similar domain composition across adjacent batches. At the same time, they allow the model to learn from easier samples at the early stage, when its capability is still limited, and gradually move to harder samples as its capability improves.

	Streetview	Remote Sensing	Terrain Plant	Map	Landmark	Roadnet	Indoor	Shape	Space	UAV
Streetview	**2782.1**	**1094.4**	479.9	716.1	**3298.2**	810.1	**1298.9**	790.3	1228.7	**1295.3**
Remote Sensing	2407.5	**1078.1**	479.1	711.5	**3182.4**	809.7	1013.6	745.2	1209.7	**1266.7**
Terrain & Plant	2059.8	991.7	**600.0**	729.4	**3165.3**	809.0	920.3	845.2	1250.9	1163.4
Map	2323.6	1007.0	488.9	**860.9**	**3149.8**	829.4	1013.2	772.6	1231.2	1119.3
Landmark	2255.2	949.8	398.9	563.6	**3267.7**	660.7	1071.8	750.0	1273.9	1098.1
Roadnet	2024.5	854.1	531.1	627.6	**2883.3**	**996.0**	880.4	775.2	1202.8	1004.2
Indoor	2292.4	957.6	475.3	669.9	**3050.5**	787.0	1045.7	763.9	1221.1	1162.4
Shape	2000.5	833.6	**553.6**	636.4	**2999.6**	866.7	836.3	**939.9**	1126.6	971.7
Space	2019.9	857.3	417.3	503.0	**2984.7**	662.0	945.0	751.9	**1431.6**	858.8
UAV	2315.4	977.5	490.6	545.5	**3119.0**	671.5	1046.9	692.8	**1338.9**	**1227.7**
Qwen3.5-9B	1998.9	889.4	471.8	717.1	3229.4	672.7	1079.7	525.1	1269.1	1202.9
GeoVerse	**2894.5**	**1286.2**	**599.7**	**966.3**	**3538.2**	**1310.6**	**1349.2**	**1129.7**	**1737.3**	**1575.8**

Figure 7: **Cross-domain transfer performance.** Rows: training domains. Columns: evaluation domains. Each cell reports the mean GeoScore. Color intensity is normalized independently within each column relative to its highest score: darker colors indicate stronger relative performance.

6 Experiments

6.1 Experimental Details

GeoVerse is trained with GRPO for one epoch on 8 $\times$ NVIDIA H20 GPUs. All evaluations are conducted on a single NVIDIA H20 GPU. Following prior work, the distance threshold τ is set to 2,500 km, 750 km, and 25 km for continent, country, and city-level evaluation, respectively.

6.2 Main Results

Performance on GeoVerse-Val. Fig. 6 compares GeoVerse with other models. Across different distance thresholds, GeoVerse ranks first in 13 of the 27 comparisons and among the top three in 22. It shows clear advantages in less common domains, such as remote sensing, plant-and-terrain, and road-network (e.g., 40.0% vs. 29.6% on road-network). In domains where specialized models receive dedicated training, such as street-view and landmark, GeoVerse performs on par with the best models and even surpasses them at some distance thresholds (e.g., 51.9% vs. 50.3% on street-view). In terms of mean GeoScore, GeoVerse also ranks first in 5 of the 10 domains and among the top three in 9. These results highlight that GeoVerse achieves balanced geolocalization performance across domains and spatial scales. In addition, general VLMs without task-specific training outperform several specialized geolocalization models in some domains. This suggests that specialized models may rely too heavily on specific training distributions, limiting their generalization.

Performance across Continents. To examine performance across geographic regions, we further compare the mean GeoScore of GeoVerse and other models across six continents and ten domains, as shown in Fig. 6(c). After aggregating results over all samples with weighting by sample count, GeoVerse achieves overall GeoScore of 1436.7 and 1773.5 for continuous and discrete domains, respectively, ranking first in both categories. Across all continent–domain combinations, GeoVerse ranks first in 12 and among the top three in 34. These results demonstrate the ability of GeoVerse to generalize across geographic distributions.

Domain-wise Analysis. We evaluate ten models, each trained on a single domain, across all ten domains, as shown in Fig. 7. These models substantially outperform the Qwen3.5-9B baseline in their respective training domains. However, their GeoScores fall below the baseline in more than half of the remaining 90 model–domain combinations. This suggests that training on a single domain can cause models to forget general cross-domain knowledge. We also observe positive transfer between domains. For example, the model trained only on street-view imagery outperforms the baseline in domains such as remote sensing and indoor scenes. Together, these results show that knowledge transfer and interference can coexist across domains and geolocalization therefore needs to cover a broader range of input domains to improve model generalization.

6.3 Ablation Analysis

Effectiveness of Components. Results in Tab. 1 show that geographic attribute sampling, reward-aware sampling, domain-balanced interleaving, and difficulty scheduling can improve performance.

Table 1: **Ablation study results reported on GeoVerse-Val.** Geo., Reward., DB, and DS denote geographic attribute sampling, reward-aware sampling, domain-balanced interleaving, and difficulty scheduling, respectively. Mean GeoScore is reported and τ is set to 2500.

Geo.	Reward.	DB	DS	Street-View		Remote Sensing		Space		UAV	
				GeoScore↑	Acc@τ ↑	GeoScore↑	Acc@τ ↑	GeoScore↑	Acc@τ ↑	GeoScore↑	Acc@τ ↑
Qwen3.5-9B				1998.85	54.20%	889.45	26.00%	1269.13	35.60%	1202.87	35.40%
Qwen2.5-VL-7B				2034.21	55.53%	996.72	29.93%	1732.80	47.80%	1344.47	38.80%
InternVL3.5-8B				1707.97	51.47%	649.05	20.00%	735.70	20.40%	611.74	17.60%
Gemma4-26B-A4B				1482.73	40.87%	547.52	16.80%	615.84	16.60%	573.73	17.40%
				2713.99	73.40%	1207.59	36.27%	1592.47	43.80%	1411.53	41.80%
✓		✓		2732.02	73.93%	1219.83	36.20%	1578.12	46.20%	1460.88	42.20%
✓	✓			2760.37	74.00%	1226.00	36.87%	1628.98	45.60%	1555.33	43.20%
✓	✓	✓		2830.36	75.27%	1262.83	37.87%	1638.66	46.20%	1568.06	**45.00%**
✓	✓		✓	2864.93	75.87%	1228.85	36.93%	1736.14	**48.60%**	1510.94	41.80%
✓	✓	✓	✓	**2894.46**	**76.40%**	**1286.23**	**38.33%**	**1737.30**	**48.60%**	**1575.76**	43.20%

Table 2: **Discussion of sampling algorithm**.

Mode	Street-View	Shape	Landmark
Qwen3.5-9B	1998.85	525.06	3229.37
Random	2713.99	908.61	3387.37
w/o Geo.	2753.00	978.13	3448.71
w/o Reward.	2732.02	953.19	3391.12
Geo. + Reward.	**2894.46**	**1129.73**	**3538.17**

Table 3: **Discussion of domain balance**.

Schedule	Street-View	Shape	Landmark
Qwen3.5-9B	1998.85	525.06	3229.37
Random	2864.93	1086.00	3469.94
Sequential	2545.97	952.02	3332.98
Equal Ratio	**2949.51**	923.90	3324.66
Balance	2894.46	**1129.73**	**3538.17**

Moreover, combining all components is the best. It is also notable that domains differ in their sensitivity to domain balancing and difficulty scheduling. Domain balancing can mitigate the negative effects of difficulty scheduling on some domains, highlighting their complementarity.

Ablation on Sampling Strategies. Tab. 2 shows that combining both strategies yields the best performance, while removing either one reduces performance across all three domains. This suggests that the three domains favor the same sampling configuration. The results also show that reward-aware sampling provides greater gains than geographic attribute sampling.

Discussion of Domain Balance. We also make a discussion about how data from different domains are balanced during training. As shown in Tab. 3, training in individual domains sequentially in a fixed order substantially degrades performance, even compared with random mixing. This decline occurs because sustained optimization on a single domain weakens the model's previously acquired capabilities in other domains, consistent with the findings from our domain-wise evaluation. In addition, mixing data from different domains in equal proportions favors continuous domains with larger datasets and substantially reduces performance on discrete domains. By contrast, mixing data within each batch in proportion to domain dataset sizes preserves performance in stronger domains while improving performance in the others, thereby yielding more balanced overall results.

7 Conclusions

In this work, we introduce Diverse Global Image Geolocalization and present GeoVerse100K and GeoVerse-Val. GeoVerse100K employs geographic attribute sampling and reward-aware sampling to improve geographic balance and data suitability for reinforcement learning. Experiments reveal limitations in existing models' cross-domain generalization, while the GeoVerse baseline achieves more balanced performance across domains. These resources provide a unified foundation for developing and evaluating geolocalization models with stronger generalization.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature discovery, language editing, and text condensation. Generative models were also used to produce candidate responses for reward-aware sampling, as described in the methodology. The authors reviewed and revised the AI-assisted text for accuracy and consistency with the methods and experimental results, and verified the cited references against their original sources. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

GeoVerse100K is constructed from publicly available datasets and websites to support research on diverse global image geolocalization. We recognize that inferring geographic locations from images may expose sensitive location information or enable unwanted surveillance. The dataset and models should not be used to identify or track individuals without their consent, and redistribution of source imagery must comply with the applicable licenses and terms of use.

REPRODUCIBILITY STATEMENT

We describe the dataset construction, sampling procedures, baseline training, and evaluation protocol in Secs. 4 to 6. Additional sampling details, training hyperparameters, and prompts are provided in the supplementary material. The code and dataset will be made publicly available to support independent reproduction and evaluation, subject to the applicable licenses and terms of use for source data.

REFERENCES

Geoguessr - let's explore the world! https://www.geoguessr.com/, 2025. Accessed: 2025-10-14.

Tuxun: explore the world. https://tuxun.fun/, 2025. Accessed: 2025-10-14.

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.09.127. URL https://www.sciencedirect.com/science/article/pii/S0925231222012188.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, and Loic Landrieu. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21967–21977, June 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4868–4878, 2022. doi: 10.1109/CVPR52688.2022.00483.

Gabriele Berton, Alex Stoken, Barbara Caputo, and Carlo Masone. Earthloc: Astronaut photography localization by indexing earth from space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12754–12764, June 2024.

Benedikt Blumenstiel, Nassim Ait Ali Braham, Conrad M Albrecht, Stefano Maurogiovanni, and Paolo Fraccaro. SSL4EO-S12 v1.1: A multimodal, multiseasonal dataset for pretraining, updated. *arXiv preprint arXiv:2503.00168*, 2025.

Ron Campos, Ashmal Vayani, Parth Parag Kulkarni, Rohit Gupta, Aizan Zafar, Aritra Dutta, and Mubarak Shah. Gaea: A geolocation aware conversational assistant. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5236–5246. IEEE, 2026.

Boyi Chen, Zhangyu Wang, Fabian Deuser, Johann Maximilian Zollner, and Martin Werner. Enhancing contrastive learning for geolocalization by discovering hard negatives on semivariograms. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '25, pp. 1162–1165, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400720864. doi: 10.1145/3748636.3763220. URL https://doi.org/10.1145/3748636.3763220.

Quan Chen, Tingyu Wang, Rongfeng Lu, Yu Liu, Bolun Zheng, and Zhedong Zheng. Scale-adaptive uav geo-localization via height-aware partition learning, 2025b. URL https://arxiv.org/abs/2412.11535.

Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3223–3230. IEEE, 2017.

Fenghua Cheng, Jinxiang Wang, Sen Wang, Zi Huang, and Xue Li. Geoexplain: Multimodal reasoning based on hierarchy of visual information in street view, 2026. URL https://arxiv.org/abs/2506.16633.

Jaeyoung Choi, Claudia Hauff, Olivier Van Laere, and Bart Thomee. The placing task at mediaeval 2016. In *Working Notes Proceedings of the MediaEval 2016 Workshop*, volume 1739. CEUR-WS.org, 2016.

Ivan Cisneros, Peng Yin, Ji Zhang, Howie Choset, and Sebastian Scherer. Alto: A large-scale dataset for uav visual place recognition and localization, 2022. URL https://arxiv.org/abs/2207.12317.

Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23182–23190, 2023. doi: 10.1109/CVPR52729.2023.02220.

Muhammad Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7132–7142, October 2025.

Angel Daruna, Nicholas Meegan, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Geosurge: Geo-localization using semantic fusion with hierarchy of geographic embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41396–41405, June 2026.

do-me. Flickr-geo. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/do-me/Flickr-Geo.

Zhiyang Dou, Zipeng Wang, Xumeng Han, Guorong Li, Zhenjun Han, and Zhipei Huang. Towards interactive global geolocation assistant. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 441–458, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37271-0.

Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14340–14349, 2021. URL https://github.com/abhimanyudubey/GeoYFCC.

David Flatow, Mor Naaman, Ke Eddie Xie, Yana Volkovich, and Yaron Kanza. On the accuracy of hyper-local geotagging of social media content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pp. 127–136, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450333177. doi: 10.1145/2684822.2685296. URL https://doi.org/10.1145/2684822.2685296.

Hari Krishna Gadi, Daniel Matos, Hongyi Luo, Lu Liu, Yongliang Wang, Yanfeng Zhang, and Liqiu Meng. Hierloc: Hyperbolic entity embeddings for hierarchical visual geolocation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 132079–132100, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d61819e9b4a607b8448de762235148c4-Paper-Conference.pdf.

Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 369–386, Cham, 2020. Springer International Publishing.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Xinyu Geng, Yanjing Xiao, Yuyang Zhang, Hanwen Wang, Xinyan Liu, Rui Min, Tianqing Fang, and Yi R. Fung. Geobrowse: A geolocation benchmark for agentic tool use with expert-annotated reasoning traces, 2026. URL https://arxiv.org/abs/2604.04017.

Narges Ghasemi, Amir Ziashahabi, Salman Avestimehr, and Cyrus Shahabi. Geotoken: Hierarchical geolocalization of images via next token prediction. In *2025 IEEE International Conference on Data Mining (ICDM)*, pp. 1214–1223, 2025. doi: 10.1109/ICDM65498.2025.00130.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocalization, February 2023.

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

Mohammad Habedi. Large dataset of geotagged images. Kaggle dataset, 2023. URL https://www.kaggle.com/datasets/habedi/large-dataset-of-geotagged-images.

Xiao Han, Chen Zhu, Hengshu Zhu, and Xiangyu Zhao. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In Luiza Antonie, Jian Pei, Xiaohui Yu, Flavio Chierichetti, Hady W. Lauw, Yizhou Sun, and Srinivasan Parthasarathy (eds.), *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pp. 814–825. ACM, 2025. doi: 10.1145/3711896.3737141. URL https://doi.org/10.1145/3711896.3737141.

James Hays and Alexei A. Efros. Im2gps: estimating geographic information from a single image. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008. doi: 10.1109/CVPR.2008.4587784.

Yujun Hou, Matias Quintana, Maxim Khomiakov, Winston Yap, Jiani Ouyang, Koichi Ito, Zeyu Wang, Tianhong Zhao, and Filip Biljecki. Global streetscapes – a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215:216–238, 2024. doi: 10.1016/j.isprsjprs.2024.06.023.

Di Hu, Xuhong Li, Lichao Mou, Pu Jin, Dong Chen, Liping Jing, Xiaoxiang Zhu, and Dejing Dou. Cross-task transfer for geotagged audiovisual aerial scene recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 68–84, Cham, 2020. Springer International Publishing.

Jingyuan Huang, Jen-tse Huang, Ziyi Liu, Xiaoyuan Liu, Wenxuan Wang, and Jieyu Zhao. AI sees your Location—But with a bias toward the wealthy world. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18019–18039, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.910. URL https://aclanthology.org/2025.emnlp-main.910/.

Ofer Idan, Dan Badur, Yosi Keller, and Yoli Shavit. Lavpr: Benchmarking language and vision for place recognition. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 356–373, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37281-9.

Mike Izbicki, Evangelos E. Papalexakis, and Vassilis J. Tsotras. Exploiting the earth's spherical geometry to geolocate images. In Ulf Brefeld, Elisa Fromont, Andreas Hotho, Arno Knobbe, Marloes Maathuis, and Céline Robardet (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 3–19, Cham, 2020. Springer International Publishing. ISBN 978-3-030-46147-8.

Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. Game4loc: A UAV geo-localization benchmark from game data. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 3913–3921. AAAI Press, 2025a. doi: 10.1609/AAAI.V39I4.32409. URL https://doi.org/10.1609/aaai.v39i4.32409.

Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. Mmgeo: Multimodal compositional geolocalization for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25165–25175, October 2025b.

Yuxiang Ji, Yong Wang, Ziyu Ma, Yiming Hu, Hailang Huang, Xuecai Hu, Guanhua Chen, Liaoni Wu, and Xiangxiang Chu. Thinking with map: Reinforced parallel map-augmented agent for geolocalization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17991–18004, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.894. URL https://aclanthology.org/2026.findings-acl.894/.

Furong Jia, Lanxin Liu, Ce Hou, Fan Zhang, Xinyan Liu, and Yu Liu. Towards Interpretable Geolocalization: A Concept-Aware Global Image-GPS Alignment Framework, September 2025a.

Furong Jia, Ling Dai, Wenjin Deng, Fan Zhang, Chen Hu, Daxin Jiang, and Yu Liu. Spotagent: Grounding visual geo-localization in large vision-language models through agentic reasoning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 2062–2073, New York, NY, USA, 2026a. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817948. URL https://doi.org/10.1145/3770855.3817948.

Pengyue Jia, Yiding Liu, Xiaopeng Li, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, Dawei Yin, and Xiangyu Zhao. G3: An effective and adaptive framework for worldwide geolocalization using large multi-modality models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 53198–53221. Curran Associates, Inc., 2024. doi: 10.52202/079017-1685. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5f2f5882d6166d814629ada0cd95f9a0-Paper-Conference.pdf.

Pengyue Jia, Seongheon Park, Song Gao, Xiangyu Zhao, and Sharon Li. Georanker: Distance-aware ranking for worldwide image geolocalization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17673–17699. Curran Associates, Inc., 2025b. doi: 10.52202/085713-0596. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19b30d80dbb996efe36fd51a702800b7-Paper-Conference.pdf.

Pengyue Jia, Derong Xu, Yingyi Zhang, Xiaopeng Li, Wenlin Zhang, Yi Wen, Yuanshao Zhu, and Xiangyu Zhao. GeoRouter: Dynamic Paradigm Routing for Worldwide Image Geolocalization, 2026b.

Pengyue Jia, Yingyi Zhang, Xiangyu Zhao, and Sharon Li. GeoArena: Evaluating open-world geographic reasoning in large vision-language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20886–20901, San Diego, California, United States, July 2026c. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.956. URL https://aclanthology.org/2026.acl-long.956/.

Modi Jin, Yiming Zhang, Boyuan Sun, Dingwen Zhang, Ming-Ming Cheng, and Qibin Hou. Geoagent: Learning to geolocate everywhere with reinforced geographic characteristics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41352–41364, June 2026.

Giorgos Kordopatis-Zilos, Panagiotis Galopoulos, Symeon Papadopoulos, and Ioannis Kompatsiaris. Leveraging efficientnet and contrastive learning for accurate global-scale location estimation. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, pp. 155–163, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384636. doi: 10.1145/3460426.3463644. URL https://doi.org/10.1145/3460426.3463644.

Oskar Kristoffersen, Alba Reinders Sánchez, Morten Rieger Hannemose, Anders Bjorholm Dahl, and Dim P. Papadopoulos. Mmlandmarks: a cross-view instance-level benchmark for geo-spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26452–26464, June 2026.

Tung-Duong Le-Duc, Hoang-Quoc Nguyen-Son, and Minh-Son Dao. Geosearch: Augmenting worldwide geolocalization with web-scale reverse image search and image matching. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3913–3918, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3809936. URL https://doi.org/10.1145/3805712.3809936.

Jinnao Li, Tingzhu Chen, and Changbo Wang. Gtpred: Benchmarking mllms for interpretable geolocalization and time-of-capture prediction, 2026a. URL https://arxiv.org/abs/2601.13207.

Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. GeoReasoner: Geo-localization with reasoning in street views using a large vision-language model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29222–29233. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/li24ch.html.

Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. In *Advances in Neural Information Processing Systems*, 2025a.

Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. From Pixels to Places: A Systematic Benchmark for Evaluating Image Geolocalization Ability in Large Language Models, 2025b.

Mingkun Li, Ziming Wang, Guang Huo, Wei Chen, and Xiaoning Zhao. A Deep Learning Framework with Geographic Information Adaptive Loss for Remote Sensing Images based UAV Self-Positioning, 2025c.

Qiujun Li, Zijin Xiao, Xulin Wang, Zhidan Ma, Cheng Yang, and Haifeng Li. LocationAgent: A Hierarchical Agent for Image Geolocation via Decoupling Strategy and Evidence from Parametric Knowledge, 2026b.

Sirun Li, Minghao Liu, Ling Dai, Yong Li, Haoxin Lyu, Junting Zhou, and Fan Zhang. SIGNPOST-Bench: Benchmarking Text-Vision Conflict Resolution in Multimodal Large Language Models, 2026c. URL https://arxiv.org/abs/2608.04244.

Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023. ISSN 1569-8432. doi: https://doi.org/10.1016/j.jag.2023.103497. URL https://www.sciencedirect.com/science/article/pii/S1569843223003217.

Zhuoxiao Li, Wenzong Ma, Taoyu Wu, Jinjing Zhu, Shuai Zhang, Jing OU, Tongyan Hua, Yinrui Ren, Rongjun Qin, Hui Xiong, and Wufan Zhao. Beyond a single light: A large-scale aerial dataset for urban scene reconstruction under varying illumination, 2026d. URL https://arxiv.org/abs/2512.14200.

Philipp Lindenberger, Paul-Edouard Sarlin, Jan Hosang, Marc Pollefeys, Simon Lynen, and Eduard Trulls. Scaling image geo-localization to continent level. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 134057–134092. Curran Associates, Inc., 2025. doi: 10.52202/085713-4469. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c2891461cb314d016da16efa06f7f1d5-Paper-Conference.pdf.

Liu Liu, Hongdong Li, and Yuchao Dai. Stochastic attraction-repulsion embedding for large scale image localization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2570–2579, Seoul, Korea (South), October 2019. IEEE.

Tao Liu, Kan Ren, and Qian Chen. Diffusionuavloc: Visually prompted diffusion for cross-view uav localization. *IEEE Internet of Things Journal*, 13(8):17169–17179, 2026a. doi: 10.1109/JIOT.2026.3661248.

Yi Liu, Gelei Deng, Junchen Ding, Yuekang Li, Tianwei Zhang, Weisong Sun, Yaowen Zheng, and Jingquan Ge. Mission: Impossible - image-based geolocation with large vision language models. *Proc. Priv. Enhancing Technol.*, 2025(4):410–428, 2025. doi: 10.56553/POPETS-2025-0137. URL https://doi.org/10.56553/popets-2025-0137.

Zeping Liu, Ni Lao, Zhangyu Wang, Junfeng Jiao, and Gengchen Mai. Gair: Location-aware self-supervised contrastive pre-training with geo-aligned implicit representations. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:166–182, 2026b. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2026.04.035. URL https://www.sciencedirect.com/science/article/pii/S092427162600208X.

Weidi Luo, Tianyu Lu, Qiming Zhang, Xiaogeng Liu, Bin Hu, Yue Zhao, Jieyu Zhao, Song Gao, Patrick McDaniel, Zhen Xiang, and Chaowei Xiao. Doxing via the lens: Revealing location-related privacy leakage on multi-modal large reasoning models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 130452–130501, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d41f8403e9bb5141bc2c81fad7658185-Paper-Conference.pdf.

Chuanlong Ma, Jiuxin Cao, Jiayu Yuan, Lijia Yu, Zixuan Li, and Bo Liu. Clue2geo: Fine-grained image geolocation via cluemap and multi-stage fine-tuning, 2025. URL https://openreview.net/forum?id=PmC9BHFy5w.

Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter. Granular privacy control for geolocation with vision language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17240–17292, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.957. URL https://aclanthology.org/2024.emnlp-main.957/.

Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In Martial Hebert, Vittorio Ferrari, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 575–592, Germany, October 2018. Springer Verlag. ISBN 9783030012571. doi: 10.1007/978-3-030-01258-8_35. Funding information: Acknowledgement. This work is financially supported by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft, project number: EW 134/4-1).; 15th European Conference on Computer Vision, ECCV 2018 ; Conference date: 08-09-2018 Through 14-09-2018.

Ogechi Blessing Onuoha, Bjørn Sand Jensen, Jan Paul Siebert, and Frank E. Pollick. Indoor 3.6m : A dataset and benchmark for indoor image geolocation, 2026. URL https://openreview.net/forum?id=Nw7vkJKHba.

Yiwei Ou, Xiaobin Ren, Ronggui Sun, Guansong Gao, Ziyi Jiang, Kaiqi Zhao, and Manfredo Manfredini. Mms-vpr: Multimodal street-level visual place recognition dataset and benchmark, 2025. URL https://arxiv.org/abs/2505.12254.

Shraman Pramanick, Ewa M. Nowara, Joshua Gleason, Carlos D. Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization inthewild. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pp. 196–215, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19838-0. doi: 10.1007/978-3-031-19839-7_12. URL https://doi.org/10.1007/978-3-031-19839-7_12.

Zhaofang Qian, Hardy Chen, Zeyu Wang, Li Zhang, Zijun Wang, Xiaoke Huang, Hui Liu, Xianfeng Tang, Zeyu Zheng, Haoqin Tu, Cihang Xie, and Yuyin Zhou. Where on earth? a vision-language benchmark for probing model geolocation skills across scales, 2025. URL https://arxiv.org/abs/2510.10880.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019. doi: 10.1109/TPAMI.2018.2846566.

Anindya Sarkar, Srikumar Sastry, Aleksis Pirinen, Chongjie Zhang, Nathan Jacobs, and Yevgeniy Vorobeychik. Gomaa-geo: Goal modality agnostic active geo-localization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 104934–104964. Curran Associates, Inc., 2024. doi: 10.52202/079017-3332. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/bd8b52c2fefdb37e3b3953a37408e9dc-Paper-Conference.pdf.

Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 544–560, Cham, 2018. Springer International Publishing.

Weimin Shi, Xiang Li, Kaige Li, Junhao Fang, Qiang Zhou, Qichuan Geng, and Zhong Zhou. Geobayes: Probabilistic image geo-localization inference via sequential bayesian updating. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 8997–9005. AAAI Press, 2026. doi: 10.1609/AAAI.V40I11.37855. URL https://doi.org/10.1609/aaai.v40i11.37855.

Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet, Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and Xiuying Chen. Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework. *CoRR*, abs/2502.13759, 2025. doi: 10.48550/ARXIV.2502.13759. URL https://doi.org/10.48550/arXiv.2502.13759.

Vedika Srivastava, Hemant Kumar Singh, and Jaisal Singh. ISS-Geo142: A Benchmark for Geolocating Astronaut Photography from the International Space Station, 2025.

Abby Stylianou, Hong Xuan, Maya Shende, Jonathan Brandt, Richard Souvenir, and Robert Pless. Hotels-50k: A global hotel recognition dataset. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. In *Proceedings of the IEEE International Conference on Robotics and Automation Workshop on Long-Term Autonomy*, 2013.

Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7199–7209, 2018.

Mohit Talreja, Joshua Diao, Jim James, Radu Casapu, Tejas Santanam, Ethan Mendes, Alan Ritter, Wei Xu, and James Hays. GeoRC: A benchmark for geolocation reasoning chains. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40540–40564, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1883. URL https://aclanthology.org/2026.acl-long.1883/.

Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable semantic photo geolocation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 750–760, January 2022.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, January 2016. ISSN 0001-0782. doi: 10.1145/2812802. URL https://doi.org/10.1145/2812802.

Akihiko Torii, Josef Sivic, Tomá Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 883–890, 2013. doi: 10.1109/CVPR.2013.119.

Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8690–8701. Curran Associates, Inc., 2023. doi: 10.52202/075280-0379. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b57aaddf85ab01a2445a79c9edc1f4b-Paper-Conference.pdf.

Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2621–2630, 2017.

Shruti Vyas, Chen Chen, and Mubarak Shah. Gama: Cross-view video geo-localization. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 440–456, Cham, 2022. Springer Nature Switzerland.

Chun Wang, Xiaojun Ye, Xiaoran Pan, Zihao Pan, Haofan Wang, and Yiren Song. Gre suite: Geo-localization inference via fine-tuned vision-language models and enhanced reasoning chains. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 92379–92409. Curran Associates, Inc., 2025a. doi: 10.52202/085713-3091. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/854a9ab0f323b841955e70ca383b27d1-Paper-Conference.pdf.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. URL https://arxiv.org/abs/2508.18265.

Xiaonan Wang, Bo Shao, and Hansaem Kim. AI knows where you are: Exposure, bias, and inference in multimodal geolocation with KoreaGEO. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9886–9903, Suzhou, China, 2025c. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.501. URL https://aclanthology.org/2025.emnlp-main.501/.

Yikun Wang, Zuyan Liu, Ziyi Wang, Han Hu, Pengfei Liu, and Yongming Rao. GeoVista: Web-Augmented Agentic Visual Reasoning for Geolocalization, 2025d.

Zhiqiang Wang, Dejia Xu, Rana Muhammad Shahroz Khan, Yanbin Lin, Zhiwen Fan, and Xingquan Zhu. Llmgeo: Benchmarking large language models on image geolocation in-the-wild, 2024. URL https://arxiv.org/abs/2405.20363.

Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2623–2632, June 2020. doi: 10.1109/CVPR42600.2020.00270. URL https://doi.org/10.1109/CVPR42600.2020.00270.

Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet - photo geolocation with convolutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 37–55, Cham, 2016. Springer International Publishing.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Biao Wu, Meng Fang, Ling Chen, Ke Xu, Tao Cheng, and Jun Wang. Vision-language reasoning for geolocalization: A reinforcement learning approach. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 33890–33898. AAAI Press, 2026. doi: 10.1609/AAAI.V40I40.40681. URL https://doi.org/10.1609/aaai.v40i40.40681.

Shuo Xing, Zezhou Sun, Shuangyu Xie, Kaiyuan Chen, Yanjia Huang, Yuping Wang, Jiachen Li, Dezhen Song, and Zhengzhong Tu. Can large vision language models read maps like a human?, 2025. URL https://arxiv.org/abs/2503.14607.

Chenhui Xu, Fuxun Yu, Mike Bianco, Jacob Kovarskiy, Raphael Tang, Qi Zhang, Zirui Xu, William Levine, Brandon Dubbs, Heming Liao, C. Burgess, Suvam Bag, Jay Patravali, Rupanjali Kukal, Mikael Figueroa, Rishi Madhok, Nikolaos Karianakis, and Jinjun Xiong. Unlocking zero-shot geospatial reasoning via indirect rewards. In *ICML 2026*, September 2025a. URL https://www.microsoft.com/en-us/research/publication/unlocking-zero-shot-geospatial-reasoning-via-indirect-rewards/.

Shixiong Xu, Chenghao Zhang, Lubin Fan, Gaofeng Meng, Shiming Xiang, and Jieping Ye. Addressclip: Empowering vision-language models for city-wide image address localization. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 76–92, Cham, 2025b. Springer Nature Switzerland.

Kaiwen Xue, Tao Wei, Guoxin Zhang, Zhonghong Ou, Kaoyan Lu, Yu Feng, Yifan Zhu, and Haoran Luo. ERGeoBench: A Comprehensive Benchmark for Embodied Reasoning and Geo-localization in Multimodal Large Language Models, 2026. URL https://arxiv.org/abs/2605.31251.

Chenjie Yang, Yutian Jiang, Yutong Deng, and Chenyu Wu. Skill-conditioned visual geolocation for vision-language models, 2026. URL https://arxiv.org/abs/2604.09025.

Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Qi Zhu, Conghui He, and Weijia Li. Where am i? cross-view geo-localization with natural language descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5890–5900, October 2025.

Sahiti Yerramilli, Nilay Pande, Rynaa Grover, and Jayant Sravan Tamarapalli. GeoChain: Multimodal chain-of-thought for geographic reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23624–23639, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1284. URL https://aclanthology.org/2025.findings-emnlp.1284/.

Qiang Yi and Lianlei Shan. GeoLocSFT: Efficient Visual Geolocation via Supervised Fine-Tuning of Multimodal Foundation Models, 2025.

Burak Yildiz, Seyran Khademi, Ronald Maria Siebes, and Jan Van Gemert. Amstertime: A visual place recognition benchmark dataset for severe domain shift. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 2749–2755, 2022. doi: 10.1109/ICPR56361.2022.9956049.

Bo Yu, Fengze Yang, Yiming Liu, Chao Wang, Xuewen Luo, Taozhe Li, Ruimin Ke, Xiaofan Zhou, and Chenxi Liu. Locatability-Guided Adaptive Reasoning for Image Geo-Localization with Vision-Language Models, 2026.

Hongyang Zhang, Yinhao Liu, and Zhenyu Kuang. Skylink: Unifying street-satellite geo-localization via uav-mediated 3d scene alignment. In Tingyu Wang, Yujiao Shi, Fabian Deuser, Shaofei Huang, Guosheng Hu, Si Liu, Zhedong Zheng, and Roger Zimmermann (eds.), *Proceedings of the 3rd International Workshop on UAVs in Multimedia: Capturing the World from a New Perspective, UAVM 2025, Dublin, Ireland, 31 October 2025*, pp. 26–30. ACM, 2025a. doi: 10.1145/3728482.3757392. URL https://doi.org/10.1145/3728482.3757392.

Zheyuan Zhang, Runze Li, Tasnim Kabir, and Jordan Boyd-Graber. Navig: Natural language-guided analysis with vision language models for image geo-localization, 2025b. URL https://arxiv.org/abs/2502.14638.

Zilun Zhang, Zian Guan, Tiancheng Zhao, Haozhan Shen, Tianyu Li, Yuxiang Cai, Zhonggen Su, Zhaojun Liu, Jianwei Yin, and Xiang Li. Geo-r1: Improving few-shot geospatial referring expression understanding with reinforcement fine-tuning. *arXiv preprint arXiv:2509.21976*, 2025c.

Yushuo Zheng, Huiyu Duan, Zicheng Zhang, Xiaohong Liu, and Xiongkuo Min. Learning to wander: Improving the global image geolocation ability of lmms via actionable reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7208–7219, June 2026a.

Yushuo Zheng, Jiangyong Ying, Huiyu Duan, Chunyi Li, Zicheng Zhang, Jing Liu, Xiaohong Liu, and Guangtao Zhai. Geox-bench: Benchmarking cross-view geo-localization and pose estimation capabilities of large multimodal models. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 13485–13493. AAAI Press, 2026b. doi: 10.1609/AAAI.V40I16.38353. URL https://doi.org/10.1609/aaai.v40i16.38353.

Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 1395–1403, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413896. URL https://doi.org/10.1145/3394171.3413896.

Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 10707–10715. AAAI Press, 2025. doi: 10.1609/AAAI.V39I10.33163. URL https://doi.org/10.1609/aaai.v39i10.33163.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40 (6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009. URL https://doi.org/10.1109/TPAMI.2017.2723009.

Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang (eds.), *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pp. 2749–2754. ACM, 2024. doi: 10.1145/3626772.3657673. URL https://doi.org/10.1145/3626772.3657673.

SUPPLEMENTARY MATERIALS OF TOWARDS DIVERSE GLOBAL IMAGE GEOLOCALIZATION: BASELINES, DATASET, AND BENCHMARK

In order to provide a more complete illustration of the capabilities of GeoVerse, a comprehensive appendix has been developed.

CONTENTS

A Detailed Comparison of Datasets and Benchmarks **22**

B Detailed Information About GeoVerse100K **22**

B.1 Data Source 22

B.2 Implementation Details of Geographic Attribute Sampling 22

B.3 Benchmark Construction 22

C More Information About Training and Evaluation **23**

C.1 Setting of Hyperparameter 23

C.2 Training and Evaluation Prompts 23

C.3 Detailed Results of Evaluation 23

C.4 Training with Another Base Model 23

D Case Study **23**

D.1 More Typical Cases 23

D.2 Failure Cases 24

E Discussion **24**

E.1 Limitations 24

E.2 Broader Impacts 24

A Detailed Comparison of Datasets and Benchmarks

As shown in Tab. 4, we systematically compare GeoVerse100K with existing image geolocalization datasets and benchmarks in terms of data scale, visual domain diversity, global coverage, geographic sampling strategies, and difficulty control. Most existing datasets focus on street-view imagery. Even those incorporating other visual sources typically cover only a limited range of scene types, making it difficult to comprehensively evaluate geolocalization across domains. In contrast, to our knowledge, GeoVerse100K is the first global image geolocalization dataset to systematically cover ten complementary visual domains under unified data and evaluation protocols. Existing data sources commonly exhibit geographic bias, and their sample difficulty distributions are often poorly suited to reinforcement learning post-training. To address these issues, we design sampling strategies that jointly account for geographic coverage and difficulty control. This yields a more balanced and challenging data resource for model training and comprehensive evaluation.

B Detailed Information About GeoVerse100K

B.1 Data Source

GeoVerse100K combines image datasets, remote-sensing products, and geospatial platforms to provide complementary visual representations across ten domains. The data sources and their candidate sample counts before sampling are summarized in Tab. 5.

B.2 Implementation Details of Geographic Attribute Sampling

The table details the geographic attribute sampling methods designed for different visual domains. Using these methods, we sample 25,000 examples from each continuous domain and 10,000 from each discrete domain to improve geographic coverage and distribution balance.

B.3 Benchmark Construction

GeoVerse-Val is independently sampled from data outside the training set and covers the same ten visual domains. Each of the four continuous domains contains 1,500 images, while each of the six discrete domains contains 500 images, yielding 9,000 images in total. Each sample is annotated with latitude and longitude, visual domain, subcategory, and continent.

As shown in Fig. 1, we further divide the ten visual domains into 37 subcategories to support finer-grained evaluation based on scene content, viewing perspective, geometric structure, and the level of geographic detail that can be inferred. Most subcategories are assigned according to the rules in Tab. 6. For some subcategories within the UAV, street-view, and space domains that cannot be reliably determined from existing metadata alone, we use Doubao-Seed-2.0-Pro for model-assisted annotation based on predefined classification criteria. The corresponding category definitions and prompt templates are shown in Fig. 8. We further manually review the model-generated category labels to reduce annotation noise and ensure the reliability of the evaluation data.

In particular, we use a unified geographically stratified procedure to select evaluation locations for the three continuous domains of remote sensing, terrain and plant, and maps. This procedure aims to prevent the benchmark from being overly concentrated around individual locations. Candidate centers are preferentially selected from land areas near annotated settlements or locations listed in gazetteers, excluding ocean grid cells and nearly uninhabited areas. Sampling quotas are allocated across three spatial scales. The fine scale covers $0.015^\circ \times 0.01^\circ$ (level 19), roughly corresponding to an urban district. The medium scale covers $0.35^\circ \times 0.25^\circ$ (level 15), roughly corresponding to a city. The coarse scale covers $3.5^\circ \times 2.0^\circ$ (level 11), roughly corresponding to a province-level region. Within each continent–spatial scale stratum, we perform greedy farthest-point sampling without replacement. At each step, we select the candidate location with the largest minimum distance to the set of previously selected locations.

C More Information About Training and Evaluation

C.1 Setting of Hyperparameter

GeoVerse is based on Qwen3.5-9B and trained with GRPO for one epoch on eight NVIDIA H20 GPUs. Evaluation is conducted on a single H20 GPU. The main hyperparameters are listed in Tab. 7.

C.2 Training and Evaluation Prompts

The task prompt used for training and evaluation is shown in Fig. 9. It covers all ten visual domains and instructs the model to first identify the image type, then extract geographic clues and infer the location from coarse to fine. The `Clues` field records key visual clues, `Reasoning` provides the reasoning process, and `FinalAnswer` returns the location in the format "Country; Region; Specific Location". The model does not directly output latitude and longitude. During evaluation, the final location is converted into coordinates using a geocoding tool before the localization error is calculated.

C.3 Detailed Results of Evaluation

Tables 8–17 report results for the nine models, overall and by continent, in each domain. Within each group, the best displayed nonzero result is **bold** and the second-best is underlined. Dashes (–) indicate values that are too low for meaningful comparison.

C.4 Training with Another Base Model

As shown in Tab. 18, we also train InternVL3.5-8B on GeoVerse100K using the same training strategy. The results show that the dataset and training strategy are also effective for this model. This indicates that their benefits are not limited to a single model and further supports the applicability of our approach across models.

D Case Study

D.1 More Typical Cases

Fig. 10 presents representative examples from continuous visual domains. In the Edinburgh street-view example, regionally distinctive stone buildings, road markings, and urban layout help the model locate the scene in Scotland, United Kingdom. In the remote-sensing example from Hwaseong, South Korea, the main cues are land-cover patterns, including solar panel arrays, regularly arranged fields, and woodland distribution. The terrain and vegetation example provides location cues through coastline shapes and island distribution along the Croatian coast, while the map of Beijing's Forbidden City shows a distinctive building layout and surrounding road structure. These examples demonstrate that geolocation evidence in continuous visual domains can come from several levels, including local natural and built features, large-scale landscape patterns, and map topology.

Fig. 11 presents representative examples from discrete visual domains. The distinctive geometry of Azadi Tower's vaulted structure and the clearly legible text "Bendigo Tramways" in the indoor scene provide direct location evidence through visual structure and textual meaning, respectively. In contrast, localization in the Chicago road-network, Massachusetts outline, and Seattle space examples relies mainly on global spatial structure. Key cues include road organization, the distinctive shape of Cape Cod, and the distribution of land and water, respectively. The UAV example from Phuket combines cues from the coastal environment, vegetation distribution, and the layout of resort buildings. These examples show that useful geographic cues vary substantially across visual domains. A unified image geolocalization model must therefore be able to extract, select, and reason about cues across domains.

D.2 Failure Cases

Figs. 12 and 13 further present several representative failure cases of GeoVerse. These cases show that when visible clues are sparse or similar clues occur across multiple geographic regions, the model may assign the target location to a visually similar but geographically distant region. For example, the model mistakes a street-view scene of dry hills in Madagascar for California, United States, and farmland and woodland in Southland, New Zealand, for North Yorkshire, United Kingdom. It also places the rice terraces of Batad, Philippines, in Ollantaytambo, Peru, and the dark beach of Kaikōura, New Zealand, in Reynisfjara, Iceland. Although these predictions share similarities with the true locations in landforms, land cover, or scene appearance, such clues often reveal only broad environmental characteristics, such as dry hills, temperate farmland, terraces, or dark coastlines. They are insufficient to uniquely identify a geographic location. The model therefore needs to combine complementary evidence from more distinctive local visual features, landmarks, text, and global spatial structure to resolve geographic ambiguity caused by visual similarities across regions.

E Discussion

E.1 Limitations

Although GeoVerse100K covers ten visual domains, its geographic distribution is still partly constrained by the availability of the original data sources. Geographic attribute sampling substantially improves spatial balance within the candidate pool, but coverage can still be expanded in regions where source data are scarce. Reward-aware sampling uses fixed reference models to provide a reproducible approximation of each sample's learning value. The reward and difficulty statistics in Figure 5 should therefore be interpreted in the context of the current reference models' capabilities. As model capabilities improve, sample difficulty can be recalibrated by updating the reference models. In addition, weighting the maximum reward among candidate responses prioritizes samples with clear learning signals, making the data distribution better suited to reinforcement learning post-training. As a result, some samples that are highly challenging for all reference models may receive relatively low sampling probabilities. Future versions could further improve geographic coverage and long-term evaluation value by expanding data sources, updating reference models, and allocating a separate quota for highly challenging samples.

E.2 Broader Impacts

GeoVerse100K provides a shared resource for studying geolocalization across diverse visual inputs. Improved generalization could support disaster assessment, environmental monitoring, and geographic image organization when reliable location metadata are unavailable. At the same time, inferring locations from images may reveal sensitive locations or enable the tracking of individuals without their consent. Model use should therefore comply with source licenses, respect individual consent, and protect sensitive location information.

Table 4: **Comparison of datasets and benchmarks. Global:** whether the dataset is a global geolocalization dataset. **Sampling:** whether a data sampling algorithm is used to reduce geographic bias. **Difficulty:** whether sample difficulty is filtered or controlled.

Name	Scale	Street	Other	Global	Sampling	Difficulty
Dataset						
IM2GPS (Hays & Efros, 2008)	6.5M	✓	×	✓	×	×
Pitts250k (Torii et al., 2013)	250K	✓	×	×	✓	×
Nordland (Sünderhauf et al., 2013)	143K	✓	×	×	×	×
Tokyo 24/7 (Torii et al., 2015)	76K	✓	×	×	✓	×
YFCC100M (Thomee et al., 2016)	48M	×	×	✓	×	×
MP-16 (Choi et al., 2016)	4.7M	✓	×	✓	✓	×
SPED (Chen et al., 2017)	2.5M	✓	×	✓	×	×
Hotels-50K (Stylianou et al., 2019)	1M	×	Indoor	✓	✓	×
ADVANCE (Hu et al., 2020)	5.1K	×	Remote Sensing	✓	×	×
MSLS (Warburg et al., 2020)	1.6M	✓	×	✓	✓	×
GLDv2 (Weyand et al., 2020)	5M	✓	Landmark	✓	×	×
Geo-YFCC (Dubey et al., 2021)	1.15M	✓	×	✓	✓	×
GAMa (Vyas et al., 2022)	1.9M	✓	Remote Sensing	×	✓	×
SF-XL (Berton et al., 2022)	41M	✓	×	×	✓	×
GSV-Cities (Ali-bey et al., 2022)	530K	✓	×	✓	✓	×
ALTO (Cisneros et al., 2022)	410km	×	UAV	×	×	×
AmsterTime (Yildiz et al., 2022)	2.5K	✓	×	×	×	×
YFCC-Geo4.2M (Habedi, 2023)	4.2M	✓	×	✓	×	×
GeoGlobe (Han et al., 2025)	293K	✓	Landmark	✓	×	×
MP16-Pro (Jia et al., 2024)	4.72M	×	Remote Sensing	✓	×	×
OSV-5M (Astruc et al., 2024)	5M	✓	×	✓	✓	×
SF-IAL-Base (Xu et al., 2025b)	184K	✓	×	×	✓	×
MG-Geo (Dou et al., 2026)	5M	✓	×	✓	✓	×
Global Streetscapes (Hou et al., 2024)	10M	✓	×	✓	✓	×
GeoComp (Song et al., 2025)	2.7M	✓	×	✓	✓	✓
GAEA-1.6M (Campos et al., 2026)	800K	✓	×	✓	✓	×
GeoReasoner-70K (Li et al., 2024)	70K	✓	×	✓	✓	×
Clue2Geo-70K (Ma et al., 2025)	70K	✓	×	✓	✓	×
SkyLume (Li et al., 2026d)	100K	×	UAV	×	✓	×
MR600K (Yi & Shan, 2025)	600K	✓	×	✓	✓	×
GRE30K (Wang et al., 2025a)	30K	✓	×	✓	✓	✓
MP16-Reason (Li et al., 2025a)	33K	✓	×	✓	×	×
NaviClues (Zhang et al., 2025b)	2K	✓	×	✓	×	×
INDOOR-3.6M (Onuoha et al., 2026)	3.6M	×	Indoor	✓	✓	×
MMS-VPR (Ou et al., 2025)	78.6K	✓	×	×	✓	×
Flickr-Geo (do-me, 2025)	217.6M	✓	×	✓	×	×
SpotAgenticCoT (Jia et al., 2026a)	6K	✓	×	✓	✓	✓
Geo-R (Wu et al., 2026)	700K	✓	×	✓	✓	✓
SSL4EO-S12 (Blumenstiel et al., 2025)	1M	×	Remote Sensing	✓	✓	×
GeoSeek (Jin et al., 2026)	30K	✓	×	✓	✓	✓
GeoVerse-Train	90K	✓	9 Domains	✓	✓	✓
Benchmark						
IM2GPS3K (Vo et al., 2017)	3K	✓	×	✓	×	×
YFCC4K (Vo et al., 2017)	4K	✓	×	✓	×	×
InLoc (Taira et al., 2018)	10K	×	Indoor	×	×	×
YFCC26K (Theiner et al., 2022)	25.6K	✓	×	✓	×	×
GWS-15K (Clark et al., 2023)	15K	✓	×	✓	✓	×
LLMGeo (Wang et al., 2024)	10K	✓	×	✓	✓	×
UrBench (Zhou et al., 2025)	11.6K	✓	Remote Sensing/UAV	×	✓	×
MM-GAG Sarkar et al. (2024)	26.6K	×	Remote Sensing	×	✓	×
GPTGeoChat (Mendes et al., 2024)	1K	✓	×	✓	×	×
ISS-Geo142 (Srivastava et al., 2025)	142	×	Space	✓	✓	×

Continued on next page

Table 4 continued

Name	Scale	Street	Other	Global	Sampling	Difficulty
GeoChain (Yerramilli et al., 2025)	1.5M	✓	×	✓	✓	×
GAEA-Bench (Campos et al., 2026)	4K	✓	×	✓	✓	×
IMAGEO-Bench (Li et al., 2025b)	9K	✓	×	✓	✓	×
MR40K (Yi & Shan, 2025)	40K	✓	×	✓	✓	×
MapBench (Xing et al., 2025)	1.5K	×	Map	×	✓	×
FairLocator (Huang et al., 2025)	1.2K	✓	×	✓	✓	×
DoxBench (Luo et al., 2026)	500	✓	Indoor	×	×	×
KoreaGEO Bench (Wang et al., 2025c)	1.1K	✓	×	×	✓	×
GEOBench-VLM (Danish et al., 2025)	10K	×	Remote Sensing	✓	×	×
GeoExplain (Cheng et al., 2026)	5.5K	✓	×	✓	×	×
WhereBench (Qian et al., 2025)	810	✓	×	✓	✓	×
CCL-Bench (Li et al., 2026b)	300	✓	×	×	✓	✓
GeoRC (Talreja et al., 2026)	800	✓	×	✓	✓	×
GTPred (Li et al., 2026a)	370	✓	×	✓	✓	×
GeoArena (Jia et al., 2026c)	1K	✓	×	✓	×	×
GeoBrowse (Geng et al., 2026)	300	✓	×	✓	✓	✓
GeoSeek-Val (Jin et al., 2026)	3K	✓	×	✓	✓	✓
WanderBench (Zheng et al., 2026a)	32K	✓	×	✓	✓	✓
LaVPR (Idan et al., 2026)	650K	✓	×	✓	×	×
SIGNPOST-Bench (Li et al., 2026c)	25.5K	✓	×	✓	×	×
ERGeoBench (Xue et al., 2026)	2.2K	✓	×	✓	×	×
GeoVerse-Val	9K	✓	9 Domains	✓	✓	✓

Table 5: **Data sources and pre-sampling sample counts for GeoVerse100K.**

Domain	Data Source	Samples	Total
Street-View	*Global Streetscapes* (Hou et al., 2024)	5M	8M
	OSV-5M (Astruc et al., 2024)	3M	
Remote Sensing	*Shuijing Weitu Global Geolocation dataset*	78K	78K
Terrain and Plant	*Sentinel2-10m-LandCover*	38K	78K
	MapBiomas	40K	
Map	*OpenStreetMap*	77K	77K
Landmark	*Google Landmarks Dataset v2 (GLDv2)* (Weyand et al., 2020)	5M	5M
Road-network	*OpenStreetMap*	100K	100K
Indoor	*Wikidata*	25K	28K
	Hotels-50K	3K	
Shape	*GADM*	200K	200K
Space	*NASA Gateway to Astronaut Photography of Earth*	500K	500K
UAV	*GeoNadir*	6K	21K
	OpenAerialMap (OAM)	15K	

Table 6: **Implementation Details of Geographic Attribute Sampling**

Continuous Domains	
Street-View	The candidate pool combines Global Streetscapes (Hou et al., 2024) and OSV-5M (Astruc et al., 2024). After removing unavailable or corrupted images and samples with invalid coordinates, candidates from the two sources are merged without imposing a fixed source ratio. We apply the country-level quota in Eq. (1) and the population-aware $1^\circ \times 1^\circ$ grid sampling in Eq. (2) to select 25,000 anchor locations. The selected coordinates are shared by the street-view, remote-sensing, terrain-and-plant, and map domains.
Remote Sensing & Terrain and Plant & Map	We first remove samples that are inaccessible, corrupted, or associated with invalid geographic coordinates, and then deduplicate the remaining samples by coordinates. Next, we use the sampled street-view locations as geographic anchors, match candidate samples to these anchors, and sample without replacement. The map and remote-sensing domains share the same geographic anchors. For each remote-sensing anchor, we define three bounding boxes with different spatial extents and generate the corresponding map images using the same minimum longitude, minimum latitude, maximum longitude, and maximum latitude. Each corresponding pair of map and remote-sensing images therefore has the same center coordinates and geographic extent, differing only in visual modality. The three spatial levels associated with each anchor are retained as a candidate group for subsequent level-aware sampling.
Discrete Domains	
Landmark	We retain GLDv2 images with valid geographic coordinates and apply the same two-stage geographic sampling strategy as in the street-view domain. Specifically, country or region quotas are computed using Eq. (1), and candidates within each country are sampled from $1^\circ \times 1^\circ$ cells according to Eq. (2). Sampling is performed without replacement over the landmark candidate pool until 10,000 images are selected for subsequent reward-aware sampling.
Road-network	The candidate pool is first geographically clustered using DBSCAN with a 2-km radius based on the Haversine distance, retaining representative locations from locally dense regions. Road-network images are then rendered at three zoom levels for each location. For candidate image i, the visual quality score is defined as $$Q_i = 0.28\,\mathrm{RZ}(b_i) + 0.17\,\mathrm{RZ}(b_i^s) + 0.22\,\mathrm{RZ}(\log v_i) + 0.14\,\mathrm{RZ}(\sigma_i) + 0.12\,\mathrm{RZ}(\log k_i) + 0.05\,\mathrm{RZ}(e_i) + 0.02\,\mathrm{RZ}(\log s_i), \quad (7)$$ where RZ denotes robust z-score normalization, and b_i, b_i^s, v_i, σ_i, k_i, e_i, and s_i denote the non-background ratio, strong non-background ratio, Laplacian variance, grayscale standard deviation, keypoint count, image entropy, and file size, respectively. The quality of location p is represented by its highest score across the three zoom levels. Greedy sampling is then performed according to $$\mathrm{Priority}(p) = \min_{s \in \mathcal{S}} d_{\mathrm{hav}}(p, s) + 6R(p), \quad (8)$$ where $\mathcal{S}$ is the set of selected locations and $R(p)$ is the relative quality rank. This strategy prioritizes visually informative locations that are geographically distant from the selected set while retaining only one zoom level per location.
Indoor	Places365 (Zhou et al., 2018) is first used to assign fine-grained labels from 365 scene categories, and outdoor images are removed using a pretrained IndoorOutdoorNet. The remaining scenes are then consolidated into six functional categories. Then, approximately balanced dynamic quotas are assigned across categories: all samples from underrepresented categories are retained, and their unused quotas are redistributed among data-rich categories. Meanwhile, the globe is divided into $1^\circ \times 1^\circ$ latitude–longitude grids, and round-robin sampling across grid cells is applied to reduce the overrepresentation of geographically dense regions.

Continued on next page

Table 6 continued

Shape	Administrative boundaries at levels 0–3 worldwide are first obtained from the data sources, and each Polygon/MultiPolygon is proportionally rendered as a black silhouette with fixed padding. Exact duplicates, duplicate enclave-like entries, and naming conflicts are then removed. For each geometry G, let C and H denote the numbers of connected components and holes, S and P its total area and perimeter, $S_{\max}$ the area of its largest component, and w and h the width and height of its bounding box. The geometric descriptors are $$a = \frac{\max(w,h)}{\min(w,h)}, \quad r = \frac{P}{2\sqrt{\pi S}}, f \quad = 1 - \frac{S_{\max}}{S}, q \quad = \frac{S}{wh}. \tag{9}$$ Samples are classified in the following priority order: $$\mathrm{type}(G) = \begin{cases} \text{Archipelago}, & C \geq 4 \vee (C \geq 2 \wedge f \geq 0.25), \\ \text{Elongated}, & a \geq 3, \\ \text{Porous}, & H \geq 1 \vee r \geq 2.2, \\ \text{Compact}, & \text{otherwise}. \end{cases} \tag{10}$$ Finally, stratified quota sampling is performed across administrative levels and geographic regions according to these four shape classes, increasing shape diversity while reducing the dominance of common silhouettes.
Space	We collect daylight Earth-facing photographs from the NASA Gateway to Astronaut Photography of Earth and retain records with an available image and a valid cataloged center-point latitude and longitude. The cataloged image center is used to assign each candidate to a country or region and a $1^\circ \times 1^\circ$ geographic cell. Because the geographic coverage of astronaut photography is constrained and uneven, sampling is restricted to non-empty geographic strata represented in the candidate pool and is performed without replacement. Unavailable quotas are redistributed among strata with remaining candidates until 10,000 images are selected for subsequent reward-aware sampling.
UAV	Images with valid GPS coordinates are first aggregated from the data sources and geographically clustered using DBSCAN with a 2.8-km radius based on the Haversine distance. Clusters overlapping with the evaluation set are removed. After within-cluster deduplication, up to three images separated by more than 1 km are selected from each cluster. Balanced sampling is then performed over $2^\circ \times 2^\circ$ geographic grids, requiring each additional sample to be more than 0.2 km from previously selected samples and limiting each cluster to eight images, resulting in 10,000 candidate training images.

Table 7: **The values of each hyperparameter.**

Symbol	Description	Value
	Batching and GRPO Sampling	
B_{dev}	Per-device training batch size, measured in completions	24
K_{grad}	Number of gradient-accumulation steps	1
G	Number of candidate responses $\{y_j\}_{j=1}^{G}$ sampled for each image x	8
S	Number of policy-training steps per generation batch	1
	Reward-Aware Sampling	
α	Weight exponent for the reward standard-deviation score s_i^{std}	1.0
β	Weight exponent for the maximum-reward score $s_i^{\max}$	1.0
	Optimization	
η	Peak learning rate	1×10^{-6}
r_{warm}	Fraction of optimization steps used for warmup	0.05
c_{grad}	Maximum gradient norm	2.0
β_{KL}	KL-divergence penalty coefficient	0.04
δ_-	Lower clipping range in the PPO-like GRPO objective	0.20
δ_+	Upper clipping range in the PPO-like GRPO objective	0.28
	Sequence and Image Processing	
L	Maximum total sequence length	8,192
L_y	Maximum length of a sampled response y_j	1,024
τ_T	Sampling temperature used to generate y_j	0.9

Continued on the next page

Table 7 continued

Symbol	Description	Value
	Reward Task	
λ_{geo}	Weight of distance-aware reward	2.0
λ_{fmt}	Weight of format reward	1.0
$D_{\max}$	Maximum distance used by GeoScore normalization	18,050

Table 8: **Detailed results on Street-View.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=1,500)				
GLOBE (Li et al., 2025a)	**2904.93**	**78.07**	50.27	**11.87**
GeoClip (Vivanco Cepeda et al., 2023)	2771.03	75.67	46.40	8.87
G3 (Jia et al., 2024)	2714.02	74.53	45.07	7.33
Geo-R1 (Zhang et al., 2025c)	2597.03	69.13	44.13	9.20
Qwen2.5-VL-7B (Bai et al., 2025)	2034.21	55.53	31.67	8.40
Qwen3.5-9B (Qwen Team, 2026)	1998.85	54.20	31.53	7.80
InternVL3.5-8B (Wang et al., 2025b)	1707.97	51.47	21.67	4.00
Gemma4-26B-A4B (Gemma Team et al., 2026)	1482.73	40.87	21.40	5.40
GeoVerse (Ours)	2894.46	76.40	**51.87**	10.80
Africa (n=284)				
GLOBE (Li et al., 2025a)	2290.73	65.14	**33.80**	10.56
GeoClip (Vivanco Cepeda et al., 2023)	**2366.42**	**69.37**	32.04	11.62
G3 (Jia et al., 2024)	2253.93	64.79	30.99	9.51
Geo-R1 (Zhang et al., 2025c)	1843.97	50.70	26.06	8.45
Qwen2.5-VL-7B (Bai et al., 2025)	1602.26	45.77	20.07	6.34
Qwen3.5-9B (Qwen Team, 2026)	1350.16	38.03	16.90	4.58
InternVL3.5-8B (Wang et al., 2025b)	1072.68	30.99	12.32	2.46
Gemma4-26B-A4B (Gemma Team et al., 2026)	1029.40	30.63	11.27	6.69
GeoVerse (Ours)	2198.30	61.27	32.04	**11.97**
Asia (n=380)				
GLOBE (Li et al., 2025a)	3124.63	80.79	**57.63**	17.63
GeoClip (Vivanco Cepeda et al., 2023)	2852.30	78.42	45.79	12.63
G3 (Jia et al., 2024)	2766.67	78.68	42.63	8.95
Geo-R1 (Zhang et al., 2025c)	2868.16	77.63	48.95	13.68
Qwen2.5-VL-7B (Bai et al., 2025)	2293.02	59.21	39.74	13.42
Qwen3.5-9B (Qwen Team, 2026)	2291.54	59.74	37.63	13.16
InternVL3.5-8B (Wang et al., 2025b)	1846.34	55.00	23.68	8.16
Gemma4-26B-A4B (Gemma Team et al., 2026)	1844.82	50.00	28.42	8.16
GeoVerse (Ours)	**3196.68**	**83.68**	55.79	**18.95**
Europe (n=327)				
GLOBE (Li et al., 2025a)	2798.72	**78.90**	45.26	**8.56**
GeoClip (Vivanco Cepeda et al., 2023)	2466.11	70.03	41.28	3.36
G3 (Jia et al., 2024)	2476.92	70.03	40.37	2.75
Geo-R1 (Zhang et al., 2025c)	2546.60	68.50	40.98	7.65
Qwen2.5-VL-7B (Bai et al., 2025)	2117.38	61.77	27.83	7.65
Qwen3.5-9B (Qwen Team, 2026)	2026.08	56.27	29.36	7.03
InternVL3.5-8B (Wang et al., 2025b)	1817.63	55.35	22.32	3.36
Gemma4-26B-A4B (Gemma Team et al., 2026)	1237.41	33.03	18.35	3.06
GeoVerse (Ours)	**2824.85**	76.45	**48.01**	6.42

Continued on next page

Table 8 continued: Street-View

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
North America (n=271)				
GLOBE (Li et al., 2025a)	3592.25	90.77	69.00	7.38
GeoClip (Vivanco Cepeda et al., 2023)	**3664.23**	**91.14**	73.06	7.38
G3 (Jia et al., 2024)	3633.30	**91.14**	**74.17**	**7.75**
Geo-R1 (Zhang et al., 2025c)	3422.26	85.24	66.79	6.27
Qwen2.5-VL-7B (Bai et al., 2025)	2349.48	61.99	40.59	5.17
Qwen3.5-9B (Qwen Team, 2026)	2411.97	67.16	42.44	3.69
InternVL3.5-8B (Wang et al., 2025b)	2554.02	81.18	33.21	1.48
Gemma4-26B-A4B (Gemma Team et al., 2026)	1768.48	49.08	26.20	3.32
GeoVerse (Ours)	3500.59	88.93	72.32	4.43
Oceania (n=70)				
GLOBE (Li et al., 2025a)	2930.88	**78.57**	47.14	**20.00**
GeoClip (Vivanco Cepeda et al., 2023)	2854.91	74.29	47.14	14.29
G3 (Jia et al., 2024)	2911.32	77.14	45.71	15.71
Geo-R1 (Zhang et al., 2025c)	2487.45	71.43	42.86	12.86
Qwen2.5-VL-7B (Bai et al., 2025)	2057.52	57.14	37.14	11.43
Qwen3.5-9B (Qwen Team, 2026)	2691.91	74.29	44.29	8.57
InternVL3.5-8B (Wang et al., 2025b)	1678.60	44.29	28.57	2.86
Gemma4-26B-A4B (Gemma Team et al., 2026)	1630.81	42.86	27.14	2.86
GeoVerse (Ours)	**3108.53**	77.14	**64.29**	14.29
South America (n=168)				
GLOBE (Li et al., 2025a)	**2533.51**	**71.43**	42.26	**11.31**
GeoClip (Vivanco Cepeda et al., 2023)	2388.97	66.67	38.69	6.55
G3 (Jia et al., 2024)	2269.08	62.50	36.31	4.76
Geo-R1 (Zhang et al., 2025c)	2069.47	55.36	33.93	6.55
Qwen2.5-VL-7B (Bai et al., 2025)	1498.84	40.48	23.81	5.95
Qwen3.5-9B (Qwen Team, 2026)	1425.24	35.71	23.81	8.93
InternVL3.5-8B (Wang et al., 2025b)	902.91	25.60	10.12	2.98
Gemma4-26B-A4B (Gemma Team et al., 2026)	1384.91	38.69	18.45	5.95
GeoVerse (Ours)	2456.25	64.88	**45.83**	7.74

Table 9: **Detailed results on Remote Sensing.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=1,500)				
GLOBE (Li et al., 2025a)	1146.75	34.53	13.67	–
GeoClip (Vivanco Cepeda et al., 2023)	1267.98	37.67	15.27	–
G3 (Jia et al., 2024)	1165.61	35.20	13.40	–
Geo-R1 (Zhang et al., 2025c)	963.98	29.27	10.53	–
Qwen2.5-VL-7B (Bai et al., 2025)	996.72	29.93	10.87	–
Qwen3.5-9B (Qwen Team, 2026)	889.45	26.00	8.93	–
InternVL3.5-8B (Wang et al., 2025b)	649.05	20.00	5.33	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	547.52	16.80	5.53	–
GeoVerse (Ours)	**1286.23**	**38.33**	**15.67**	–
Africa (n=243)				
GLOBE (Li et al., 2025a)	**1016.76**	**32.10**	7.00	–
GeoClip (Vivanco Cepeda et al., 2023)	901.24	27.57	**7.41**	–
G3 (Jia et al., 2024)	799.59	25.10	5.35	–
Geo-R1 (Zhang et al., 2025c)	702.71	22.63	4.12	–
Qwen2.5-VL-7B (Bai et al., 2025)	918.91	28.81	6.17	–
Qwen3.5-9B (Qwen Team, 2026)	715.59	20.16	4.94	–
InternVL3.5-8B (Wang et al., 2025b)	473.59	13.58	2.47	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	311.90	8.64	2.06	–
GeoVerse (Ours)	916.62	25.51	6.58	–

Continued on next page

Table 9 continued: Remote Sensing

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Asia (n=353)				
GLOBE (Li et al., 2025a)	1265.15	38.53	**16.15**	–
GeoClip (Vivanco Cepeda et al., 2023)	1014.87	30.03	9.35	–
G3 (Jia et al., 2024)	1023.88	31.73	8.50	–
Geo-R1 (Zhang et al., 2025c)	1102.03	32.29	10.48	–
Qwen2.5-VL-7B (Bai et al., 2025)	906.89	27.76	8.78	–
Qwen3.5-9B (Qwen Team, 2026)	925.29	25.78	10.76	–
InternVL3.5-8B (Wang et al., 2025b)	732.05	25.21	4.82	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	610.07	18.13	6.52	–
GeoVerse (Ours)	**1354.49**	**40.23**	15.86	–
Europe (n=321)				
GLOBE (Li et al., 2025a)	1023.97	27.41	15.89	–
GeoClip (Vivanco Cepeda et al., 2023)	**1482.36**	**42.06**	19.31	–
G3 (Jia et al., 2024)	1273.35	36.76	17.13	–
Geo-R1 (Zhang et al., 2025c)	985.13	28.97	13.71	–
Qwen2.5-VL-7B (Bai et al., 2025)	953.53	26.79	14.02	–
Qwen3.5-9B (Qwen Team, 2026)	1013.19	30.53	10.59	–
InternVL3.5-8B (Wang et al., 2025b)	654.34	19.00	6.23	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	464.63	13.08	5.92	–
GeoVerse (Ours)	1335.32	38.63	**19.94**	–
North America (n=313)				
GLOBE (Li et al., 2025a)	1617.30	48.88	20.45	–
GeoClip (Vivanco Cepeda et al., 2023)	1839.41	52.40	**30.35**	–
G3 (Jia et al., 2024)	**1848.53**	54.31	26.52	–
Geo-R1 (Zhang et al., 2025c)	1439.21	43.77	17.89	–
Qwen2.5-VL-7B (Bai et al., 2025)	1659.70	50.80	19.49	–
Qwen3.5-9B (Qwen Team, 2026)	1159.45	33.55	12.78	–
InternVL3.5-8B (Wang et al., 2025b)	988.08	29.71	9.58	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	908.35	30.03	8.63	–
GeoVerse (Ours)	1809.79	**55.59**	23.64	–
Oceania (n=79)				
GLOBE (Li et al., 2025a)	778.88	24.05	**7.59**	–
GeoClip (Vivanco Cepeda et al., 2023)	800.05	25.32	6.33	–
G3 (Jia et al., 2024)	**860.80**	**27.85**	6.33	–
Geo-R1 (Zhang et al., 2025c)	714.65	26.58	6.33	–
Qwen2.5-VL-7B (Bai et al., 2025)	508.09	13.92	3.80	–
Qwen3.5-9B (Qwen Team, 2026)	589.56	18.99	3.80	–
InternVL3.5-8B (Wang et al., 2025b)	143.76	2.53	1.27	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	318.34	10.13	1.27	–
GeoVerse (Ours)	678.67	21.52	6.33	–
South America (n=191)				
GLOBE (Li et al., 2025a)	680.68	23.04	5.24	–
GeoClip (Vivanco Cepeda et al., 2023)	**1099.19**	**38.22**	8.38	–
G3 (Jia et al., 2024)	719.11	23.56	7.85	–
Geo-R1 (Zhang et al., 2025c)	330.02	9.95	3.14	–
Qwen2.5-VL-7B (Bai et al., 2025)	449.99	13.09	4.19	–
Qwen3.5-9B (Qwen Team, 2026)	517.96	16.75	3.66	–
InternVL3.5-8B (Wang et al., 2025b)	363.43	11.52	3.14	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	374.49	12.04	4.19	–
GeoVerse (Ours)	941.10	29.32	**10.47**	–

Table 10: **Detailed results on Terrain and Plant.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=1,500)				
GLOBE (Li et al., 2025a)	483.89	14.47	3.00	–
GeoClip (Vivanco Cepeda et al., 2023)	476.50	15.47	3.47	–
G3 (Jia et al., 2024)	424.80	12.73	3.73	–
Geo-R1 (Zhang et al., 2025c)	438.76	13.67	2.67	–
Qwen2.5-VL-7B (Bai et al., 2025)	532.26	16.47	3.67	–
Qwen3.5-9B (Qwen Team, 2026)	471.84	14.73	3.73	–
InternVL3.5-8B (Wang et al., 2025b)	429.73	12.53	2.53	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	245.09	7.60	0.93	–
GeoVerse (Ours)	**599.70**	**18.27**	**4.80**	–
Africa (n=237)				
GLOBE (Li et al., 2025a)	188.16	**3.80**	**0.84**	–
GeoClip (Vivanco Cepeda et al., 2023)	134.37	2.11	–	–
G3 (Jia et al., 2024)	109.13	2.11	–	–
Geo-R1 (Zhang et al., 2025c)	119.21	1.69	0.42	–
Qwen2.5-VL-7B (Bai et al., 2025)	**194.96**	2.95	–	–
Qwen3.5-9B (Qwen Team, 2026)	143.83	2.53	**0.84**	–
InternVL3.5-8B (Wang et al., 2025b)	127.46	1.27	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	52.21	0.84	–	–
GeoVerse (Ours)	151.40	3.38	0.42	–
Asia (n=351)				
GLOBE (Li et al., 2025a)	207.59	2.28	1.14	–
GeoClip (Vivanco Cepeda et al., 2023)	290.33	7.69	0.85	–
G3 (Jia et al., 2024)	392.47	8.83	3.70	–
Geo-R1 (Zhang et al., 2025c)	195.02	2.85	0.57	–
Qwen2.5-VL-7B (Bai et al., 2025)	246.41	4.84	0.57	–
Qwen3.5-9B (Qwen Team, 2026)	412.29	10.83	1.99	–
InternVL3.5-8B (Wang et al., 2025b)	417.92	11.11	0.57	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	244.96	4.56	0.28	–
GeoVerse (Ours)	**842.57**	**22.79**	**5.41**	–
Europe (n=315)				
GLOBE (Li et al., 2025a)	**1594.19**	53.97	11.11	–
GeoClip (Vivanco Cepeda et al., 2023)	828.81	26.98	7.62	–
G3 (Jia et al., 2024)	640.94	19.37	6.67	–
Geo-R1 (Zhang et al., 2025c)	1326.42	46.67	10.16	–
Qwen2.5-VL-7B (Bai et al., 2025)	1571.01	**54.60**	**15.24**	–
Qwen3.5-9B (Qwen Team, 2026)	711.35	21.27	8.57	–
InternVL3.5-8B (Wang et al., 2025b)	1207.89	40.32	10.79	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	251.39	7.62	2.86	–
GeoVerse (Ours)	931.19	28.89	10.48	–
North America (n=297)				
GLOBE (Li et al., 2025a)	284.39	8.42	1.35	–
GeoClip (Vivanco Cepeda et al., 2023)	553.63	18.18	5.39	–
G3 (Jia et al., 2024)	701.34	25.59	6.06	–
Geo-R1 (Zhang et al., 2025c)	428.09	14.48	1.68	–
Qwen2.5-VL-7B (Bai et al., 2025)	478.01	15.82	1.68	–
Qwen3.5-9B (Qwen Team, 2026)	**936.34**	**36.36**	**6.73**	–
InternVL3.5-8B (Wang et al., 2025b)	206.43	5.05	0.67	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	549.19	22.22	1.35	–
GeoVerse (Ours)	842.88	31.65	6.40	–

Continued on next page

Table 10 continued: Terrain and Plant

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Oceania (n=111)				
GLOBE (Li et al., 2025a)	67.01	–	–	–
GeoClip (Vivanco Cepeda et al., 2023)	108.38	**2.70**	**1.80**	–
G3 (Jia et al., 2024)	**181.66**	**2.70**	**1.80**	–
Geo-R1 (Zhang et al., 2025c)	115.36	0.90	–	–
Qwen2.5-VL-7B (Bai et al., 2025)	112.83	0.90	–	–
Qwen3.5-9B (Qwen Team, 2026)	112.10	0.90	–	–
InternVL3.5-8B (Wang et al., 2025b)	63.95	–	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	132.13	1.80	–	–
GeoVerse (Ours)	121.66	–	–	–
South America (n=189)				
GLOBE (Li et al., 2025a)	75.66	2.65	–	–
GeoClip (Vivanco Cepeda et al., 2023)	**759.07**	**30.69**	**3.70**	–
G3 (Jia et al., 2024)	228.67	7.94	1.06	–
Geo-R1 (Zhang et al., 2025c)	19.43	–	–	–
Qwen2.5-VL-7B (Bai et al., 2025)	86.41	1.59	–	–
Qwen3.5-9B (Qwen Team, 2026)	75.95	0.53	–	–
InternVL3.5-8B (Wang et al., 2025b)	99.48	2.12	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	65.15	2.12	–	–
GeoVerse (Ours)	56.91	0.53	–	–

Table 11: **Detailed results on Map.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=1,500)				
GLOBE (Li et al., 2025a)	883.20	27.67	8.67	1.67
GeoClip (Vivanco Cepeda et al., 2023)	1165.42	32.80	14.87	2.73
G3 (Jia et al., 2024)	**1247.49**	**35.33**	**17.47**	**3.40**
Geo-R1 (Zhang et al., 2025c)	849.52	26.00	9.27	1.87
Qwen2.5-VL-7B (Bai et al., 2025)	879.97	28.13	8.33	1.67
Qwen3.5-9B (Qwen Team, 2026)	717.11	21.20	8.27	1.13
InternVL3.5-8B (Wang et al., 2025b)	478.71	14.00	3.67	0.07
Gemma4-26B-A4B (Gemma Team et al., 2026)	393.35	11.40	4.47	1.20
GeoVerse (Ours)	966.29	28.80	12.13	1.73
Africa (n=271)				
GLOBE (Li et al., 2025a)	261.80	7.01	0.37	–
GeoClip (Vivanco Cepeda et al., 2023)	453.76	11.07	3.32	**2.21**
G3 (Jia et al., 2024)	**517.27**	**13.65**	**6.27**	**2.21**
Geo-R1 (Zhang et al., 2025c)	252.13	7.01	0.74	–
Qwen2.5-VL-7B (Bai et al., 2025)	270.82	9.59	–	–
Qwen3.5-9B (Qwen Team, 2026)	196.79	4.80	0.74	–
InternVL3.5-8B (Wang et al., 2025b)	106.62	2.21	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	93.59	3.69	–	–
GeoVerse (Ours)	368.29	9.96	2.58	–
Asia (n=400)				
GLOBE (Li et al., 2025a)	762.92	20.75	8.75	2.75
GeoClip (Vivanco Cepeda et al., 2023)	808.53	24.00	6.75	1.25
G3 (Jia et al., 2024)	**975.14**	**29.75**	**10.25**	1.50
Geo-R1 (Zhang et al., 2025c)	805.19	24.00	7.00	3.75
Qwen2.5-VL-7B (Bai et al., 2025)	585.32	15.25	7.00	3.25
Qwen3.5-9B (Qwen Team, 2026)	617.18	15.00	8.25	3.00
InternVL3.5-8B (Wang et al., 2025b)	255.96	5.75	1.00	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	306.73	7.75	3.25	2.25
GeoVerse (Ours)	695.37	20.50	7.75	**4.50**

Continued on next page

Table 11 continued: Map

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Europe (n=325)				
GLOBE (Li et al., 2025a)	1342.50	**48.00**	8.62	–
GeoClip (Vivanco Cepeda et al., 2023)	1451.28	44.31	15.38	**3.38**
G3 (Jia et al., 2024)	**1599.52**	47.38	**20.31**	**3.38**
Geo-R1 (Zhang et al., 2025c)	1281.07	40.62	14.15	–
Qwen2.5-VL-7B (Bai et al., 2025)	1293.50	44.62	9.23	0.62
Qwen3.5-9B (Qwen Team, 2026)	648.20	19.69	6.77	0.31
InternVL3.5-8B (Wang et al., 2025b)	418.26	12.31	2.15	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	352.59	10.15	4.92	0.31
GeoVerse (Ours)	1070.30	32.62	13.54	0.31
North America (n=286)				
GLOBE (Li et al., 2025a)	1754.28	54.20	22.73	4.90
GeoClip (Vivanco Cepeda et al., 2023)	**2339.83**	63.99	40.56	4.20
G3 (Jia et al., 2024)	2314.86	63.29	**40.91**	**5.24**
Geo-R1 (Zhang et al., 2025c)	1603.10	49.65	21.68	4.55
Qwen2.5-VL-7B (Bai et al., 2025)	2003.00	65.73	23.08	3.50
Qwen3.5-9B (Qwen Team, 2026)	1915.26	62.59	23.43	1.40
InternVL3.5-8B (Wang et al., 2025b)	1496.23	48.25	15.03	0.35
Gemma4-26B-A4B (Gemma Team et al., 2026)	1106.58	33.57	12.94	2.80
GeoVerse (Ours)	2283.61	**70.63**	31.12	2.10
Oceania (n=64)				
GLOBE (Li et al., 2025a)	76.86	1.56	1.56	–
GeoClip (Vivanco Cepeda et al., 2023)	1475.61	31.25	**25.00**	10.94
G3 (Jia et al., 2024)	**1673.24**	**39.06**	21.88	**15.63**
Geo-R1 (Zhang et al., 2025c)	26.94	–	–	–
Qwen2.5-VL-7B (Bai et al., 2025)	91.21	1.56	1.56	–
Qwen3.5-9B (Qwen Team, 2026)	64.52	–	–	–
InternVL3.5-8B (Wang et al., 2025b)	64.15	1.56	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	78.85	1.56	1.56	–
GeoVerse (Ours)	142.49	4.69	1.56	–
South America (n=154)				
GLOBE (Li et al., 2025a)	37.18	0.65	–	–
GeoClip (Vivanco Cepeda et al., 2023)	**431.45**	**12.34**	3.25	–
G3 (Jia et al., 2024)	337.79	9.09	4.55	**1.95**
Geo-R1 (Zhang et al., 2025c)	47.56	0.65	0.65	–
Qwen2.5-VL-7B (Bai et al., 2025)	86.67	0.65	–	–
Qwen3.5-9B (Qwen Team, 2026)	83.77	1.30	–	–
InternVL3.5-8B (Wang et al., 2025b)	122.19	1.30	0.65	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	37.99	–	–	–
GeoVerse (Ours)	398.77	7.79	**6.49**	0.65

Table 12: **Detailed results on Landmark.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=500)				
GLOBE (Li et al., 2025a)	**4089.85**	**92.80**	**81.60**	**40.00**
GeoClip (Vivanco Cepeda et al., 2023)	3373.59	84.80	61.60	21.40
G3 (Jia et al., 2024)	3250.67	83.20	58.60	17.00
Geo-R1 (Zhang et al., 2025c)	3668.56	85.00	73.60	31.60
Qwen2.5-VL-7B (Bai et al., 2025)	3698.53	86.60	70.80	35.60
Qwen3.5-9B (Qwen Team, 2026)	3229.37	80.40	59.40	19.60
InternVL3.5-8B (Wang et al., 2025b)	2371.87	66.00	36.40	10.60
Gemma4-26B-A4B (Gemma Team et al., 2026)	1731.41	47.00	27.60	8.80
GeoVerse (Ours)	3538.17	85.20	68.80	25.40

Continued on next page

Table 12 continued: Landmark

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Africa (n=20)				
GLOBE (Li et al., 2025a)	**3900.88**	**80.00**	**75.00**	**50.00**
GeoClip (Vivanco Cepeda et al., 2023)	2723.27	65.00	50.00	30.00
G3 (Jia et al., 2024)	2308.35	55.00	35.00	30.00
Geo-R1 (Zhang et al., 2025c)	3312.03	70.00	65.00	**50.00**
Qwen2.5-VL-7B (Bai et al., 2025)	3284.87	75.00	60.00	45.00
Qwen3.5-9B (Qwen Team, 2026)	2229.19	45.00	40.00	30.00
InternVL3.5-8B (Wang et al., 2025b)	1159.82	30.00	15.00	10.00
Gemma4-26B-A4B (Gemma Team et al., 2026)	689.02	10.00	10.00	10.00
GeoVerse (Ours)	2722.08	55.00	55.00	25.00
Asia (n=144)				
GLOBE (Li et al., 2025a)	**4277.14**	**97.92**	**84.72**	**42.36**
GeoClip (Vivanco Cepeda et al., 2023)	3421.38	84.72	63.89	22.22
G3 (Jia et al., 2024)	3347.94	85.42	58.33	18.75
Geo-R1 (Zhang et al., 2025c)	3814.44	89.58	75.69	36.81
Qwen2.5-VL-7B (Bai et al., 2025)	3855.96	90.97	73.61	38.89
Qwen3.5-9B (Qwen Team, 2026)	3549.24	86.81	65.28	23.61
InternVL3.5-8B (Wang et al., 2025b)	2715.91	75.00	43.75	15.28
Gemma4-26B-A4B (Gemma Team et al., 2026)	1739.33	47.22	25.69	8.33
GeoVerse (Ours)	3749.08	87.50	72.22	33.33
Europe (n=140)				
GLOBE (Li et al., 2025a)	**4059.78**	95.00	**81.43**	**35.71**
GeoClip (Vivanco Cepeda et al., 2023)	3378.56	90.00	58.57	12.86
G3 (Jia et al., 2024)	3351.16	87.86	62.14	10.00
Geo-R1 (Zhang et al., 2025c)	3747.10	87.14	75.71	29.29
Qwen2.5-VL-7B (Bai et al., 2025)	3782.06	90.71	72.14	30.00
Qwen3.5-9B (Qwen Team, 2026)	3584.82	92.86	68.57	15.71
InternVL3.5-8B (Wang et al., 2025b)	2830.95	76.43	42.86	11.43
Gemma4-26B-A4B (Gemma Team et al., 2026)	1966.68	56.43	31.43	7.86
GeoVerse (Ours)	3801.12	**95.71**	76.43	20.00
North America (n=122)				
GLOBE (Li et al., 2025a)	**3983.07**	88.52	**79.51**	**36.89**
GeoClip (Vivanco Cepeda et al., 2023)	3768.31	**90.98**	70.49	26.23
G3 (Jia et al., 2024)	3457.05	86.89	63.11	18.85
Geo-R1 (Zhang et al., 2025c)	3689.01	83.61	73.77	27.87
Qwen2.5-VL-7B (Bai et al., 2025)	3572.60	83.61	68.85	30.33
Qwen3.5-9B (Qwen Team, 2026)	3070.49	77.05	54.10	14.75
InternVL3.5-8B (Wang et al., 2025b)	2281.75	68.85	31.97	4.92
Gemma4-26B-A4B (Gemma Team et al., 2026)	1705.35	47.54	26.23	6.56
GeoVerse (Ours)	3364.58	83.61	63.93	15.57
Oceania (n=21)				
GLOBE (Li et al., 2025a)	**4601.60**	**100.00**	**95.24**	52.38
GeoClip (Vivanco Cepeda et al., 2023)	3143.05	76.19	61.90	23.81
G3 (Jia et al., 2024)	3152.23	85.71	66.67	19.05
Geo-R1 (Zhang et al., 2025c)	3746.84	90.48	85.71	23.81
Qwen2.5-VL-7B (Bai et al., 2025)	4354.41	95.24	85.71	**57.14**
Qwen3.5-9B (Qwen Team, 2026)	2999.33	76.19	57.14	28.57
InternVL3.5-8B (Wang et al., 2025b)	1494.99	42.86	28.57	4.76
Gemma4-26B-A4B (Gemma Team et al., 2026)	1557.76	42.86	33.33	9.52
GeoVerse (Ours)	3517.95	76.19	71.43	42.86

Continued on next page

Table 12 continued: Landmark

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
South America (n=53)				
GLOBE (Li et al., 2025a)	**3774.81**	**84.91**	**75.47**	**43.40**
GeoClip (Vivanco Cepeda et al., 2023)	2658.73	67.92	47.17	26.42
G3 (Jia et al., 2024)	2640.47	66.04	45.28	20.75
Geo-R1 (Zhang et al., 2025c)	3121.17	73.58	60.38	28.30
Qwen2.5-VL-7B (Bai et al., 2025)	3236.26	71.70	62.26	41.51
Qwen3.5-9B (Qwen Team, 2026)	2255.65	52.83	39.62	22.64
InternVL3.5-8B (Wang et al., 2025b)	1236.74	30.19	20.75	11.32
Gemma4-26B-A4B (Gemma Team et al., 2026)	1610.63	35.85	30.19	16.98
GeoVerse (Ours)	2986.06	69.81	54.72	33.96

Table 13: **Detailed results on Road-network.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=500)				
GLOBE (Li et al., 2025a)	844.23	28.40	6.20	0.40
GeoClip (Vivanco Cepeda et al., 2023)	935.67	26.60	10.40	0.80
G3 (Jia et al., 2024)	990.06	29.60	10.80	**1.20**
Geo-R1 (Zhang et al., 2025c)	860.10	27.60	8.40	0.60
Qwen2.5-VL-7B (Bai et al., 2025)	894.55	28.00	7.40	0.40
Qwen3.5-9B (Qwen Team, 2026)	672.66	20.20	6.80	0.80
InternVL3.5-8B (Wang et al., 2025b)	681.00	19.00	5.60	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	388.78	11.40	4.00	0.60
GeoVerse (Ours)	**1310.63**	**40.00**	**15.80**	**1.20**
Africa (n=38)				
GLOBE (Li et al., 2025a)	429.25	18.42	**2.63**	–
GeoClip (Vivanco Cepeda et al., 2023)	**708.56**	**23.68**	–	–
G3 (Jia et al., 2024)	421.55	13.16	**2.63**	–
Geo-R1 (Zhang et al., 2025c)	380.56	7.89	–	–
Qwen2.5-VL-7B (Bai et al., 2025)	488.55	18.42	–	–
Qwen3.5-9B (Qwen Team, 2026)	597.98	**23.68**	–	–
InternVL3.5-8B (Wang et al., 2025b)	230.29	7.89	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	121.82	5.26	–	–
GeoVerse (Ours)	518.66	15.79	**2.63**	–
Asia (n=141)				
GLOBE (Li et al., 2025a)	675.51	21.28	3.55	–
GeoClip (Vivanco Cepeda et al., 2023)	776.13	23.40	6.38	–
G3 (Jia et al., 2024)	**1003.49**	**33.33**	6.38	–
Geo-R1 (Zhang et al., 2025c)	607.77	16.31	6.38	0.71
Qwen2.5-VL-7B (Bai et al., 2025)	416.95	12.06	2.13	–
Qwen3.5-9B (Qwen Team, 2026)	305.19	8.51	0.71	–
InternVL3.5-8B (Wang et al., 2025b)	304.18	4.26	0.71	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	142.06	4.26	2.13	–
GeoVerse (Ours)	780.73	22.70	**8.51**	**1.42**
Europe (n=143)				
GLOBE (Li et al., 2025a)	1026.07	34.97	5.59	–
GeoClip (Vivanco Cepeda et al., 2023)	668.74	14.69	5.59	–
G3 (Jia et al., 2024)	670.30	14.69	8.39	**0.70**
Geo-R1 (Zhang et al., 2025c)	920.73	32.87	6.99	–
Qwen2.5-VL-7B (Bai et al., 2025)	1187.00	38.46	9.79	–
Qwen3.5-9B (Qwen Team, 2026)	467.16	12.59	5.59	**0.70**
InternVL3.5-8B (Wang et al., 2025b)	539.12	16.08	3.50	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	117.53	1.40	1.40	–
GeoVerse (Ours)	**1366.12**	**45.45**	**13.29**	–

Continued on next page

Table 13 continued: Road-network

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
North America (n=124)				
GLOBE (Li et al., 2025a)	1303.77	43.55	13.71	1.61
GeoClip (Vivanco Cepeda et al., 2023)	1873.99	56.45	28.23	3.23
G3 (Jia et al., 2024)	1905.59	59.68	25.00	**4.03**
Geo-R1 (Zhang et al., 2025c)	1569.04	51.61	18.55	1.61
Qwen2.5-VL-7B (Bai et al., 2025)	1552.42	48.39	15.32	1.61
Qwen3.5-9B (Qwen Team, 2026)	1629.86	50.00	20.16	2.42
InternVL3.5-8B (Wang et al., 2025b)	1691.94	50.81	17.74	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	1211.55	37.10	12.10	2.42
GeoVerse (Ours)	**2647.44**	**78.23**	**37.90**	3.23
Oceania (n=7)				
GLOBE (Li et al., 2025a)	3.38	–	–	–
GeoClip (Vivanco Cepeda et al., 2023)	140.25	–	–	–
G3 (Jia et al., 2024)	**662.06**	**14.29**	**14.29**	–
Geo-R1 (Zhang et al., 2025c)	2.86	–	–	–
Qwen2.5-VL-7B (Bai et al., 2025)	14.87	–	–	–
Qwen3.5-9B (Qwen Team, 2026)	25.77	–	–	–
InternVL3.5-8B (Wang et al., 2025b)	44.97	–	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	–	–	–	–
GeoVerse (Ours)	10.64	–	–	–
South America (n=47)				
GLOBE (Li et al., 2025a)	45.48	**2.13**	–	–
GeoClip (Vivanco Cepeda et al., 2023)	52.99	–	–	–
G3 (Jia et al., 2024)	15.77	–	–	–
Geo-R1 (Zhang et al., 2025c)	77.67	**2.13**	–	–
Qwen2.5-VL-7B (Bai et al., 2025)	**161.18**	**2.13**	**2.13**	–
Qwen3.5-9B (Qwen Team, 2026)	31.67	–	–	–
InternVL3.5-8B (Wang et al., 2025b)	35.11	–	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	57.24	**2.13**	–	–
GeoVerse (Ours)	38.47	–	–	–

Table 14: **Detailed results on Indoor.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=500)				
GLOBE (Li et al., 2025a)	1348.56	37.00	**20.80**	5.00
GeoClip (Vivanco Cepeda et al., 2023)	**1370.93**	**39.80**	17.60	3.60
G3 (Jia et al., 2024)	1251.23	35.80	17.40	2.80
Geo-R1 (Zhang et al., 2025c)	1216.81	34.20	16.60	4.80
Qwen2.5-VL-7B (Bai et al., 2025)	1112.46	29.80	16.00	4.00
Qwen3.5-9B (Qwen Team, 2026)	1079.68	30.80	14.20	3.80
InternVL3.5-8B (Wang et al., 2025b)	840.31	22.80	12.00	3.00
Gemma4-26B-A4B (Gemma Team et al., 2026)	661.11	18.20	10.00	1.80
GeoVerse (Ours)	1349.17	36.20	**20.80**	**5.20**
Africa (n=34)				
GLOBE (Li et al., 2025a)	838.11	23.53	**14.71**	**2.94**
GeoClip (Vivanco Cepeda et al., 2023)	**1019.42**	**29.41**	**14.71**	**2.94**
G3 (Jia et al., 2024)	716.92	17.65	8.82	–
Geo-R1 (Zhang et al., 2025c)	644.04	20.59	8.82	–
Qwen2.5-VL-7B (Bai et al., 2025)	815.18	20.59	**14.71**	–
Qwen3.5-9B (Qwen Team, 2026)	816.81	20.59	**14.71**	–
InternVL3.5-8B (Wang et al., 2025b)	841.21	23.53	11.76	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	634.34	20.59	5.88	**2.94**
GeoVerse (Ours)	791.14	17.65	**14.71**	–

Continued on next page

Table 14 continued: Indoor

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Asia (n=135)				
GLOBE (Li et al., 2025a)	2355.29	**61.48**	**40.74**	6.67
GeoClip (Vivanco Cepeda et al., 2023)	2112.91	56.30	33.33	4.44
G3 (Jia et al., 2024)	2050.75	57.78	35.56	1.48
Geo-R1 (Zhang et al., 2025c)	2242.24	**61.48**	36.30	7.41
Qwen2.5-VL-7B (Bai et al., 2025)	1926.56	48.15	31.85	4.44
Qwen3.5-9B (Qwen Team, 2026)	1984.01	57.78	29.63	4.44
InternVL3.5-8B (Wang et al., 2025b)	1840.82	48.15	30.37	5.19
Gemma4-26B-A4B (Gemma Team et al., 2026)	1349.50	35.56	24.44	2.22
GeoVerse (Ours)	**2377.98**	**61.48**	38.52	**8.15**
Europe (n=13)				
GLOBE (Li et al., 2025a)	2094.53	66.67	**25.00**	**8.33**
GeoClip (Vivanco Cepeda et al., 2023)	1558.61	58.33	8.33	–
G3 (Jia et al., 2024)	1877.94	58.33	16.67	–
Geo-R1 (Zhang et al., 2025c)	1529.41	58.33	16.67	–
Qwen2.5-VL-7B (Bai et al., 2025)	2215.23	75.00	16.67	–
Qwen3.5-9B (Qwen Team, 2026)	2193.85	83.33	8.33	**8.33**
InternVL3.5-8B (Wang et al., 2025b)	1709.53	75.00	8.33	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	1169.98	41.67	8.33	–
GeoVerse (Ours)	**2406.93**	**100.00**	–	–
North America (n=42)				
GLOBE (Li et al., 2025a)	1788.99	42.86	26.19	14.29
GeoClip (Vivanco Cepeda et al., 2023)	1977.17	54.76	23.81	9.52
G3 (Jia et al., 2024)	1895.96	54.76	26.19	11.90
Geo-R1 (Zhang et al., 2025c)	2084.55	57.14	23.81	**21.43**
Qwen2.5-VL-7B (Bai et al., 2025)	**2196.13**	**66.67**	**28.57**	19.05
Qwen3.5-9B (Qwen Team, 2026)	1814.07	47.62	21.43	19.05
InternVL3.5-8B (Wang et al., 2025b)	1458.31	42.86	19.05	11.90
Gemma4-26B-A4B (Gemma Team et al., 2026)	994.04	28.57	16.67	7.14
GeoVerse (Ours)	1863.61	42.86	**28.57**	16.67
Oceania (n=60)				
GLOBE (Li et al., 2025a)	649.34	18.33	8.33	3.33
GeoClip (Vivanco Cepeda et al., 2023)	**1057.09**	**26.67**	**13.33**	5.00
G3 (Jia et al., 2024)	1054.53	**26.67**	10.00	**8.33**
Geo-R1 (Zhang et al., 2025c)	621.66	13.33	8.33	3.33
Qwen2.5-VL-7B (Bai et al., 2025)	244.75	5.00	1.67	1.67
Qwen3.5-9B (Qwen Team, 2026)	536.14	11.67	8.33	1.67
InternVL3.5-8B (Wang et al., 2025b)	264.76	5.00	5.00	3.33
Gemma4-26B-A4B (Gemma Team et al., 2026)	206.91	5.00	3.33	1.67
GeoVerse (Ours)	620.50	15.00	10.00	3.33
South America (n=216)				
GLOBE (Li et al., 2025a)	873.10	26.39	11.57	**2.78**
GeoClip (Vivanco Cepeda et al., 2023)	**927.73**	**31.02**	8.80	1.85
G3 (Jia et al., 2024)	735.88	22.69	7.87	0.93
Geo-R1 (Zhang et al., 2025c)	650.93	19.44	6.48	1.39
Qwen2.5-VL-7B (Bai et al., 2025)	624.63	17.13	7.87	2.31
Qwen3.5-9B (Qwen Team, 2026)	507.13	14.81	5.09	1.39
InternVL3.5-8B (Wang et al., 2025b)	210.14	5.09	1.39	0.46
Gemma4-26B-A4B (Gemma Team et al., 2026)	271.25	7.41	2.31	0.46
GeoVerse (Ours)	843.85	24.54	**13.43**	**2.78**

Table 15: **Detailed results on Shape.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=500)				
GLOBE (Li et al., 2025a)	821.34	23.20	9.40	–
GeoClip (Vivanco Cepeda et al., 2023)	699.27	22.00	7.20	–
G3 (Jia et al., 2024)	623.36	16.00	7.20	–
Geo-R1 (Zhang et al., 2025c)	748.65	19.40	10.20	–
Qwen2.5-VL-7B (Bai et al., 2025)	952.73	26.60	10.20	–
Qwen3.5-9B (Qwen Team, 2026)	525.06	14.60	3.20	–
InternVL3.5-8B (Wang et al., 2025b)	428.08	13.20	2.40	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	644.57	15.60	9.60	–
GeoVerse (Ours)	**1129.73**	**33.40**	**12.80**	–
Africa (n=84)				
GLOBE (Li et al., 2025a)	802.01	19.05	9.52	–
GeoClip (Vivanco Cepeda et al., 2023)	694.93	20.24	5.95	–
G3 (Jia et al., 2024)	532.95	16.67	2.38	–
Geo-R1 (Zhang et al., 2025c)	819.72	19.05	10.71	–
Qwen2.5-VL-7B (Bai et al., 2025)	977.14	27.38	8.33	–
Qwen3.5-9B (Qwen Team, 2026)	504.00	15.48	1.19	–
InternVL3.5-8B (Wang et al., 2025b)	197.00	5.95	1.19	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	774.74	17.86	**13.10**	–
GeoVerse (Ours)	**1163.58**	**34.52**	8.33	–
Asia (n=159)				
GLOBE (Li et al., 2025a)	680.88	16.35	10.06	–
GeoClip (Vivanco Cepeda et al., 2023)	352.67	6.92	5.66	–
G3 (Jia et al., 2024)	356.62	8.18	4.40	–
Geo-R1 (Zhang et al., 2025c)	627.44	12.58	10.06	–
Qwen2.5-VL-7B (Bai et al., 2025)	**810.21**	18.87	**11.95**	–
Qwen3.5-9B (Qwen Team, 2026)	518.30	11.32	5.03	–
InternVL3.5-8B (Wang et al., 2025b)	500.11	13.21	1.89	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	586.06	15.09	7.55	–
GeoVerse (Ours)	740.28	**20.13**	6.92	–
Europe (n=125)				
GLOBE (Li et al., 2025a)	1342.47	45.60	11.20	–
GeoClip (Vivanco Cepeda et al., 2023)	1259.43	44.80	12.00	–
G3 (Jia et al., 2024)	1066.20	27.20	13.60	–
Geo-R1 (Zhang et al., 2025c)	1114.41	35.20	13.60	–
Qwen2.5-VL-7B (Bai et al., 2025)	1459.20	47.20	12.00	–
Qwen3.5-9B (Qwen Team, 2026)	701.51	23.20	3.20	–
InternVL3.5-8B (Wang et al., 2025b)	692.24	26.40	4.80	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	645.58	15.20	9.60	–
GeoVerse (Ours)	**2258.69**	**70.40**	**27.20**	–
North America (n=62)				
GLOBE (Li et al., 2025a)	740.15	19.35	9.68	–
GeoClip (Vivanco Cepeda et al., 2023)	734.10	22.58	8.06	–
G3 (Jia et al., 2024)	801.46	20.97	11.29	–
Geo-R1 (Zhang et al., 2025c)	589.15	16.13	4.84	–
Qwen2.5-VL-7B (Bai et al., 2025)	**995.56**	**24.19**	12.90	–
Qwen3.5-9B (Qwen Team, 2026)	541.90	14.52	4.84	–
InternVL3.5-8B (Wang et al., 2025b)	252.17	8.06	1.61	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	737.05	17.74	11.29	–
GeoVerse (Ours)	755.79	19.35	**14.52**	–

Continued on next page

Table 15 continued: Shape

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Oceania (n=18)				
GLOBE (Li et al., 2025a)	466.43	11.11	5.56	–
GeoClip (Vivanco Cepeda et al., 2023)	612.18	**16.67**	5.56	–
G3 (Jia et al., 2024)	449.36	11.11	–	–
Geo-R1 (Zhang et al., 2025c)	**661.54**	**16.67**	**11.11**	–
Qwen2.5-VL-7B (Bai et al., 2025)	184.15	5.56	–	–
Qwen3.5-9B (Qwen Team, 2026)	204.93	5.56	–	–
InternVL3.5-8B (Wang et al., 2025b)	198.46	5.56	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	309.58	5.56	5.56	–
GeoVerse (Ours)	231.66	11.11	–	–
South America (n=52)				
GLOBE (Li et al., 2025a)	249.02	5.77	3.85	–
GeoClip (Vivanco Cepeda et al., 2023)	408.13	**17.31**	1.92	–
G3 (Jia et al., 2024)	368.32	7.69	5.77	–
Geo-R1 (Zhang et al., 2025c)	345.59	7.69	7.69	–
Qwen2.5-VL-7B (Bai et al., 2025)	346.57	9.62	3.85	–
Qwen3.5-9B (Qwen Team, 2026)	246.37	5.77	–	–
InternVL3.5-8B (Wang et al., 2025b)	235.35	1.92	1.92	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	**616.43**	15.38	**9.62**	–
GeoVerse (Ours)	308.74	7.69	5.77	–

Table 16: **Detailed results on Space.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=500)				
GLOBE (Li et al., 2025a)	1699.81	47.00	25.20	2.20
GeoClip (Vivanco Cepeda et al., 2023)	1697.49	46.80	27.00	**3.40**
G3 (Jia et al., 2024)	1686.35	43.80	**27.80**	2.00
Geo-R1 (Zhang et al., 2025c)	1379.90	38.60	20.80	3.00
Qwen2.5-VL-7B (Bai et al., 2025)	1732.80	47.80	26.60	2.60
Qwen3.5-9B (Qwen Team, 2026)	1269.13	35.60	16.60	1.80
InternVL3.5-8B (Wang et al., 2025b)	735.70	20.40	8.20	0.60
Gemma4-26B-A4B (Gemma Team et al., 2026)	615.84	16.60	8.40	1.40
GeoVerse (Ours)	**1737.30**	**48.60**	25.60	2.60
Africa (n=85)				
GLOBE (Li et al., 2025a)	838.47	25.88	8.24	–
GeoClip (Vivanco Cepeda et al., 2023)	1005.14	**34.12**	11.76	–
G3 (Jia et al., 2024)	805.52	20.00	**14.12**	–
Geo-R1 (Zhang et al., 2025c)	808.18	22.35	10.59	–
Qwen2.5-VL-7B (Bai et al., 2025)	1000.67	32.94	5.88	–
Qwen3.5-9B (Qwen Team, 2026)	**1012.84**	29.41	**14.12**	–
InternVL3.5-8B (Wang et al., 2025b)	734.05	23.53	5.88	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	402.62	11.76	2.35	–
GeoVerse (Ours)	868.52	25.88	9.41	–
Asia (n=81)				
GLOBE (Li et al., 2025a)	**1938.02**	**60.49**	19.75	**2.47**
GeoClip (Vivanco Cepeda et al., 2023)	1445.65	41.98	18.52	1.23
G3 (Jia et al., 2024)	1619.61	48.15	19.75	–
Geo-R1 (Zhang et al., 2025c)	1441.68	48.15	12.35	**2.47**
Qwen2.5-VL-7B (Bai et al., 2025)	1606.26	45.68	18.52	1.23
Qwen3.5-9B (Qwen Team, 2026)	1666.69	50.62	18.52	–
InternVL3.5-8B (Wang et al., 2025b)	836.56	23.46	9.88	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	654.39	18.52	7.41	–
GeoVerse (Ours)	1708.66	54.32	**22.22**	1.23

Continued on next page

Table 16 continued: Space

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Europe (n=71)				
GLOBE (Li et al., 2025a)	2435.21	59.15	42.25	4.23
GeoClip (Vivanco Cepeda et al., 2023)	2354.76	61.97	38.03	**11.27**
G3 (Jia et al., 2024)	2244.67	54.93	39.44	5.63
Geo-R1 (Zhang et al., 2025c)	2030.70	57.75	32.39	5.63
Qwen2.5-VL-7B (Bai et al., 2025)	**2585.24**	**71.83**	**46.48**	–
Qwen3.5-9B (Qwen Team, 2026)	1216.28	29.58	15.49	4.23
InternVL3.5-8B (Wang et al., 2025b)	639.17	14.08	8.45	1.41
Gemma4-26B-A4B (Gemma Team et al., 2026)	501.28	11.27	7.04	2.82
GeoVerse (Ours)	1963.10	59.15	26.76	2.82
North America (n=176)				
GLOBE (Li et al., 2025a)	2107.15	56.25	34.66	3.41
GeoClip (Vivanco Cepeda et al., 2023)	2230.89	56.82	41.48	4.55
G3 (Jia et al., 2024)	2439.75	63.07	41.48	3.41
Geo-R1 (Zhang et al., 2025c)	1779.61	46.59	30.68	4.55
Qwen2.5-VL-7B (Bai et al., 2025)	2343.11	61.93	40.34	**6.82**
Qwen3.5-9B (Qwen Team, 2026)	1756.51	48.30	25.57	3.41
InternVL3.5-8B (Wang et al., 2025b)	968.74	27.27	10.80	1.14
Gemma4-26B-A4B (Gemma Team et al., 2026)	917.06	25.00	14.77	2.84
GeoVerse (Ours)	**2747.61**	**72.16**	**44.32**	5.68
Oceania (n=22)				
GLOBE (Li et al., 2025a)	827.04	**27.27**	4.55	–
GeoClip (Vivanco Cepeda et al., 2023)	**852.44**	22.73	13.64	–
G3 (Jia et al., 2024)	770.84	22.73	13.64	–
Geo-R1 (Zhang et al., 2025c)	709.27	18.18	**18.18**	–
Qwen2.5-VL-7B (Bai et al., 2025)	240.73	4.55	–	–
Qwen3.5-9B (Qwen Team, 2026)	220.42	9.09	–	–
InternVL3.5-8B (Wang et al., 2025b)	159.44	4.55	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	47.52	–	–	–
GeoVerse (Ours)	82.47	–	–	–
South America (n=65)				
GLOBE (Li et al., 2025a)	918.49	26.15	**16.92**	–
GeoClip (Vivanco Cepeda et al., 2023)	**1040.48**	**33.85**	10.77	–
G3 (Jia et al., 2024)	581.40	12.31	10.77	–
Geo-R1 (Zhang et al., 2025c)	484.37	12.31	6.15	**1.54**
Qwen2.5-VL-7B (Bai et al., 2025)	769.26	20.00	13.85	–
Qwen3.5-9B (Qwen Team, 2026)	201.83	6.15	–	–
InternVL3.5-8B (Wang et al., 2025b)	281.67	6.15	4.62	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	348.47	9.23	4.62	–
GeoVerse (Ours)	486.89	12.31	7.69	–

Table 17: **Detailed results on UAV.**

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Overall (n=500)				
GLOBE (Li et al., 2025a)	**1810.93**	**49.40**	**28.20**	**3.80**
GeoClip (Vivanco Cepeda et al., 2023)	1527.30	42.40	23.00	3.00
G3 (Jia et al., 2024)	1426.69	41.40	18.40	2.00
Geo-R1 (Zhang et al., 2025c)	1587.11	43.60	25.00	**3.80**
Qwen2.5-VL-7B (Bai et al., 2025)	1344.47	38.80	19.00	2.00
Qwen3.5-9B (Qwen Team, 2026)	1202.87	35.40	14.60	0.80
InternVL3.5-8B (Wang et al., 2025b)	611.74	17.60	5.60	0.80
Gemma4-26B-A4B (Gemma Team et al., 2026)	573.73	17.40	6.40	0.60
GeoVerse (Ours)	1575.76	43.20	25.40	2.40

Continued on next page

Table 17 continued: UAV

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
Africa (n=83)				
GLOBE (Li et al., 2025a)	**987.51**	**25.30**	**14.46**	**2.41**
GeoClip (Vivanco Cepeda et al., 2023)	871.38	24.10	13.25	1.20
G3 (Jia et al., 2024)	629.33	20.48	4.82	1.20
Geo-R1 (Zhang et al., 2025c)	678.04	21.69	7.23	–
Qwen2.5-VL-7B (Bai et al., 2025)	767.49	**25.30**	6.02	–
Qwen3.5-9B (Qwen Team, 2026)	497.61	12.05	1.20	–
InternVL3.5-8B (Wang et al., 2025b)	66.74	–	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	210.56	7.23	1.20	–
GeoVerse (Ours)	597.32	10.84	8.43	1.20
Asia (n=92)				
GLOBE (Li et al., 2025a)	**2475.91**	**65.22**	**40.22**	**6.52**
GeoClip (Vivanco Cepeda et al., 2023)	1480.51	44.57	18.48	4.35
G3 (Jia et al., 2024)	1385.55	39.13	15.22	1.09
Geo-R1 (Zhang et al., 2025c)	2173.87	58.70	36.96	3.26
Qwen2.5-VL-7B (Bai et al., 2025)	1689.11	47.83	26.09	**6.52**
Qwen3.5-9B (Qwen Team, 2026)	1990.71	58.70	25.00	3.26
InternVL3.5-8B (Wang et al., 2025b)	1015.86	27.17	10.87	4.35
Gemma4-26B-A4B (Gemma Team et al., 2026)	826.08	23.91	8.70	1.09
GeoVerse (Ours)	2267.17	61.96	38.04	5.43
Europe (n=89)				
GLOBE (Li et al., 2025a)	**2681.44**	**78.65**	41.57	1.12
GeoClip (Vivanco Cepeda et al., 2023)	2004.79	57.30	30.34	–
G3 (Jia et al., 2024)	2066.41	60.67	31.46	1.12
Geo-R1 (Zhang et al., 2025c)	2328.11	67.42	35.96	**4.49**
Qwen2.5-VL-7B (Bai et al., 2025)	2155.40	62.92	34.83	–
Qwen3.5-9B (Qwen Team, 2026)	1793.04	51.69	28.09	–
InternVL3.5-8B (Wang et al., 2025b)	977.24	30.34	8.99	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	627.55	16.85	10.11	–
GeoVerse (Ours)	2520.56	71.91	**47.19**	1.12
North America (n=95)				
GLOBE (Li et al., 2025a)	1859.03	48.42	30.53	5.26
GeoClip (Vivanco Cepeda et al., 2023)	**2278.48**	**57.89**	**38.95**	**7.37**
G3 (Jia et al., 2024)	2027.61	56.84	28.42	5.26
Geo-R1 (Zhang et al., 2025c)	1642.65	43.16	27.37	**7.37**
Qwen2.5-VL-7B (Bai et al., 2025)	1609.99	44.21	23.16	3.16
Qwen3.5-9B (Qwen Team, 2026)	1432.44	44.21	16.84	–
InternVL3.5-8B (Wang et al., 2025b)	1075.34	34.74	9.47	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	888.57	28.42	9.47	1.05
GeoVerse (Ours)	1975.02	55.79	29.47	4.21
Oceania (n=52)				
GLOBE (Li et al., 2025a)	1027.70	30.77	11.54	–
GeoClip (Vivanco Cepeda et al., 2023)	1447.21	**44.23**	17.31	1.92
G3 (Jia et al., 2024)	1409.45	38.46	17.31	1.92
Geo-R1 (Zhang et al., 2025c)	**1499.12**	38.46	**19.23**	**3.85**
Qwen2.5-VL-7B (Bai et al., 2025)	820.53	25.00	11.54	1.92
Qwen3.5-9B (Qwen Team, 2026)	736.86	21.15	9.62	1.92
InternVL3.5-8B (Wang et al., 2025b)	184.85	5.77	1.92	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	227.37	7.69	1.92	–
GeoVerse (Ours)	981.40	28.85	15.38	1.92

Continued on next page

Table 17 continued: UAV

Model	Mean GeoScore	Continent Acc. (%)	Country Acc. (%)	City Acc. (%)
South America (n=89)				
GLOBE (Li et al., 2025a)	**1427.21**	**38.20**	**22.47**	**5.62**
GeoClip (Vivanco Cepeda et al., 2023)	954.85	24.72	15.73	2.25
G3 (Jia et al., 2024)	941.74	29.21	11.24	1.12
Geo-R1 (Zhang et al., 2025c)	1079.50	28.09	19.10	3.37
Qwen2.5-VL-7B (Bai et al., 2025)	738.07	20.22	7.87	–
Qwen3.5-9B (Qwen Team, 2026)	483.25	15.73	3.37	–
InternVL3.5-8B (Wang et al., 2025b)	91.29	–	–	–
Gemma4-26B-A4B (Gemma Team et al., 2026)	464.07	14.61	4.49	1.12
GeoVerse (Ours)	749.79	20.22	7.87	–

Table 18: **GeoVerse with different base models.** Mean GeoScore is reported.

Model	Street-View	Remote Sensing	Terrain and Plant	Map	Landmark	Road-network	Indoor	Shape	Space	UAV
Qwen3.5-9B (Qwen Team, 2026)	1998.85	889.45	471.84	717.11	3229.37	672.66	1079.68	525.06	1269.13	1202.87
InternVL3.5-8B (Wang et al., 2025b)	1707.97	649.05	429.73	478.71	2371.87	681.00	840.31	428.08	735.70	611.74
GeoVerse (InternVL3.5-8B)	2006.40	871.53	534.82	841.58	3160.63	1095.92	947.01	832.70	1054.83	889.20
GeoVerse (Ours)	**2894.46**	**1286.23**	**599.70**	**966.29**	**3538.17**	**1310.63**	**1349.17**	**1129.73**	**1737.30**	**1575.76**

City/ Wild Prompt of UAV

Prompt: ” Classify this UAV or aerial image. Return ONLY one valid and compact JSON object in the following format:
{
  "scene_class": "dense_urban",
  "confidence": 0.0
}
Allowed scene_class values:
- "dense_urban"
- "suburban_residential"
- "industrial_transport"
- "rural_farmland"
- "forest_vegetation"
- "water_coast"
- "barren_mountain"
Scene definitions:
- dense_urban: Compact city blocks, high-density roads, downtown areas,dense housing, or extensive built-up urban fabric.
- suburban_residential: Lower-density houses, villages, small towns, or residential blocks mixed with yards, local roads, and light vegetation.
- industrial_transport: Airports, ports, railway yards, highways, parking lots, warehouses, factories, mines, construction si tes, solar or wind farms, and major utility infrastructure.
- rural_farmland: Crop fields, orchards, pastures, rural tracks, sparse farms, or predominantly agricultural landscapes.
- forest_vegetation: Dominant tree cover, dense vegetation, grassland, wetland vegetation, parks, or natural green cover with little built structure.
- water_coast: Seas, beaches, lakes, rivers, lagoons, marinas, reefs, or coastlines that constitute the dominant visual cue.
- barren_mountain: Bare soil, desert, quarry-like barren ground, rocky hills, mountains, or dry open land.
If the image contains multiple scene types, select the single dominant class according to its visible area and geolocation va lue. If uncertain between a built-environment class and a natural or rural class, select the class that provides the most use ful evidence for geolocating the image. The confidence value must be a number between 0.0 and 1.0. Do not return Markdown, ex planations, or any text outside the JSON object"

View Judgement Prompt of UAV

Prompt: ” You are a rigorous geospatial benchmark evaluation expert. Determine whether the given image is suitable for visual geolocation evaluation. Examples of invalid images include pure desert, open ocean, boundaryless dense
forest canopy, pure cloud cover, and severely blurred images.
Return ONLY one valid JSON object in the following format:
{
  "is_valid": true,
  "view_angle": "nadir",
  "reason": "A one-sentence explanation of the decision."
}
Allowed values:
- is_valid: true or false
- view_angle: "nadir" or "oblique"
Viewpoint definitions:
- nadir: A vertical or approximately top-down view. Buildings, roads, fields, and water bodies are observed primarily from above, with little visible facade information or perspective distortion.
- oblique: A tilted aerial view with visible building facades, strong depth perspective, or a visible horizon.
The reason must concisely identify the visual evidence supporting the decision. Do not return Markdown, explanations, or any text outside the JSON object.“

Human / Physical Judgement Prompt of Streetview

Prompt: ”You are given a street-view image for visual geolocation. Your task is to make a binary decision: determine whether the main evidence for geolocating this image relies more on human geography or physical geography.
Definitions:
- Human geography: man-made or culture-related cues, such as road signs, language, architecture, vehicles, road markings, infrastructure, shops, utility poles, traffic rules, clothing, urban layout, or cultural symbols.
- Physical geography: natural environmental cues, such as vegetation, terrain, mountains, rivers, coastline, climate, soil, desert, snow, sunlight, landforms, or natural landscape.
You must choose exactly one of the following two labels:
- "human_geography"
- "physical_geography"
Even if both types of cues are present, select the one that contributes more to geolocation.
Return your answer strictly in JSON format:
{
 "primary_geography_type": "human_geography" | "physical_geography",
 "confidence": 0.0-1.0,
 "key_evidence": ["brief evidence 1", "brief evidence 2", "brief evidence 3"],
 "reasoning_summary": "A concise explanation of why this image relies more on the selected geography type for geolocation."
}”

Level Judgement Prompt of Space

Prompt: ”You are given an space image for visual geolocation. Your task is to classify the geolocation resolution level that can reasonably be inferred from this image.
Use the following levels:
- Level 1: Country or continent level. The image only provides broad geographic clues, such as climate zone, large-scale terrain, coastline shape, desert, forest, mountain range, or general land-cover patterns. It is not enough to identify a specific province, city, or local area.
- Level 2: Province or city level. The image provides enough cues to infer a more specific administrative region or city-level area, such as urban structure, road network, river layout, agricultural patterns, coastline configuration, or regional landscape features.
- Level 3: Specific local area level. The image contains highly distinctive local cues that may support identifying a specific district, neighborhood, landmark, facility, airport, stadium, port, campus, industrial area, or other recognizable local region.
Analyze the image carefully and determine the most appropriate resolution level. Return your answer strictly in JSON format:
{
  "resolution_level": 1 | 2 | 3, "level_name": "country_or_continent" | "province_or_city" | "specific_local_area",
  "confidence": 0.0-1.0,
  "key_evidence": ["brief evidence 1", "brief evidence 2", "brief evidence 3"],
  "reasoning_summary": "A concise explanation of why the image supports this resolution level.”
}”

Figure 8: **Prompts used to construct benchmarks for selected domains.**

Train / Evaluation Prompt

```
Prompt: "You are a geolocation expert. Your task is to infer the single most likely geographic location corresponding to the input image and output it in a fixed format.
The image may be a street-view photo, ground-level photo, drone/UAV image, remote-sensing or satellite image, astronaut/space-view image, topographic/elevation map, vector map, road-network image, administrative-boundary silhouette, indoor scene, landmark image, land-cover raster, or another visual input related to geographic location.
Reason carefully from visual evidence before choosing the final location, and include the concise reasoning fields inside the JSON object.
Output exactly the following structure and no extra text:
{
"Clues": ["key visual clue 1", "key visual clue 2"],
"Reasoning": "Briefly identify the image type, extract the strongest location clues, localize from coarse to fine, and justify one final location.",
"FinalAnswer": "Country; Region; Specific Location"
}
Requirements:
1. The output must be valid JSON.
2. The JSON must contain exactly three keys: "Clues", "Reasoning", and "FinalAnswer".
3. Do not output coordinates, latitude/longitude, bounding boxes, numeric geolocation fields, multiple candidates, or extra notes.
4. Do not use refusal or uncertainty wording such as "cannot determine", "unable to locate", "don't know", "uncertain", "unclear", "unknown", "insufficient information", "cannot confirm", or similar phrases.
5. Keep the reasoning focused and decisive. Do not repeat observations, backtrack, list many rejected candidates, or continue after a clear best guess.
6. "Clues" must be a non-empty list of useful visual clues.
7. "Reasoning" must be a concise string that identifies the image type, uses the strongest clues, localizes from coarse to fine, and justifies one final location.
8. FinalAnswer must contain exactly three semicolon-separated parts and follow: "Country; Region; Specific Location".
Useful visual clues include:
- Street/ground/drone images: language, signs, license plates, traffic side, road markings, architecture, vegetation, landforms, climate, infrastructure, cultural elements.
- Satellite/remote-sensing/space images: coastlines, rivers, lakes, mountains, deserts, glaciers, urban shape, road grids, farmland patterns, ports, airports, reservoirs, vegetation cover, aridity/humidity.
- Topographic/elevation maps: mountain ranges, valleys, basins, plateaus, contour density, relief patterns.
- Vector maps/road networks: road hierarchy, interchanges, bridges, railways, road-grid style, relationships with rivers/lakes/bays, administrative boundaries.
- Boundary silhouettes: outline shape, coastlines, protrusions, narrow corridors, concavities, exclaves.
- Indoor/landmark images: signage language, interior style, brands, directional signs, numbering systems, distinctive structures.
Localization method:
First identify the image type, then extract relevant clues, then localize from coarse to fine:
continent or climate zone -> country -> region/state/province/county -> city/landmark/specific place.
For Esri land-cover / land-use raster images, interpret common colors as:
water = blue, trees = dark green, flooded vegetation = purple-blue, crops = orange, built area = red, bare ground = gray-brown, snow/ice = pale cyan, clouds = gray, rangeland = pale beige.
Example output:
{
  "Clues": ["desert coastal metropolis", "cluster of very tall modern towers", "wide multilane roads", "Arabic-style urban environment", "dense waterfront skyline"],
  "Reasoning": "This is a modern urban scene with a dense high-rise skyline, wide roads, arid lighting, and Gulf-style infrastructure. The cues point from an arid coastal metropolitan environment to the Arabian Peninsula, then to the United Arab Emirates. The tower density and urban texture most strongly match Dubai, so the best single location is Dubai.",
  "FinalAnswer": "United Arab Emirates; Dubai; Dubai"
}"
```

Figure 9: **Prompt of Training and Evaluation.**

United Kingdom; Scotland; Edinburgh



Answer

Clues: ["Scottish-style stone architecture", "overcast grey sky", "wet pavement", "people wearing coats", "large ornate building in background", "road markings and traffic signs", "urban street scene"],
Reasoning: "The image shows a street scene with distinctive Scottish-style stone buildings, overcast weather, and wet pavement, suggesting a location in Scotland. The large ornate building in the background appears to be a significant landmark, possibly a hotel or historic building. The road markings and traffic signs are consistent with UK standards. The overall atmosphere and architectural style point towards a major Scottish city. Given the prominence of the large building and the urban setting, this is likely Edinburgh, Scotland, United Kingdom."
FinalAnswer: "United Kingdom; Scotland; Edinburgh"

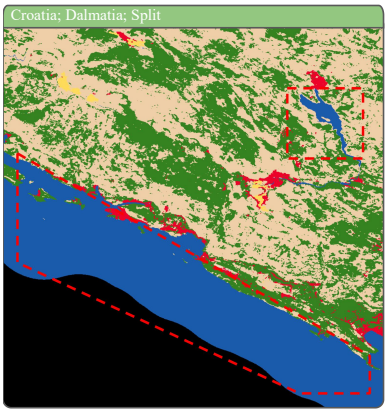
South Korea; Gyeonggi Province; Hwaseong



Answer

Clues: ["rectangular solar panel arrays", "irregularly shaped forested patches", "patchwork agricultural fields", "narrow dirt roads"]
Reasoning: "This is a satellite or aerial image showing a rural landscape with a large solar farm, agricultural fields, and a winding waterway adjacent to a larger body of water. The presence of a significant solar installation suggests a developed region with investment in renewable energy. The vegetation appears to be temperate deciduous forest, ruling out tropical or arid regions. The patchwork of fields and the specific layout of the solar farm are characteristic of certain regions in Europe or East Asia. The combination of a large solar farm, the specific shape of the forested areas, and the proximity to a coastline points towards a location in South Korea, where such solar installations are common and the landscape matches the visual evidence."
FinalAnswer: "South Korea; Gyeonggi Province; Hwaseong"

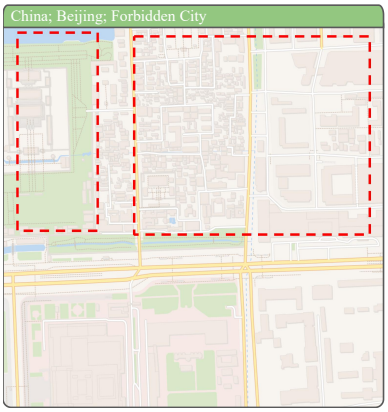
Croatia; Dalmatia; Split



Answer

Clues: ["coastal landmass with numerous islands", "long narrow bay", "distinctive lake-like water body inland", "green vegetation patches", "urbanized areas in red/yellow", "overall shape matches Croatia"]
Reasoning: "The image shows a coastal region with a long, narrow bay and numerous islands, which is characteristic of the Adriatic coast. The distinctive lake-like water body inland is Lake Vrana, located in Croatia. The green patches represent vegetation, while the red and yellow areas indicate urbanized regions. The overall shape and features strongly match the geography of Croatia, particularly the Dalmatian coast and the surrounding areas
FinalAnswer: "Croatia; Dalmatia; Split"

China; Beijing; Forbidden City

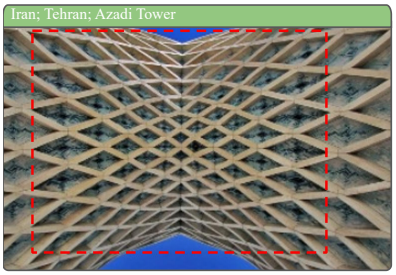


Answer

Clues: ["Large rectangular complex with central building and surrounding walls", "Wide boulevards and grid-like street pattern", "Green spaces and water features"]
Reasoning: "The image shows a large, rectangular complex with a central building and surrounding walls, which is characteristic of a major historical or cultural site. The wide boulevards and grid-like street pattern suggest a planned urban area, likely in a major city. The presence of green spaces and water features indicates a well-developed urban environment. The overall layout and design are consistent with a significant historical or cultural site in a major city, possibly in East Asia. The specific configuration of the complex and the surrounding urban area points to a location in Beijing, China, where such large-scale historical and cultural sites are common."
FinalAnswer: "China; Beijing; Forbidden City"

Figure 10: **Cases of continuous domains.**

Iran; Tehran; Azadi Tower



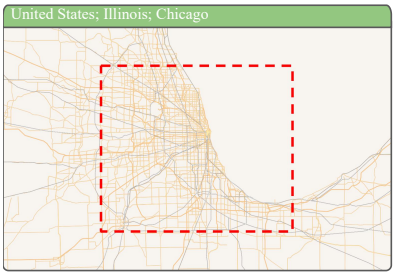
Answer

Clues: ["distinctive geometric lattice archway","Persian-style muqarnas-inspired design","sandstone-colored stone construction","architectural style typical of Iranian monumental buildings"]

Reasoning: "The image shows a highly stylized, geometric archway with intricate lattice patterns and muqarnas-like detailing, characteristic of Persian architecture. The use of sandstone-colored stone and the symmetrical, monumental scale suggest a major public or commemorative structure in Iran. This architectural style is most commonly found in Tehran, particularly around the Azadi Tower and surrounding areas. The specific design matches the interior or underside of the Azadi Tower's main arch, which is a well-known landmark in Tehran."

FinalAnswer: "Iran; Tehran; Azadi Tower"

United States; Illinois; Chicago



Answer

Clues: ["dense urban road network with grid pattern","major highway interchanges and expressways","suburban sprawl extending westward"]

Reasoning: "This is a road network map showing a major metropolitan area with a dense grid of streets, multiple highway interchanges, and a large body of water to the east. The distinctive curved coastline and the pattern of urban development extending westward from the waterfront are characteristic of Chicago, Illinois. The road network density and the presence of major highways like I-90 and I-55 match Chicago's infrastructure. The large body of water is Lake Michigan, and the curved coastline is the Chicago shoreline. The suburban sprawl extending westward is consistent with the Chicago metropolitan area."

FinalAnswer: "United States; Illinois; Chicago"

Australia; Victoria; Bendigo



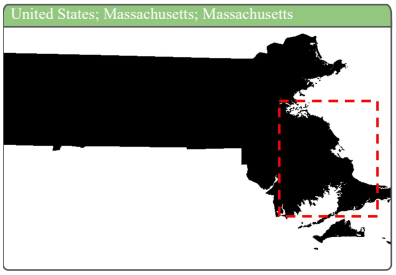
Answer

Clues: ["Bendigo Tramways sign","red and cream tram car","indoor museum or workshop setting","Australian-style tram design"]

Reasoning: "The image shows a vintage tram car with clear signage reading 'Bendigo Tramways' and 'Bendigo Pioneer Centre'. This directly identifies the location as Bendigo, Victoria, Australia. The tram is displayed indoors, likely within the Bendigo Tramways Museum, which is a known heritage site in the city. The tram's design and livery are consistent with early 20th-century Australian trams, further confirming the Australian context. Therefore, the specific location is Bendigo, Victoria, Australia."

FinalAnswer: "Australia; Victoria; Bendigo"

United States; Massachusetts; Massachusetts



Answer

Clues: ["administrative boundary silhouette","distinctive elongated shape with a narrow western end","coastal geography matching the northeastern United States"]

Reasoning: "This is an administrative boundary silhouette showing a state with a very distinctive shape. The long, narrow western portion tapers to a point, while the eastern section features a complex, irregular coastline with many islands and peninsulas. The most prominent feature is the large, hook-shaped peninsula extending eastward, which is characteristic of Cape Cod. This specific combination of a long mainland body and the Cape Cod peninsula is unique to the state of Massachusetts in the northeastern United States."

FinalAnswer: "United States; Massachusetts; Massachusetts"

United States; Washington; Seattle



Answer

Clues: ["complex network of waterways and islands","dense urban development along waterways","distinctive fjord-like inlets","coastal mountainous terrain"]

Reasoning: "This is a satellite image showing a complex network of waterways, islands, and dense urban development along the coast. The distinctive fjord-like inlets and the pattern of urbanization strongly suggest a location in the Pacific Northwest of the United States. The combination of deep-water inlets, numerous islands, and extensive urban development is characteristic of the Seattle metropolitan area. The specific arrangement of water bodies and landmasses matches the geography of Puget Sound and the surrounding areas in Washington state."

FinalAnswer: "United States; Washington; Seattle"

Thailand; Phuket; Phuket



Answer

Clues: ["tropical beachfront resort with palm trees and sandy beach","multi-story hotel buildings","large swimming pool with blue lounge chairs and white umbrellas"]

Reasoning: "This is an aerial view of a tropical beachfront resort. The presence of palm trees, sandy beach, and multi-story hotel buildings with brown roofs suggests a warm, humid climate typical of Southeast Asia or the Caribbean. The hilly terrain in the background and the specific layout of the resort with its large swimming pool and beach access point to a location in Southeast Asia, particularly Thailand. The architecture and landscaping are consistent with high-end resorts found in Phuket, Thailand. The combination of the beach, hills, and resort style strongly indicates this is a location in Phuket, Thailand."

FinalAnswer: "Thailand; Phuket; Phuket"

Figure 11: **Cases of discrete domains.**

Madagascar; Antananarivo; Analamanga



Answer

Clues: ["charred vegetation and blackened tree trunks", "dry, arid landscape with sparse vegetation", "paved road winding through hilly terrain"],
Reasoning: "The image shows a rural, arid landscape with clear signs of recent wildfire damage—charred trees and scorched earth. The terrain is hilly with distant mountains, and the road is unpaved or minimally maintained, suggesting a remote area. The vegetation type (sparse shrubs and scattered trees) and the dry, dusty ground are consistent with Mediterranean or semi-arid climates. The lack of urban features and the presence of a single road indicate a non-urbanized region. These conditions are typical of wildfire-prone areas in the western United States, particularly in states like California, Oregon, or Washington. Given the topography and vegetation, California is the most likely location, as it frequently experiences wildfires in such landscapes."
FinalAnswer: "United States; California; Sierra Nevada Foothills"

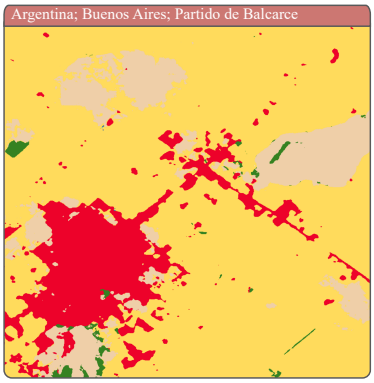
New Zealand; Southland; Southland



Answer

Clues: ["rectangular agricultural fields with hedgerows","dark coniferous forest patches","narrow winding waterways or drainage channels","overcast lighting typical of temperate climates"]
Reasoning: "The image displays a rural landscape characterized by a mix of agricultural fields, dense coniferous forests, and wetland areas. The presence of rectangular fields separated by hedgerows is a strong indicator of the British Isles, particularly England or Wales. The dark, boggy terrain suggests a peatland environment, which is common in northern England and Scotland. The small settlement with a grid-like layout and the overall topography point towards a specific region within the UK. The combination of these features—agricultural land, coniferous forests, and wetlands—is highly characteristic of the North Pennines or the Yorkshire Dales in northern England. The specific pattern of fields and the type of vegetation align well with areas in North Yorkshire."
FinalAnswer: "United Kingdom; England; North Yorkshire"

Argentina; Buenos Aires; Partido de Balcarce



Answer

Clues: ["Esri land-cover classification map", "large red built-up area", "yellow background (water)", "green patches (trees)", "beige patches (bare ground)", "irregular coastline", "long narrow landform extending eastward"]
Reasoning: "This is an Esri land-cover classification map showing a large built-up area (red) surrounded by water (yellow), with scattered vegetation (green) and bare ground (beige). The shape of the landmasses, particularly the large central urban area and the long, narrow peninsula extending to the east, strongly resembles the geography of the Arabian Peninsula. The configuration matches the urban sprawl of Riyadh, Saudi Arabia, with the Red Sea to the west and the Rub' al Khali desert to the east. The long eastern extension corresponds to the Diriyah and surrounding areas. The color scheme and pattern are consistent with Esri's land-cover classification for urban areas in arid regions."
FinalAnswer: "Saudi Arabia; Riyadh; Riyadh "

United Kingdom; Scotland; Edinburgh



Answer

Clues: ["administrative boundary silhouette","distinctive coastline with a large northern protrusion","narrow eastern corridor connecting to a larger landmass","coastline shape matches the Iberian Peninsula"]
Reasoning: "The image is an administrative boundary silhouette showing a landmass with a distinctive coastline. The overall shape, including the large northern protrusion and the narrow eastern corridor, matches the Iberian Peninsula. The specific configuration of the coastlines and the relative positions of the landmasses confirm this is Spain and Portugal. The narrow eastern corridor is the Algarve region of Portugal, and the large northern protrusion is the northern part of Spain. The overall shape and coastline features are consistent with the Iberian Peninsula, specifically Spain and Portugal."
FinalAnswer: "Spain; Andalusia; Almería"

Figure 12: **Failure cases of continuous domains.**

Philippines; Ifugao; Batad



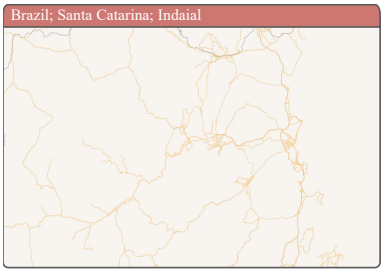
Answer

Clues: ["extensive agricultural terraces on steep mountain slopes","lush green vegetation and tropical/subtropical climate","no modern infrastructure or urban development in immediate view"]

Reasoning: "The image displays a classic example of ancient agricultural terracing on steep mountainous terrain, which is characteristic of pre-Columbian civilizations in the Andes region. The terracing pattern, with its stepped design following the contours of the mountains, is typical of Inca or pre-Inca agricultural systems. The combination of extensive terracing, mountainous terrain, and tropical climate points to the Cusco region of Peru."

FinalAnswer: "Peru; Cusco Region; Ollantaytambo"

Brazil; Santa Catarina; Indaial



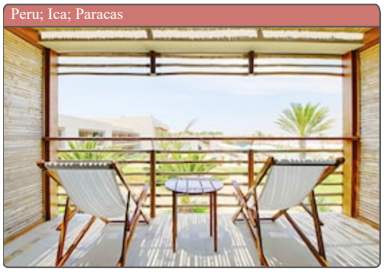
Answer

Clues: ["river network with major confluence","dense road grid in northeast corner","distinctive river delta pattern","administrative boundary lines"]

Reasoning: "This is a vector map showing a river network and road infrastructure. The key features are a major river system with a prominent confluence and a dense urban area in the northeast corner. The river pattern and coastal location strongly suggest a region in Southeast Asia. The specific configuration of the river network and the urban cluster in the northeast matches the geography of Thailand, particularly the area around Bangkok and the Chao Phraya River delta. The road network density and the shape of the urban area are consistent with Bangkok and its surrounding provinces. The administrative boundaries further confirm this is a region within Thailand."

FinalAnswer: "Thailand; Bangkok; Bangkok"

Peru; Ica; Paracas



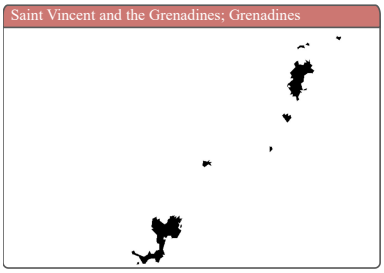
Answer

Clues: ["tropical vegetation including palm trees","modern resort-style architecture with bamboo walls and ceiling","overcast sky suggesting coastal or island environment"]

Reasoning: "The image shows a tropical resort setting with palm trees, bamboo construction, and a coastal view, which is characteristic of luxury vacation destinations in warm climates. The architectural style and vegetation suggest a location in Southeast Asia or the Caribbean. Given the specific combination of bamboo architecture, tropical vegetation, and resort design, this is most likely a location in Thailand, known for its tropical resorts and bamboo construction. The specific location is likely a resort in Phuket or Krabi, which are famous for such settings."

FinalAnswer: "Thailand; Phuket; Phuket"

Saint Vincent and the Grenadines; Grenadines



Answer

Clues: ["long north-south chain of islands","distinctive large northern island with complex coastline","multiple smaller islands scattered along the chain"]

Reasoning: "The image is an administrative boundary silhouette of an archipelago. The overall shape is a long, narrow chain of islands oriented north-south. There are numerous smaller islands scattered along the chain, particularly in the southern half. This pattern is characteristic of the Philippines, which is an archipelago of over 7,000 islands. The specific configuration of the main islands (Luzon, Visayas, Mindanao) and the surrounding smaller islands matches the Philippines' geography. The silhouette clearly shows the elongated shape of the country from north to south, with the major islands positioned correctly relative to each other. "

FinalAnswer: "Philippines; Philippines; Philippines"

Colombia; La Guajira; Uribia



Answer

Clues: ["distinctive peninsula shape with a large inland lake","arid terrain with mountainous eastern edge","coastline features including a prominent bay and headland]

Reasoning: "The image shows a distinctive peninsula with a large inland lake, arid terrain, and mountainous eastern edge. These features are characteristic of the Australian state of Queensland, specifically the area around the Great Sandy Desert and the coastal regions. The prominent bay and headland on the western side, along with the overall shape, strongly suggest this is the northern part of Queensland, near the border with the Northern Territory. The cloud patterns over the eastern mountains further confirm this location, as they are typical of the region's climate and topography."

FinalAnswer: "Australia; Queensland; Northern Queensland"

New Zealand; Canterbury; Kaikōura



Answer

Clues: ["black volcanic sand beach","turbulent turquoise ocean water","coastal erosion patterns","aerial view of shoreline"]

Reasoning: "The image displays a distinctive black sand beach adjacent to turquoise ocean water, which is a hallmark of volcanic coastal environments. This combination is most famously found in Iceland, particularly along the south coast where glaciers meet the Atlantic Ocean. The specific color contrast and wave action are characteristic of Reynisfjara, a well-known black sand beach in Iceland. The aerial perspective confirms the unique geological features of this location."

FinalAnswer: "Iceland; South Region; Reynisfjara"

Figure 13: **Failure cases of discrete domains.**