

HoloGeo: Mitigating Landmark Bias in Geo-localization via Evidence-Driven Reasoning

Pengcheng Zhou
National University of Singapore
Singapore, Singapore
e1554357@u.nus.edu

Xuanyu Liu
Shandong University of Science and
Technology
Qingdao, China
202211081013@sdust.edu.cn

Yanchen Yin
Shandong University of Science and
Technology
Qingdao, China
yyc@sdust.edu.cn

Bobo Li
National University of Singapore
Singapore, Singapore
libobo@nus.edu.sg

Shengqiong Wu*
University of Oxford
Oxford, United Kingdom
shengqiongwu@gmail.com

Mong-Li Lee
National University of Singapore
Singapore, Singapore
dcsleml@nus.edu.sg

Wynne Hsu
National University of Singapore
Singapore, Singapore
dcshsuw@nus.edu.sg

Abstract

Recent advances in Vision-Language Models (VLMs) have significantly improved image geo-localization, yet existing models remain susceptible to **landmark bias**, causing them to overlook geographical cues or form spurious correlations, ultimately resulting in inaccurate localization. To systematically investigate this issue, we first design two quantitative metrics, **Bias Intensity (BI)** and **Bias Harmfulness (BH)**, to characterize the impact of landmarks exerted on model reasoning, and establish a comprehensive benchmark, **LandmarkBias-3K**. To mitigate landmark bias, we further propose an evidence-driven reasoning framework, **HoloGeo**, to improve the reliability of geo-localization. HoloGeo is supported by a high-quality dataset, **BF-30k**, annotated with structured multi-evidence bias-free reasoning chains. By incorporating multi-dimensional rewards, HoloGeo explicitly encourages balanced attention over diverse visual cues and achieves evidence-driven joint reasoning. Extensive experiments demonstrate that HoloGeo not only maintains excellent performance on IM2GPS3K and YFCC4k but also significantly outperforms existing VLMs on LandmarkBias-3K, validating its effectiveness for robust geospatial reasoning. Project page: <https://hologeo.github.io/>.

Keywords

Geo-localization, Large Vision-Language Models, Visual Bias, Multimodal Reasoning, Reinforcement Learning

1 Introduction

Image geo-localization aims to infer geographic coordinates or specific shooting locations from visual content, and plays a crucial role in a wide range of applications such as autonomous navigation, social media geotagging, and crisis response [5, 16, 51, 53]. Compared to traditional classification- or retrieval-based methods that rely on fixed grid partitioning or large-scale reference databases [25, 31, 42],

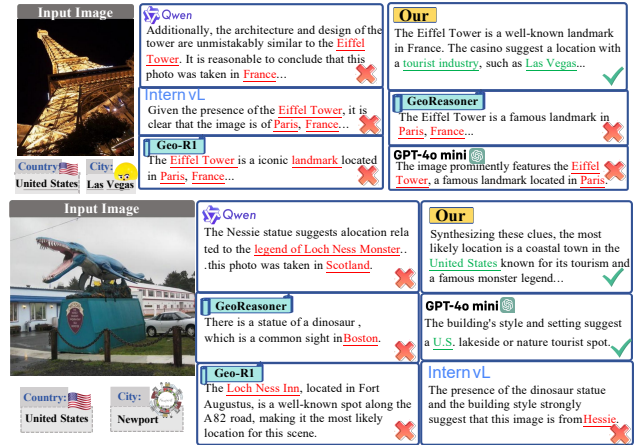


Figure 1: Illustration of landmark bias in geospatial reasoning. Existing models over-rely on salient landmarks (e.g., Eiffel Tower-like structure, Loch Ness Monster statue) and ignore complementary cues, leading to erroneous localization (actual locations: Las Vegas, USA; Newport, USA).

Large Vision-Language Models (VLMs) [2, 7] have shown remarkable progress in geo-localization by leveraging strong cross-modal understanding and integrated world knowledge. Beyond predicting final locations, emerging systems [28, 33, 57, 67] further attempt to generate interpretable reasoning trajectories, typically by identifying salient landmarks (e.g., iconic buildings) and then mapping them to geographic hypotheses, moving from black-box prediction toward explainable decision-making.

Despite progress, existing VLM-based methods still face a critical challenge: **landmark bias**, in which models may over-rely on visually salient landmark-like cues and become over-confident even when such cues are ambiguous, replicated, or weakly informative. This bias is amplified by the strong saliency and spurious correlations learned from web-scale data, causing the model to treat a

*Corresponding author.

single landmark as sufficient evidence while under-utilizing complementary geographic cues, as the *Eiffel Tower* or the *Loch Ness* in Figure 1. Importantly, recent reasoning-oriented approaches [31, 33, 67] still lack explicit evidence reliability calibration and mechanisms to discourage single-cue dominance; consequently, generated rationales may be fluent yet weakly grounded, failing to guarantee reliable localization. Moreover, conventional evaluations largely focus on top-line accuracy, leaving it unclear *how strongly* landmarks dominate the reasoning process and *how harmful* this dominance becomes under landmark ambiguity. This highlights the need for more fine-grained diagnostic tools to systematically analyze and address landmark bias.

To enable systematic investigation of landmark bias, we first formalize this phenomenon from a reasoning perspective. Formally, landmark bias refers to the tendency of models to over-rely on visually salient landmark cues during geospatial reasoning by assigning disproportionate weight to such cues at the expense of underutilizing complementary geographic features, thereby leading to over-confident yet erroneous predictions when landmarks are ambiguous, replicated, or misleading. Grounded in this formulation, we propose two quantitative metrics, **Bias Intensity (BI)** and **Bias Harmfulness (BH)**, to measure the dominance of landmark evidence and its adverse impact under ambiguity. BI measures the degree to which landmarks anchor model predictions by evaluating the confidence shift between original and landmark-occluded images; BH measures the extent to which landmarks mislead the model’s preference for the ground truth. To facilitate fine-grained evaluation, we introduce **LandmarkBias-3K**, a challenging benchmark specifically curated with scenarios where landmark cues are insufficient or deceptive, thereby exposing the inherent vulnerabilities in biased reasoning behaviors.

To mitigate landmark bias, we further propose **HoloGeo**, a novel evidence-driven framework that explicitly enforces holistic multi-cue aggregation for geo-localization rather than landmark-centric shortcuts. Firstly, we construct a high-quality multi-evidence reasoning dataset, **BF-30K**, by decomposing images into semantically meaningful regions, organizing geographically grounded analyses into structured multi-evidence reasoning chains, and further refining it through cross-model verification and manual inspection to reduce reasoning errors and hallucinated explanations [18, 24, 73]. Building on this dataset, HoloGeo incorporates multi-dimensional rewards over *prediction correctness* and *evidence coverage*, explicitly encouraging balanced attention across heterogeneous geographic cues. Through this design, the model learns to jointly evaluate diverse evidence, align intermediate reasoning with final predictions, and avoid over-reliance on isolated landmark cues.

Extensive experiments on standard benchmarks and custom bias benchmarks verify the effectiveness of the proposed method. Existing general-purpose vision-language models and mainstream geo-localization methods perform poorly on LandmarkBias-3K, with most achieving below 20% city-level accuracy; even the state-of-the-art domain-specific model only reaches 23.57%. In contrast, our proposed HoloGeo attains 27.27% city-level accuracy on this dataset while maintaining strong performance on conventional geo-localization benchmarks including IM2GPS [22], IM2GPS3K [55], and YFCC4K [55], which fully demonstrates its effectiveness. Through saliency map visualization, we find that existing models suffer from

landmark bias due to over-reliance on salient landmarks and neglect of broader geographic context, while HoloGeo mitigates this issue via balanced attention distribution. Finally, this validates its capability to effectively mitigate landmark bias while enhancing the robustness of geospatial reasoning.

In all, our contribution can be summarized in threefold:

- We present the first systematic study on landmark bias in geospatial reasoning, along with a quantitative metric suite (BI and BH) for fine-grained analysis.
- We introduce LandmarkBias-3K and BF-30K, the first benchmark and training dataset specifically designed to evaluate and mitigate landmark bias with structured multi-evidence reasoning for geo-localization tasks.
- We propose HoloGeo, an evidence-driven framework, which encourages balanced multi-cue reasoning through multi-dimensional reward design, effectively reducing landmark bias while preserving strong geo-localization performance.

2 Related Work

Image Geo-localization. Image geo-localization aims to predict the geographic location of a given image and has been widely studied in urban analysis [28, 68, 72], individual trajectories [9], and geospatial data mining [17, 35, 46, 64]. Early approaches primarily rely on *retrieval-based* methods [19, 20, 53, 58, 66, 69, 77], which match query images against large-scale reference databases, and *classification-based* methods [13, 42, 48, 50, 62], which discretize the earth into predefined regions [22, 61]. While effective, some research [25, 31, 75] has demonstrated that retrieval-based methods require a large reference database, and classification-based approaches are limited by coarse spatial granularity. To overcome these limitations, recent works explore *generation-based* approaches [15, 25, 75] that directly predict geographic coordinates or location descriptions, improving flexibility and scalability. With the emergence of large vision-language models (VLMs) [1, 2, 76], geo-localization has further benefited from strong cross-modal understanding and embedded world knowledge, enabling more generalizable and semantically rich predictions [28, 39]. More recently, researchers [8, 31, 41] have focused on enhancing the *reasoning capabilities* of VLM-based geo-localization systems. Techniques such as chain-of-thought prompting [32, 57, 70], agent-based reasoning [4, 21], and reinforcement learning [33, 67, 72] have been introduced to generate interpretable reasoning trajectories and improve decision quality. However, despite improved reasoning interpretability, existing methods still lack mechanisms to ensure *balanced and reliable evidence utilization*. In particular, they tend to over-rely on visually salient landmark cues while under-utilizing complementary geographic signals, leading to biased and over-confident predictions. This limitation motivates our work on systematically analyzing and mitigating landmark bias in geo-localization.

Bias and Robustness in VLMs. Recent studies have revealed that VLMs are prone to various forms of bias and unreliable reasoning behaviors [54]. In particular, models often rely on *spurious correlations* or dataset-specific shortcuts, leading to overconfident predictions that are not well grounded in visual evidence. Such issues have been extensively studied in the context of hallucination [37, 47],

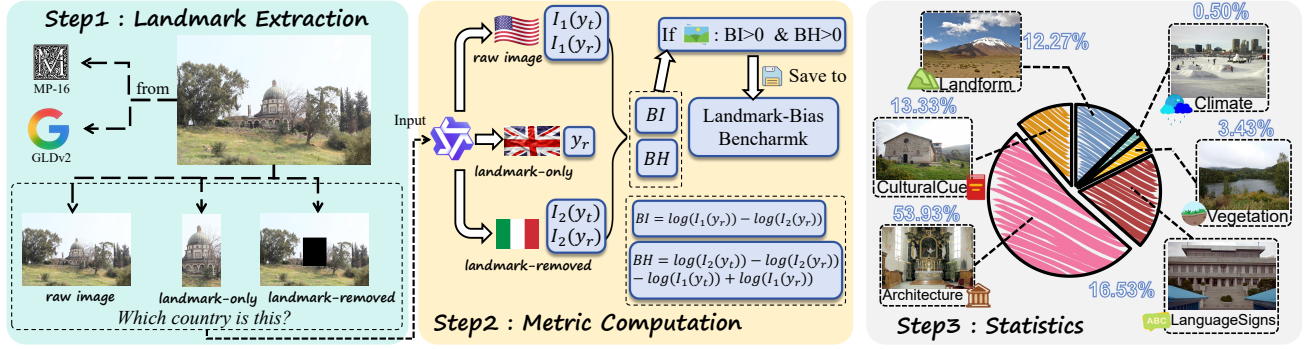


Figure 2: Pipeline for constructing LandmarkBias-3K. For each image, landmark regions are first extracted to derive landmark-only and landmark-removed views. Model probabilities under the three views are then used to compute Bias Intensity (BI) and Bias Harmfulness (BH), measuring the strength of landmark influence and its harmful effect on the ground-truth prediction, respectively. Samples with BI and BH above predefined thresholds are retained to form LandmarkBias-3K. The rightmost sub-figure shows the landmark category distribution of the benchmark.

shortcut learning [3, 11], and visual bias [54], where models generate plausible yet incorrect outputs due to over-reliance on partial or misleading cues. Several works [6, 30, 65] have attempted to improve robustness by encouraging better alignment between visual inputs and generated outputs, for example, through improved training objectives, data curation, or reasoning supervision. However, these approaches primarily focus on general-purpose robustness or hallucination mitigation, without explicitly modeling how different types of visual evidence contribute to the final decision. In the context of geo-localization, this limitation becomes particularly critical. Existing VLM-based approaches often rely heavily on salient landmark cues during reasoning, yet the dominance of such cues and their potential to mislead predictions under ambiguity have not been systematically studied. Moreover, prior work lacks quantitative metrics to measure the extent of such bias, as well as dedicated benchmarks to evaluate its impact on reasoning reliability. In contrast, we provide a principled formulation and introduce quantitative metrics to measure both the strength and harmfulness of landmark-induced bias, along with a dedicated benchmark and mitigation framework.

3 Landmark-Bias Benchmark

Our goal is to characterize not only whether landmark cues influence model predictions, but more importantly, whether such influence leads to *harmful* geospatial reasoning errors. However, not all salient landmark attraction is detrimental. Thus, a key challenge lies in distinguishing between benign reliance on informative cues and biased over-reliance that misleads prediction. To address this, we design two complementary metrics that capture both the strength of landmark influence and its adverse impact on prediction correctness. Based on the two metrics, we further construct a challenging benchmark that focuses on cases where landmark cues are highly dominant but unreliable.

3.1 Quantitation Metrics

Given an input raw image x , we first detect its landmark region r using GroundingDINO [38]. We then construct a landmark-removed

image x_r by masking out the detected landmark region. To quantify the influence of landmark cues on geo-localization, we evaluate the model under three complementary input conditions.

For a location label y , we first define

$$I_1(y) = P(y | x), \quad I_2(y) = P(y | x_r), \quad I_3(y) = P(y | r), \quad (1)$$

where $I_1(y)$ is the model’s predicted probability for label y given the full image, $I_2(y)$ is the probability given the landmark-removed image, and $I_3(y)$ is the probability given only the landmark region. This formulation allows us to isolate the contribution of landmark cues from the remaining contextual evidence. Then, we define y_t as the ground-truth location label, and y_r as the *landmark-induced label* prediction obtained from the landmark region alone:

$$y_r = \arg \max_y I_3(y) = \arg \max_y P(y | r), \quad (2)$$

Building on these definitions, we quantify landmark-induced bias from two complementary perspectives: *intensity* and *harmlessness*. The corresponding metrics are defined as follows:

Bias Intensity (BI). Bias intensity measures how strongly the presence of the landmark increases the model’s preference for the landmark-induced label y_r :

$$BI = \log I_1(y_r) - \log I_2(y_r). \quad (3)$$

A positive BI indicates that the landmark increases the model’s confidence in y_r , suggesting a strong anchoring effect. Values close to zero imply limited influence, while $BI < 0$ indicates that the landmark suppresses support for y_r .

Bias Harmfulness (BH). Bias harmfulness measures whether the landmark shifts the model away from the ground-truth label y_t and toward the landmark-induced label y_r :

$$BH = (\log I_2(y_t) - \log I_2(y_r)) - (\log I_1(y_t) - \log I_1(y_r)). \quad (4)$$

where $\log I_1(y_t) - \log I_1(y_r)$ reflects the relative advantage of the true label over the landmark-induced label when the full image is observed, while $\log I_2(y_t) - \log I_2(y_r)$ measures the same advantage after removing the landmark. A positive BH means that the landmark reduces the model’s relative preference for the true label and therefore has a more harmful effect on correct geo-localization.

To ensure a stable probability estimation, Qwen2.5-VL-72B [2] is adopted for all probability estimates of $I_1(\cdot)$, $I_2(\cdot)$, and $I_3(\cdot)$. Additionally, we employ InternVL3-78B [76] for cross-validation to alleviate potential biases inherent in a single large vision-language model. The results are analyzed in the appendix G.

3.2 Benchmark Construction

Built upon the proposed metrics, we systematically select instances from two large-scale geo-localization datasets, MP-16 [29] and GLDv2 [61], to construct a dedicated benchmark that explicitly targets landmark-induced bias in real-world scenarios. As shown in Figure 2, for each image, we first detect its landmark region and construct the landmark-only and landmark-removed views. Then, we compute the corresponding BI and BH scores, and retain samples with positive BI and BH values that exceed predefined thresholds. The detailed threshold settings and selection criteria are provided in Appendix B. Finally, this procedure yields **3,000** images spanning diverse landmark types (e.g., *landform*, *architecture*, *climate*, and *sings*), forming the proposed *LandmarkBias-3K* benchmark. Additional dataset statistics are provided in Appendix B.

4 Methodology: HoloGeo

To mitigate landmark bias and enhance localization robustness, we propose **HoloGeo**, a unified evidence-driven reasoning framework. Technically, we construct the high-quality multi-evidence reasoning-driven dataset, **BF-30k**, and adopt a two-stage learning paradigm (Figure 3) to guide the model in learning geospatial reasoning patterns, encouraging balanced attention across diverse visual cues and reducing over-reliance on misleading landmarks.

4.1 BF-30K Dataset Construction

Recent works [33, 67, 72] have explored annotating intermediate reasoning processes to improve model interpretability and performance. However, despite considering multiple aspects of an image (e.g., textual cues or scene attributes), the reasoning process may be implicitly anchored to initially salient but potentially misleading cues, introducing early-stage bias and leading to incorrect predictions. To address this issue, we argue that geospatial reasoning should be grounded in *region-level evidence decomposition* followed by global aggregation [50, 62]. Instead of centering reasoning around a single dominant cue, we propose to construct *structured multi-evidence reasoning chains*, where diverse geographic elements are first analyzed independently and then integrated into a coherent global inference, establishing a complete logical pathway from local visual cues to final localization. To this end, we design a rigorous construction pipeline, shown in Figure 3, consisting of three stages: data filtering, multi-VLM reasoning annotation, and multi-dimensional validation.

Data Filtering. Large-scale web-crawled datasets provide diverse visual data for geo-localization but often suffer from significant noise [34, 59, 60], such as close-up views with insufficient context, generic objects without discriminative localization cues, and images dominated by repetitive or uninformative patterns. To ensure high-quality training data, we design a filtering pipeline to retain samples with reliable geographic signals while preserving diverse landmark bias scenarios. Following the previous work [4],

we first adopt GeoCLIP [53] to filter out geographically indistinguishable images from the MP-16 [29] dataset, ensuring that all selected samples contain valid geographic cues for geolocation. For GLDv2 [61], we extract the GPS coordinates of each sample via the Wikimedia API [63]. We then leverage the geocoder Python library to reversely map these coordinates to the corresponding city and country information. We then categorize the filtered samples into three groups: *landmark-biased*, *landmark-containing*, and *non-landmark*. For landmark-containing images, we further compute BI and BH scores to identify instances exhibiting strong landmark bias characteristics. This results in a balanced dataset covering diverse visual conditions and bias scenarios.

Multi-VLM Reasoning Annotation. At this stage, we employ GroundingDINO[38] to detect visual elements in pre-filtered images, automatically identifying bounding boxes for discriminative key regions, thereby providing precise localization for subsequent fine-grained reasoning. We then use Qwen2.5-VL-72B[2] to filter the detected boxes, describing and analyzing each box independently with contextual information to ensure only highly localization-relevant boxes are retained. Furthermore, to improve the reliability of geographic element analysis, inspired by recent visual reasoning studies [10, 40, 71], we feed both the original image and cropped patches of filtered bounding boxes into VLMs for fine-grained annotation, where Qwen2.5-VL-72B and ChatGPT-o3 [45] serve as our annotation models. Specifically, we guide the model to complete annotation through a standardized reasoning process: in the **<Analyze>** stage, the model is required to independently parse the geographic visual content within each valid bounding box for fine-grained feature analysis. In the **<Think>** stage, the model aggregates feature information from all bounding boxes, resolves potential ambiguities and contradictions, thereby constructing a unified geographic feature evidence system. Based on this evidence chain, the model performs hierarchical reasoning, gradually narrowing the scope from continent and country to the target city. Finally, it outputs the final result in the **<Answer>** stage, forming a coherent logical connection from local visual cues to global geographic localization. This design enforces structured multi-evidence reasoning and mitigates over-reliance on single salient cues, providing reliable supervision for geo-localization.

Multi-Dimensional Validation. To ensure data quality, mitigate reasoning trajectory errors, and prevent bias propagation from a single model[18, 56], we design a multi-dimensional validation process. We leverage InternVL3-78B [76] as an independent reviewer to perform cross-model verification of the annotated data. By enforcing visual-semantic consistency across different models, we filter out samples exhibiting hallucinations or logical inconsistencies, thereby improving the reliability of the reasoning chains. In addition, we conduct manual spot checks to further identify and remove samples with annotation errors or reasoning contradictions. More details are provided in Appendix C. The resulting dataset is thus carefully curated to support robust and trustworthy training.

4.2 Warm-up with SFT

To equip the model with preliminary structured geospatial reasoning ability, avoid cold-start issues, and establish a solid foundation for the subsequent reinforcement learning stage [12, 18, 74], we

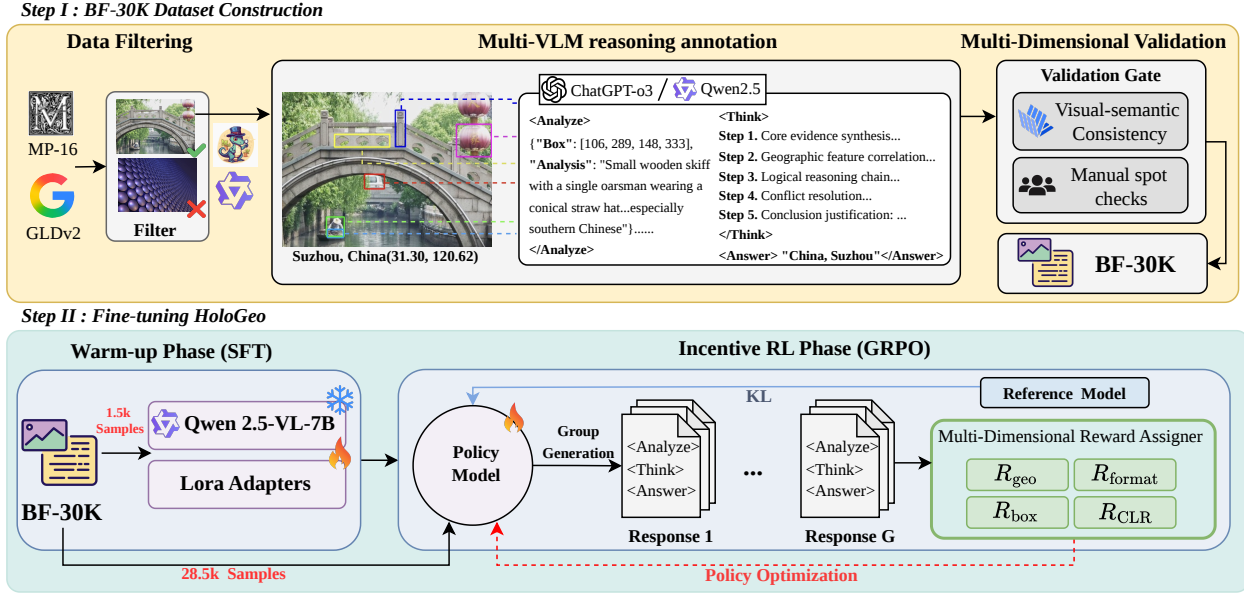


Figure 3: Overview of the proposed HoloGeo. Stage I constructs the BF-30K dataset by filtering samples from MP-16 [29] and GLDv2 [61], annotating structured multi-evidence reasoning chains, and validating them via cross-model visual-semantic consistency and manual checks. In Stage II, built upon Qwen2.5-VL-7B [2], HoloGeo adopts a two-stage learning paradigm: a warm-up phase with SFT, followed by GRPO-based reinforcement learning with rewards for geo-location accuracy (R_{geo}), visual evidence grounding (R_{box}), and comprehensive logical reasoning (R_{CLR}).

perform SFT using the LoRA strategy [23] for model warm-up. Specifically, we randomly sample 1.5K instances from the BF-30K dataset and fine-tune the pre-trained Qwen2.5-VL-7B model.

4.3 Scaffolding with GRPO

Drawing inspiration from [18], we apply GRPO-based RL optimization to further empower geospatial reasoning capability and enhance its generalization [12]. The pivot part of this stage lies in the design of reward functions, which guide the model toward balanced and reliable reasoning. Beyond standard format R_{format} and accuracy-based rewards R_{geo} , we introduce additional objectives to address landmark bias. Specifically, to discourage overreliance on salient regions and encourage broader evidence utilization, we design a *visual evidence grounding reward* R_{box} . In addition, to ensure the correctness of the reasoning process itself, we introduce a *comprehensive logical reasoning reward* R_{CLR} that evaluates the consistency and validity of the generated reasoning chains.

Geo-localization Accuracy Reward R_{geo} . This reward evaluates the matching degree between the model’s predicted result $\hat{g}_i = (\hat{c}_i, \hat{t}_i)$ and the ground-truth geographic location $g_i = (c_i, t_i)$, where c_i and t_i denote the country and city, respectively. To account for the hierarchical structure of geographic labels, we adopt a hierarchical scoring scheme defined as:

$$R_{geo}(\hat{g}_i, g_i) = \mathbb{I}[\hat{c}_i = c_i] \cdot (\alpha \cdot \mathbb{I}[\hat{t}_i = t_i] + (1 - \alpha)) \quad (5)$$

where $\mathbb{I}[\cdot]$ is the indicator function and $\alpha \in [0, 1]$ controls the weight of city-level accuracy. The reward is 0 if the country prediction is incorrect. If the country is correct but the city is not, a partial reward of $1 - \alpha$ is assigned. The full reward of 1.0 is obtained only when both country and city predictions are correct.

This design encourages the model to first capture coarse-grained geographic information before refining fine-grained localization. In our experiments, we set $\alpha = 0.8$.

Format Reward R_{format} . This reward ensures the model’s output contains the $\langle \text{Analyze} \rangle$, $\langle \text{Think} \rangle$, and $\langle \text{Answer} \rangle$ structure: a score of 1.0 is awarded for compliance, and 0 otherwise.

Visual Evidence Grounding Reward R_{box} . This reward measures how well the model grounds its reasoning in relevant visual evidence by evaluating the alignment between predicted and ground-truth bounding boxes, so as to encourage the model to cover diverse geographic cues rather than over-relying on a single salient region. For the i -th sample, let $\hat{B}_i$ denote the set of predicted bounding boxes parsed by the model from its generation module, and B_i denote the set of ground-truth annotated boxes for key regions corresponding to the image. We adopt an IoU-based matching criterion with threshold τ_{IoU} . Specifically, for each ground-truth box $b_k \in B_i$, we compute its maximum overlap with predicted boxes. A ground-truth box is considered successfully covered if the maximum IoU exceeds τ_{IoU} . The reward is defined as:

$$R_{box} = \frac{1}{|B_i|} \sum_{b_k \in B_i} \mathbb{I} \left(\max_{b_j \in \hat{B}_i} \text{IoU}(b_j, b_k) \geq \tau_{IoU} \right), \quad (6)$$

where $\mathbb{I}(\cdot)$ is the indicator function. This reward score is positively correlated with the ratio of successfully matched ground-truth boxes, encouraging the model to accurately locate and cover key geographic regions in images.

Comprehensive Logical Reasoning Reward R_{CLR} . This reward evaluates the validity and reliability of the model’s reasoning process, encouraging consistent, well-grounded, and multi-evidence

Table 1: Comparison with existing geo-localization models. Evaluation on IM2GPS, IM2GPS3k, and YFCC4k benchmarks. Values indicate accuracy (%) at different distance thresholds.

Model	IM2GPS			IM2GPS3k			YFCC4k		
	City	Region	Country	City	Region	Country	City	Region	Country
• Traditional Domain-Specific Model									
PlaNet [62]	24.5	37.6	53.6	24.8	34.3	48.4	-	-	-
CPlaNet [50]	37.1	46.4	62.0	26.5	34.6	48.6	-	-	-
ISNs[42]	43.0	51.9	66.7	28.0	36.6	49.7	-	-	-
TransLocator[48]	48.1	64.6	75.6	31.1	46.7	58.9	18.6	27.0	41.1
GeoCLIP [53]	41.8	60.8	77.2	34.5	50.7	69.7	-	-	-
GeoDecoder [13]	50.2	69.0	80.0	33.5	45.9	61.0	-	-	-
PIGEON [20]	40.9	63.3	82.3	36.7	53.8	72.4	-	-	-
• General VLM Model									
InternVL2-8B [7]	0.8	3.0	3.8	4.2	6.8	9.4	0.4	0.9	2.0
LLaVA-Next-Mistral-7B [36]	7.6	8.4	11.4	4.1	5.1	7.4	2.0	2.7	4.7
Qwen2.5-VL-7B [2]	35.0	46.8	59.1	27.9	38.9	49.9	11.5	16.6	24.1
• VLM-based Domain-Specific Model									
GeoReasoner [31]	24.9	48.1	65.8	26.5	40.4	57.7	6.3	10.4	17.3
GaGA [14]	38.8	54.8	75.1	33.0	48.0	67.1	-	-	-
GAEA [4]	43.0	57.4	77.2	36.9	56.0	73.2	-	-	-
Geo-R1 [67]	42.2	59.5	77.2	35.4	52.3	69.7	17.0	29.0	48.9
GeoChat [28]	38.8	59.1	76.8	34.7	51.7	69.4	17.0	29.2	49.8
GLOBE [33]	44.3	59.4	76.3	40.2	56.2	71.5	18.0	30.7	50.6
HoloGeo-SFT(Ours)	40.9	56.5	73.4	36.9	52.3	70.8	18.1	27.0	44.5
HoloGeo(Ours)	47.3	60.3	76.8	38.5	53.7	70.8	18.9	31.7	51.5

reasoning. We employ Qwen3.5-35B-A3B [49] as a reward model to score each sample pair $(I_i, \hat{r}_i)$, where I_i is the input image and $\hat{r}_i$ is the generated reasoning trajectory. Specifically, the reward model evaluates the reasoning according to the structured three-stage process. 1) **<Analyze> Visual Factuality**: The described visual attributes must be consistent with the content of the corresponding bounding box regions. 2) **<Think> Logical Consistency**: The reasoning process should be coherent and avoid over-reliance on a single geographic cue, unless no other informative evidence is available. 3) **<Answer> Answer Specificity**: The predicted location should be expressed as specific country and city names, without vague or ambiguous descriptions. The reward model outputs the probabilities of yes (valid reasoning) and no (invalid reasoning). We define the reward as the normalized probability of valid reasoning:

$$R_{\text{CLR}} = \frac{\mathbb{P}(\text{yes} \mid I_i, \hat{r}_i; \theta_{\text{reason}})}{\mathbb{P}(\text{yes} \mid I_i, \hat{r}_i; \theta_{\text{reason}}) + \mathbb{P}(\text{no} \mid I_i, \hat{r}_i; \theta_{\text{reason}})}, \quad (7)$$

where θ_{reason} denotes the predictive distribution of the reward model, the numerator is the probability that the reasoning is valid, and the denominator is the sum of the probabilities of validity and invalidity. This score serves not only as a soft metric for reasoning validity, but also guides the model to realize visual fact-based multi-element geographic reasoning.

Finally, the total reward is:

$$r_i = R_{\text{format}} + \lambda_1 R_{\text{geo}} + \lambda_2 R_{\text{box}} + \lambda_3 R_{\text{CLR}}, \quad (8)$$

where we set the reward weights to $\lambda_1 = 1.2$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.5$. **GRPO-based Reinforcement Learning**. Based on the reward signals defined above, we adopt the GRPO [18] to fine-tune the model π_θ optimized via supervised fine-tuning in Section 4.2 on the BF-30K (D_{r_i}) dataset. GRPO introduces an intra-group normalization mechanism that optimizes the relative preferences A_i among candidate outputs $O = \{o_i\}$ for each input sample I , thereby enhancing

robustness against fluctuations in reward scales. A KL-divergence regularization term $\mathbb{D}_{\text{KL}}$ is introduced to prevent the optimized policy π_θ from diverging excessively from the reference model π_{ref} :

$$\max_{\pi_\theta} \mathbb{E}_{[I \sim D_{r_i}, \{o_i\}_{i=1}^G \sim \pi_\theta(O|I)]} \left[\frac{1}{G} \sum_{i=1}^G \min \left(\frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)} A_i, \right. \right. \\ \left. \left. \text{clip} \left(\frac{\pi_\theta(o_i)}{\pi_{\theta_{\text{old}}}(o_i)}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right], \quad (9)$$

where ϵ, β are the hyper-parameters.

5 Experiments

5.1 Setups

Datasets. We evaluate all models on four geo-localization benchmarks: IM2GPS [22], IM2GPS3K, [55] YFCC4K [55], and our proposed **Landmark-Bias-3K Benchmark**. IM2GPS and IM2GPS3K are widely used benchmarks that primarily consist of landmark-centric images, while YFCC4K contains more diverse scenes. In contrast, the Landmark-Bias-3K benchmark is explicitly curated to highlight landmark-induced bias.

Benchmark Models. We evaluate a diverse set of geo-localization models, including (i) classical image-based geo-localization methods, (ii) recent vision language models, and (iii) reasoning-oriented geo-localization models. For classical methods, we include PlaNet[62], CPlaNet[50], ISNs[42], TransLocator[48] and so on. For vision-language models, we evaluate InternVL2-8B[7], LLaVA-Next-Mistral-7B[36] and Qwen2.5-VL-7B[2] using their official checkpoints and inference protocols. We further include several reasoning-oriented models, including GeoReasoner [31], GeoChat [28], and Geo-R1 [67], among others, which explicitly incorporate structured or multi-step geographic reasoning.

Table 2: Evaluation results on LandmarkBias-3K. + indicates gains over backbone Qwen2.5-VL-7B [2].

Model	City (25 km)	Region (200 km)	Country (750 km)
• General VLM Model			
GPT-4o-mini [43]	19.47	35.77	59.00
GPT-4o [44]	30.53	47.50	67.83
Qwen2.5-VL-7B [2]	16.83	28.67	44.57
InternVL2-8B [7]	5.77	11.54	32.40
LLaVA-Next-Mistral-7B [36]	0.74	1.59	5.10
• Domain-Specific Model			
TransLocator [48]	18.60	27.00	41.10
GeoCLIP [53]	17.96	30.45	60.88
GeoDecoder [13]	18.23	34.27	59.60
GeoChat [28]	19.67	38.77	63.37
GeoReasoner [31]	10.67	22.77	47.13
GLOBE [33]	23.27	44.27	63.90
Geoagent [27]	23.57	45.27	67.07
HoloGeo-SFT	20.03	30.33	64.05
HoloGeo-SFT+RL	27.27 (+10.44)	47.07 (+18.40)	68.20 (+23.63)

Evaluation Metrics. For geo-localization benchmarks, we follow previous work [26, 31, 67] and report the percentage of predictions whose geographic distance to the ground-truth coordinate falls within fixed thresholds of **25 km**, **200 km**, and **750 km**. Some models output discrete place names; we convert these predictions into their corresponding geographic center coordinates using external geocoding tools for distance-based evaluation. All metrics are reported as accuracy percentages, and higher values indicate better geo-localization performance.

Implementations. We adopt Qwen2.5-VL-7B-Instruct as the base model and perform parameter-efficient fine-tuning using LoRA [23] with a rank of 64 and an alpha of 128. AdamW optimizer is applied with a learning rate of $1e-5$ for LoRA parameters, a cosine learning rate scheduler with a warmup ratio of 0.01, a weight decay of 0.1, and a global batch size of 32. Details are provided in Appendix E.

5.2 Main Results

Results on Existing Benchmarks. Table 1 summarizes the comparison with existing geo-localization methods. Our SFT-only model already exhibits strong and competitive performance across multiple geographic scales, achieving 40.9% (City) on IM2GPS and 18.1% (City) on YFCC4k with more complex and diverse scenes, surpassing the recent GLOBE model [33]. Although the SFT variant does not always yield the highest accuracy among all models, its performance remains consistently close to the state-of-the-art level, indicating that the basic representation and training design are effective. Notably, further performance improvements are achieved by introducing GRPO-based reinforcement learning. Specifically, HoloGeo attains the best accuracy of 47.3% (City) and 60.3% (Region) on the IM2GPS dataset and achieves state-of-the-art results across city, region, and country levels on YFCC4k. Although its performance on IM2GPS3k is slightly inferior to the state-of-the-art, this is mainly attributed to the severe label distribution redundancy in this test set and the targeted optimization adopted by competing methods. In contrast, HoloGeo focuses more on building global generalization ability, thus exhibiting stronger robustness in diverse real-world scenarios. These improvements stem from HoloGeo’s

Table 3: Ablation study on HoloGeo’s training components and reward configurations. Evaluated on the IM2GPS dataset with Qwen2.5-VL-7B-Instruct [2] as the base model.

Models	SFT	Reward Function			City	Region
		R_{geo}	R_{box}	R_{CLR}		
Qwen2.5-VL-7B	-	-	-	-	35.02	46.83
Qwen2.5-VL-7B	✓	-	-	-	40.93	56.54
HoloGeo (Full)	✓	✓	✓	✓	47.26	60.34
w/o R_{geo}	✓	-	✓	✓	41.35	59.49
w/o R_{box}	✓	✓	-	✓	43.03	59.07
w/o R_{CLR}	✓	✓	✓	-	45.99	59.49

multi-dimensional reward mechanism and evidence-driven reasoning, which explicitly guide the model to focus on and leverage diverse effective geographic visual elements in images.

Results on Landmark-bias Datasets. Table 2 shows the comparison results of various models on our constructed LandmarkBias-3K benchmark, which is specifically designed to test the reasoning robustness of models when encountering misleading salient landmarks. It can be seen that both general VLM models and existing geo-localization methods are affected by landmark bias, resulting in low overall accuracy. Among them, models such as InternVL2-8B achieve less than 10% (City) accuracy. Even the strongest domain-specific models such as GLOBE and Geoagent only reach 23.27% and 23.57% (City) accuracy, respectively. In contrast, our HoloGeo-SFT model achieves 20.03% at the city level, demonstrating preliminary anti-bias ability. After incorporating GRPO-based reinforcement learning, HoloGeo-SFT+RL further improves the accuracy to 27.27% (City), while maintaining leading performance on region-level and country-level metrics. These results fully prove that HoloGeo can alleviate over-reliance on a single landmark and conduct robust reasoning by integrating effective geographic visual elements in challenging scenarios with misleading landmarks.

5.3 Ablation Study

In Table 3, we further investigate the impact of different training paradigms and reward configurations on the performance of HoloGeo. Specifically, we compare LoRA SFT with our proposed GRPO-based reinforcement learning approach. For SFT, the city-level accuracy is improved from 35.02% to 40.93%, indicating that structured geographic reasoning supervision can endow the model with basic geo-localization capabilities. For GRPO, we conduct a reward ablation study by evaluating different combinations of the three reward components. It shows that GRPO with the full set of rewards achieves the best overall performance of 47.26% in city-level accuracy. Removing any one of R_{geo} , R_{box} , and R_{CLR} leads to significant performance drops, which underscores the importance of supervision on both location-level correctness and comprehensive reasoning over multiple visual elements.

5.4 In-depth Analyses

Why landmark-bias occurs? Landmark-induced bias mainly arises from models concentrating visual attention on salient landmarks while ignoring broader geographic context. To this end, we



Figure 4: Comparison of Attention Distributions Across Qwen2.5-VL-7B-Instruct, GLOBE, GeoAgent, GeoReasoner, and HoloGeo.

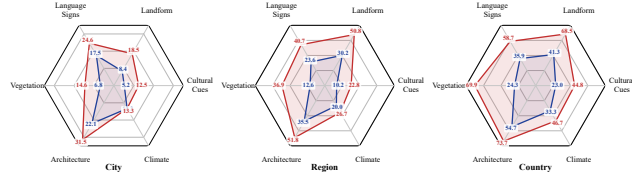


Figure 5: Prediction of Qwen2.5-VL-7B and HoloGeo on different landmark types.



Figure 6: Case study: Reasoning comparison of three different models on the same input image. Reliable geographic information identified by the models is annotated in the text.

employ gradient-based saliency analysis to generate saliency maps of the original images, which intuitively visualize the key visual regions that models focus on during geographic reasoning. As shown in Figure 4, for Qwen2.5-VL-7B-Instruct and other VLM-based geolocation models, the saliency maps are highly concentrated on a few visually prominent landmarks (e.g., iconic statues, distinctive buildings, or aircraft), while most surrounding regions of the scene receive little to no attention. Such a landmark-centered attention pattern causes models to overemphasize globally recognizable objects and neglect complementary geographic cues, including environmental context. Consequently, model predictions become strongly anchored to the landmarks themselves, leading to a significantly higher risk of confusion when similar landmarks appear in

different locations. In contrast, HoloGeo fully considers both landmark and non-landmark regions such as surrounding architecture and environmental details, exhibiting a more balanced attention distribution. This holistic attention pattern enables HoloGeo to integrate diverse geographic evidence and reduce over-reliance on any single visual cue. As a result, landmark bias is effectively mitigated, leading to more robust and reliable geolocation performance.

How does landmark bias vary across landmark types? Figure 5 analyzes how landmark-induced bias varies across landmark types by reporting geo-localization accuracy at the city, region, and country levels. The key observation is that the severity of bias depends on how specific and misleading a cue is for localization. For Qwen2.5-VL-7B, Architecture and Language Signs show relatively low city-level accuracy but noticeably higher country-level accuracy, suggesting that the model often locks onto globally recognizable styles or text patterns that provide a coarse geographic prior but are not distinctive enough to pin down the exact city, leading to plausible yet incorrect fine-grained predictions. In contrast, Climate yields comparatively higher accuracy, which we attribute to climate cues being less shortcut-like: climate-related evidence typically constrains the scene to broad latitudinal or regional bands rather than pointing to a single iconic place, so it induces a milder form of bias and is less likely to anchor the model to a specific wrong location. Finally, weak or transferable cues such as Vegetation, Landform, and Cultural Cues result in low accuracy across granularities because these signals are often shared across many regions and require careful integration with surrounding context; when the model’s attention is dominated by any salient element, it struggles to accumulate sufficient discriminative evidence for reliable localization.

Case Study. Figure 6 presents a representative case of landmark-induced bias, where salient landmarks lead baseline models to rely on superficial similarity and produce incorrect predictions. In contrast, HoloGeo mitigates this effect by integrating complementary contextual cues beyond the landmark, enabling correct geo-localization. We further provide additional case studies in Appendix F.3 to qualitatively illustrate how biased landmarks influence model behavior and decision-making.

6 Conclusion

This paper addresses the landmark bias issue in image geolocalization by proposing a holistic reasoning-driven framework, HoloGeo. On the one hand, we introduce two quantitative metrics, Bias Intensity (BI) and Bias Harmfulness (BH), to systematically characterize the impact of landmark cues on model decisions, and construct both a diagnostic evaluation benchmark (LandmarkBias-3K) and a high-quality training dataset (BF-30K) with structured multi-evidence reasoning chains. On the other hand, to mitigate landmark bias, HoloGeo integrates supervised fine-tuning with GRPO-based reinforcement learning, guided by a multi-dimensional reward design that promotes visual evidence grounding and logical reasoning consistency. This framework encourages models to shift from over-reliance on single salient landmarks to balanced reasoning over diverse geographic cues. Extensive experiments on IM2GPS, IM2GPS3K, YFCC4K, and LandmarkBias-3K demonstrate that HoloGeo improves both robustness under landmark ambiguity and overall geo-localization performance. In future work, we plan to explore tool-augmented and agent-based reasoning to incorporate external geographic knowledge, as well as more data-efficient learning paradigms such as few-shot and zero-shot generalization.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).
- [3] Maurits Bleeker, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. 2024. Demonstrating and Reducing Shortcuts in Vision-Language Representation Learning. *Transactions on Machine Learning Research* (2024).
- [4] Ron Campos, Ashmal Vayani, Parth Parag Kulkarni, Rohit Gupta, Aizan Zafar, Aritra Dutta, and Mubarak Shah. 2026. Gaea: A geolocation aware conversational assistant. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5236–5246.
- [5] Athanasios Chalvatzaras, Ioannis Pratikakis, and Angelos A. Amanatiadis. 2022. A survey on map-based localization techniques for autonomous vehicles. *IEEE Transactions on intelligent vehicles* 8, 2 (2022), 1574–1596.
- [6] Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025. PerturboLLaVA: Reducing Multimodal Hallucinations with Perturbative Visual Training. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=j4LITBSUjs>
- [7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.
- [8] Fenghua Cheng, Jinxian Wang, Sen Wang, Zi Huang, and Xue Li. 2025. GeoGuess: Multimodal Reasoning based on Hierarchy of Visual Information in Street View. *arXiv preprint arXiv:2506.16633* (2025).
- [9] Wenqing Cheng, Ruxue Wen, Haojun Huang, Wang Miao, and Chen Wang. 2022. OPTDP: Towards optimal personalized trajectory differential privacy for trajectory data publishing. *Neurocomputing* 472 (2022), 201–211.
- [10] Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. 2025. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510* (2025).
- [11] Ziheng Chi, Yifan Hou, Chenxi Pang, Shaobo Cui, Mubashara Akhtar, and Mrinmaya Sachan. 2025. Chimera: Diagnosing Shortcut Learning in Visual-Language Understanding. *arXiv preprint arXiv:2509.22437* (2025).
- [12] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. In *International Conference on Machine Learning*. PMLR, 10818–10838.
- [13] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. 2023. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23182–23190.
- [14] Zhiyang Dou, Zipeng Wang, Xumeng Han, Guorong Li, Zhipei Huang, and Zhenjun Han. 2024. Gaea: Towards interactive global geolocation assistant. *arXiv preprint arXiv:2412.08907* (2024).
- [15] Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. 2025. Around the world in 80 timesteps: A generative approach to global visual geolocation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23016–23026.
- [16] Aritra Dutta, Srijan Das, Jacob Nielsen, Rajatshubra Chakraborty, and Mubarak Shah. 2024. Multiview aerial visual recognition (mavrec): Can multi-view improve aerial visual perception?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22678–22690.
- [17] Yu Feng, Puzhen Zhang, Guohui Xiao, Linfang Ding, and Liqiu Meng. 2025. Towards a barrier-free GeoQA portal: Natural language interaction with geospatial data using multi-agent LLMs and semantic search. *Int. J. Appl. Earth Obs. Geoinformation* 144 (2025), 104825.
- [18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
- [19] Lukas Haas, Silas Alberti, and Michal Skreta. 2023. Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv preprint arXiv:2302.00275* (2023).
- [20] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. 2024. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12893–12902.
- [21] Xiao Han, Chen Zhu, Hengshu Zhu, and Xiangyu Zhao. 2025. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 814–825.
- [22] James Hays and Alexei A Efros. 2008. Im2gps: estimating geographic information from a single image. In *2008 IEEE conference on computer vision and pattern recognition*. 1–8.
- [23] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=nZEVKeeFY9>
- [24] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large Language Models Cannot Self-Correct Reasoning Yet. In *The Twelfth International Conference on Learning Representations*.
- [25] Pengyue Jia, Yiding Liu, Xiaopeng Li, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, Dawei Yin, and Xiangyu Zhao. 2024. G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. *Advances in Neural Information Processing Systems* 37 (2024), 53198–53221.
- [26] Pengyue Jia, Seongheon Park, Song Gao, Xiangyu Zhao, and Yixuan Li. 2025. GeoRanker: Distance-Aware Ranking for Worldwide Image Geolocalization. *arXiv preprint arXiv:2505.13731* (2025).
- [27] Modi Jin, Yiming Zhang, Boyuan Sun, Dingwen Zhang, MingMing Cheng, and Qibin Hou. 2026. GeoAgent: Learning to Geolocate Everywhere with Reinforced Geographic Characteristics. *CVPR* (2026).
- [28] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. 2024. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 27831–27840.
- [29] Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. 2017. The benchmarking initiative for multimedia evaluation: MediaEval 2016. *IEEE MultiMedia* 24, 1 (2017), 93–96.
- [30] Bin Li, Dehong Gao, Yeyuan Wang, Linbo Jin, Shanqing Yu, Xiaoyan Cai, and Libin Yang. 2025. Instruction-Aligned Visual Attention for Mitigating Hallucinations in Large Vision-Language Models. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.
- [31] Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. GeoReasoner: Geolocalization with Reasoning in Street Views using a Large Vision-Language Model. In *International Conference on Machine Learning*. PMLR, 29222–29233.
- [32] Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. 2025. From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models. *arXiv preprint arXiv:2508.01608* (2025).

- [33] Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. 2025. Recognition through Reasoning: Reinforcing Image Geo-localization with Large Vision-Language Models. In *Advances in Neural Information Processing Systems*.
- [34] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862* (2017).
- [35] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. 2018. Geoman: Multi-level attention networks for geo-sensory time series prediction.. In *Ijcai*, Vol. 2018. 3428–3434.
- [36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.
- [37] Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2025. PhD: A ChatGPT-Prompted Visual Hallucination Evaluation Dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11–15, 2025*. 19857–19866.
- [38] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*.
- [39] Yi Liu, Junchen Ding, Gelei Deng, Yuekang Li, Tianwei Zhang, Weisong Sun, Yaowen Zheng, Jingquan Ge, and Yang Liu. 2024. Image-based geolocation using large vision-language models. *arXiv preprint arXiv:2408.09474* (2024).
- [40] Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-Centric Reasoning with Grounded Chain-of-Thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 14268–14280.
- [41] Tessa Masis and Brendan O'Connor. 2026. Coordinates from Context: Using LLMs to Ground Complex Location References. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1591–1606.
- [42] Eric Muller-Budack, Kader Pustiu-Iren, and Ralph Ewerth. 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*. 563–579.
- [43] OpenAI. 2024. GPT-4o mini: Advancing Cost-Efficient Intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- [44] OpenAI. 2024. GPT-4o System Card. <https://openai.com/index/gpt-4o-system-card/>.
- [45] OpenAI. 2025. Introducing o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- [46] Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. 2019. Urban Traffic Prediction from Spatio-Temporal Data Using Deep Meta Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019*. 1720–1730.
- [47] Nhi Pham and Michael Schott. 2024. H-POPE: Hierarchical Polling-based Probing Evaluation of Hallucinations in Large Vision-Language Models. In *Workshop on Statistical Foundations of LLMs and Foundation Models*.
- [48] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. 2022. Where in the world is this image? transformer-based geo-localization in the wild. In *European Conference on Computer Vision*. Springer, 196–215.
- [49] Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents. <https://qwen.ai/blog?id=qwen3.5>
- [50] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. 2018. CPLaNet: Enhancing Image Geolocation by Combinatorial Partitioning of Maps. In *Proceedings of the European Conference on Computer Vision*.
- [51] Qiaomu Shen, Wei Zeng, Yu Ye, Stefan Müller Arisona, Simon Schubiger, Remo Burkhard, and Huamin Qu. 2017. StreetVizor: Visual exploration of human-scale urban forms based on street views. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 1004–1013.
- [52] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Commun. ACM* (2016).
- [53] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2023. Geolclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems* 36 (2023), 8690–8701.
- [54] An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. 2025. Vision language models are biased. *arXiv preprint arXiv:2505.23941* (2025).
- [55] Nam Vo, Nathan Jacobs, and James Hays. 2017. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*. 2621–2630.
- [56] Beining Wang, Weihang Su, Hongtao Tian, Tao Yang, Yujia Zhou, Ting Yao, Qingyao Ai, and Yiqun Liu. 2025. From< Answer> to< Think>: Multidimensional Supervision of Reasoning Process for LLM Optimization. *arXiv preprint arXiv:2510.11457* (2025).
- [57] Chun Wang, Xiaojun Ye, Xiaoran Pan, Zihao Pan, Haofan Wang, and Yiren Song. 2025. Gre suite: Geo-localization inference via fine-tuned vision-language models and enhanced reasoning chains. *Advances in Neural Information Processing Systems* (2025).
- [58] Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. 2023. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances in Neural Information Processing Systems* 36 (2023), 5301–5319.
- [59] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. 2022. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. In *International Conference on Learning Representations*.
- [60] Jiaheng Wei, Zhaowei Zhu, Tianyi Luo, Ehsan Amid, Abhishek Kumar, and Yang Liu. 2023. To aggregate or not? learning with separate noisy labels. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2523–2535.
- [61] T. Weyand, A. Araujo, B. Cao, and J. Sim. 2020. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*.
- [62] Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet-photo geolocation with convolutional neural networks. In *European conference on computer vision*.
- [63] Wikimedia Foundation. [n. d.]. Wikimedia API. <https://commons.wikimedia.org/w/api.php>.
- [64] Jiayang Wu, Wensheng Gan, Han-Chieh Chao, and Philip S. Yu. 2024. Geospatial Big Data: Survey and Challenges. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 17 (2024), 17007–17020.
- [65] Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating multimodal llm hallucination via bottom-up holistic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.
- [66] Zimin Xia and Alexandre Alahi. 2025. FG²: Fine-Grained Cross-View Localization by Fine-Grained Feature Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6362–6372.
- [67] Chenhui Xu, Fuxun Yu, Michael J. Bianco, Jacob Kovarskiy, Raphael Tang, Qi Zhang, Zirui Xu, Will LeVine, Brandon Dubbs, Heming Liao, Cassandra Burgess, Suvam Bag, Jay Patravali, Rupanjali Kukal, Mikael Figueroa, Rishi Madhok, Nikolaos Karianakis, and Jinjun Xiong. 2026. Unlocking Zero-Shot Geospatial Reasoning via Indirect Rewards. In *Forty-third International Conference on Machine Learning*. <https://openreview.net/forum?id=MJBAmHLr8E>
- [68] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. UrbanCLIP: Learning Text-enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13–17, 2024*. 4006–4017.
- [69] Hongji Yang, Xiufan Lu, and Yingying Zhu. 2021. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems* 34 (2021), 29009–29020.
- [70] Sahiti Yerramilli, Nilay Pande, Rynaa Grover, and Jayant Sravan Tamarapalli. 2025. GeoChain: Multimodal Chain-of-Thought for Geographic Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 23624–23639. doi:10.18653/v1/2025.findings-emnlp.1284
- [71] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. 2025. Chain-of-Focus: Adaptive Visual Search and Zooming for Multimodal Reasoning via RL. *arXiv preprint arXiv:2505.15436* (2025).
- [72] Zilun Zhang, Zian Guan, Tiancheng Zhao, Haozhan Shen, Tianyu Li, Yuxiang Cai, Zhonggen Su, Zhaojun Liu, Jianwei Yin, and Xiang Li. 2025. Geo-R1: Improving Few-Shot Geospatial Referring Expression Understanding with Reinforcement Fine-Tuning. *arXiv preprint arXiv:2509.21976* (2025).
- [73] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.
- [74] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems* 36 (2023), 55006–55021.
- [75] Zhongliang Zhou, Jieli Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. 2024. Img2Loc: Revisiting image geolocation using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval*. 2749–2754.
- [76] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025).
- [77] Sijie Zhu, Mubarak Shah, and Chen Chen. 2022. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF*

Appendix

A	Detailed Landmark-Bias Quantitation Metrics	11
A.1	Distribution of BI and BH	11
B	LandmarkBias-3K Benchmark	11
B.1	Source Data	11
B.2	Threshold Selection	11
B.3	More Examples	11
C	Detailed BF-30K Annotation	13
D	Training Prompt Template	13
E	Detailed Experimental Settings	13
E.1	Detailed Evaluation Benchmarks	13
E.2	Detailed Implementation	15
F	Extended Experiments	15
F.1	Analysis of Landmark Bias	15
F.2	BI and BH Distribution Analysis	16
F.3	More Case Examples	16
G	Model Cross-Validation	17
H	Limitation	17

A Detailed Landmark-Bias Quantitation Metrics

A.1 Distribution of BI and BH

In this appendix, we provide further analysis of the proposed bias metrics. We first examine the empirical distributions of Bias Intensity (BI) and Bias Harmfulness (BH) on LandmarkBias-3K, and then discuss a small set of counterintuitive cases with negative BI values.

Figure 7 illustrates the distributions of BI and BH. Both exhibit clear right-skewed patterns, with most samples concentrated in low-to-moderate positive ranges and a long tail extending toward higher values. This suggests that while many samples exhibit moderate landmark-induced bias, a substantial portion involves stronger and more harmful effects. The median values of BI and BH are 2.37 and 0.28, respectively, indicating that the benchmark predominantly consists of samples with positive landmark influence and non-trivial harmfulness. Importantly, the distribution is not dominated by extreme cases, but instead spans a wide range of bias levels. Overall, these results demonstrate that LandmarkBias-3K provides a diverse and challenging evaluation set, enabling robust assessment of geo-localization models under misleading landmark cues.

B LandmarkBias-3K Benchmark

In this appendix, we provide additional details on the construction of the proposed **LandmarkBias-3K** benchmark, a diagnostic dataset designed to evaluate the robustness of geo-localization models under misleading landmark cues.

B.1 Source Data

We construct LandmarkBias-3K from two large-scale geo-localization datasets, **MP-16** [29] and **GLDv2** [61], which offer broad geographic coverage and diverse real-world scenes. For each image, we use GroundingDINO [38] to detect the prominent landmark region r . Based on this region, we construct two auxiliary inputs: the landmark-removed image x_r , obtained by masking out r , and the landmark-only crop r , which isolates the detected landmark.

B.2 Threshold Selection

For each image, we perform three forward passes using the full image x , the landmark-removed image x_r , and the landmark-only region r . Following the definitions in the main text, we obtain the corresponding prediction probabilities $I_1(\cdot)$, $I_2(\cdot)$, and $I_3(\cdot)$, and compute the two bias metrics, **Bias Intensity** (BI) and **Bias Harmfulness** (BH). These metrics capture complementary aspects of landmark-induced bias. A positive BI indicates that the landmark increases the model’s confidence in the landmark-induced label, while a positive BH indicates that such influence is harmful, as it reduces the model’s relative preference for the ground-truth label. We therefore retain only samples satisfying:

$$BI > 0, \quad BH > 0, \quad (10)$$

ensuring that the selected images exhibit both positive and harmful landmark-induced bias. We further remove geographically ambiguous samples through manual verification. The resulting benchmark contains **3,000** images, forming the **LandmarkBias-3K**.

B.3 More Examples

Figure 8 presents representative cases with different BI/BH sign combinations, illustrating that landmark cues can influence geo-localization models in qualitatively different ways.

- **(a) $BI > 0, BH > 0$: Harmful Landmark Bias.** This case represents the primary failure mode targeted by LandmarkBias-3K. In the first example, the image contains a structure resembling the Eiffel Tower, which strongly anchors the model to France when focusing on the landmark alone, while the true location is Hangzhou, China. As a result, the landmark exerts a strong positive pull toward the landmark-induced hypothesis, yielding a positive BI, and simultaneously weakens the model’s relative preference for the ground truth, leading to a positive BH. A similar phenomenon appears in the second example from the United States, where a highly salient landmark cue again dominates the prediction and drives the model toward an incorrect but visually plausible location hypothesis. Together, these cases illustrate the standard harmful landmark-bias pattern: the landmark is visually prominent enough to attract the model, but this attraction is misleading and overrides broader contextual evidence.
- **(b) $BI > 0, BH < 0$: Strong but Beneficial Landmark Influence.** In the first example, the landmark is the Merlion, a highly distinctive symbol of Singapore. The model therefore assigns strong confidence to the landmark-induced label, producing a large positive BI; however, because this landmark cue is genuinely reliable and consistent with the full scene, it supports rather than harms the correct prediction, yielding a negative BH. The second example from Japan shows the same pattern: highly discriminative visual cues strongly influence the model, but this influence remains aligned with the ground-truth location. These cases demonstrate that strong landmark influence is not inherently harmful. Instead, when the landmark is trustworthy and geographically discriminative, it can facilitate correct localization. They also highlight the complementary roles of BI and BH: BI measures

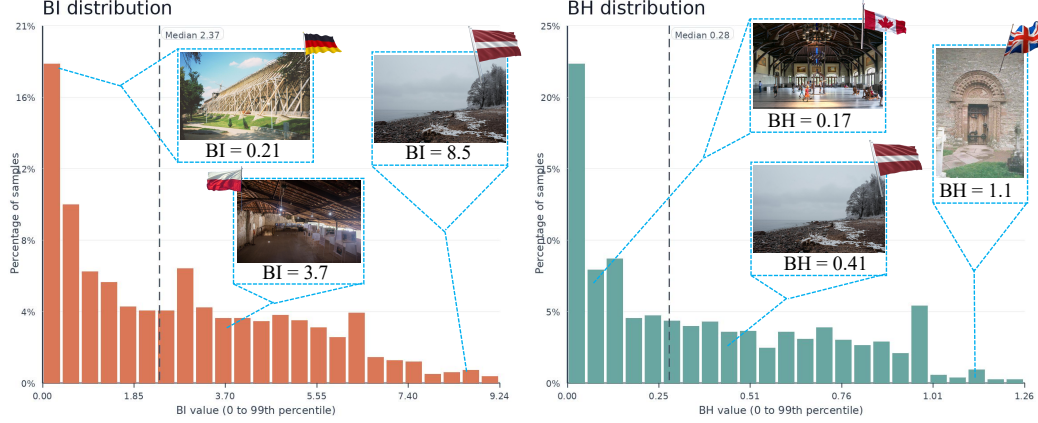


Figure 7: Empirical distributions of Bias Intensity (BI) and Bias Harmfulness (BH) in LandmarkBias-3K. The dashed lines mark the median values.

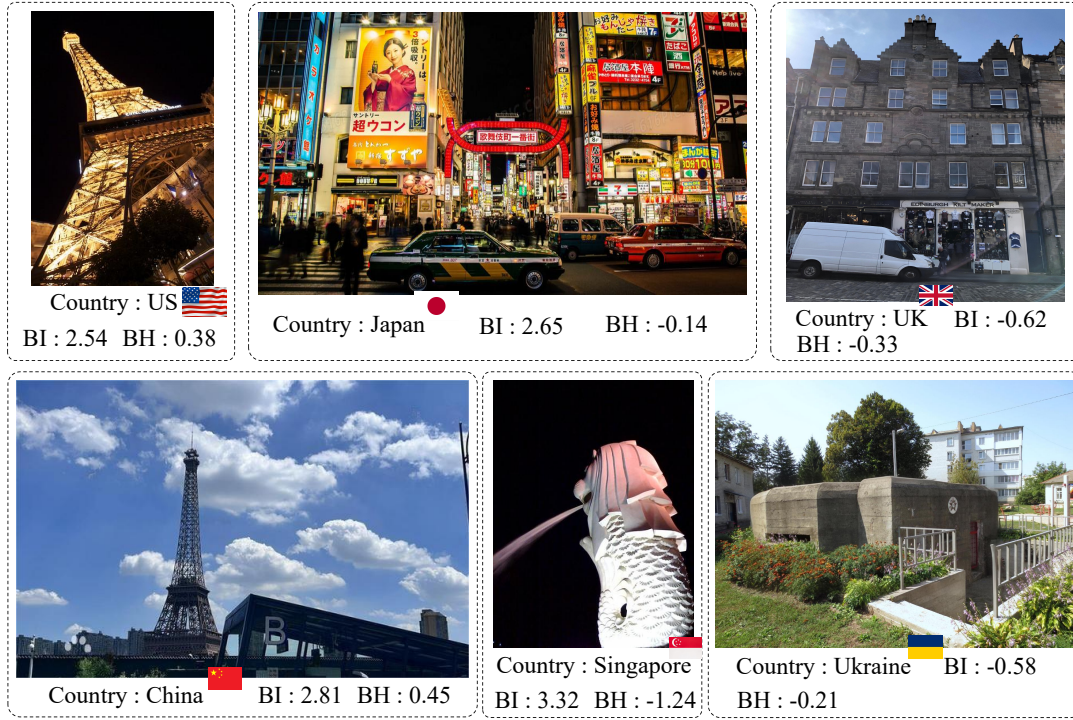


Figure 8: Representative cases with different BI/BH sign combinations: (a) $BI > 0, BH > 0$, harmful landmark bias; (b) $BI > 0, BH < 0$, strong but non-harmful landmark influence; (c) $BI < 0, BH < 0$, contextual reinterpretation.

the strength of landmark influence, while BH determines whether that influence is actually detrimental.

- (c) $BI < 0, BH < 0$: **Contextual Reinterpretation.** In the first example, the model predicts Russia when observing either the landmark alone or the landmark-removed image, while the ground truth is Ukraine. However, once the full image is provided, the model shifts away from the landmark-induced prediction, resulting in a negative BI. At the same time, this shift moves the prediction relatively closer to the ground truth, yielding a negative BH. A similar effect is observed in the second example from the United Kingdom, where the landmark-only interpretation is weakened after

being placed back into the full scene. These cases suggest that the meaning of an isolated landmark can be reinterpreted in the context of the entire image: rather than simply reinforcing the landmark-only hypothesis, the global scene can reshape or partially override it. In this sense, landmark influence is not always additive, but can depend strongly on its interaction with surrounding contextual evidence.

Cases with $BI < 0$ and $BH > 0$ are relatively rare in our analysis. Overall, these examples show that landmark cues can be harmful, beneficial, or contextually corrected depending on their interaction with the surrounding visual context, which further highlights

GroundingDINO Prompt Template	Reasoning Prompt Template
Any object, including buildings, people, characters, animals, flags, landscapes.	The confirmed geographical location corresponding to the original image is: 'country': {country}, 'city': {city}. Your final task is to INDEPENDENTLY determine the country and city where the original image is located based solely on visual evidence. Please structure your reasoning in three distinct parts: <Analyze> (geographic clue analysis), <Think> (comprehensive reasoning), and <Answer> (final conclusion). All analyses must be based on visual evidence. In <Analyze> part, DO NOT mention the country and city of the final answer, only conduct analysis. Please output strictly in the following format:
Box Filtering Prompt Template	
You are an expert in geographic information filtering. To ensure the diversity, variety, and representativeness of the selected box elements, you need to remove boxes that represent similar geographical and cultural symbols, contain excessive content, or correspond to panoramic images. Please carefully read the image areas corresponding to the multiple boxes entered and think carefully. Please output the results strictly in the following format, retaining only boxes that meet the criteria. The format is: <pre>{ "geo_boxes": [{ "Box": [x0, y0, x1, y1], "Object description": "...Here is your brief description of the content contained in Box...", "Reason": "...this is the brief reasons why you chose to keep this Box..." }, ...] }</pre> List of Boxes to be filtered: [List of candidate bounding boxes]	<pre>{ "Box": [x0, y0, x1, y1], "Analysis": "Detailed analysis(20-50 words) of the association between the content of this area and the geographical location under the original image context." }, ...]</pre> </Analyze> <Think> Provide a detailed, step-by-step breakdown of your comprehensive reasoning process to infer the most likely country and city. Strictly follow these requirements for your thinking: 1. Core evidence synthesis: Synthesize and aggregate the unique visual features and geographic association clues from ALL boxes in <Analyze> (avoid repeating single-box details; focus on integrating overlapping/ complementary/ key features into a set of core evidence points). List each core evidence point with specific, consolidated details (e.g., landscape types, architectural styles, man-made facilities, vegetation, climate-related traits). 2. Geographic feature correlation: For each synthesized core evidence point, explicitly link it to the specific geographic, climatic, cultural or infrastructural characteristics of a certain region/country (explain why this feature is typical of the region, not just a general trait). 3. Logical reasoning chain: Connect the correlated geographic features step by step to narrow down the scope from "continent/hemisphere" to "country", then to "specific city/coastal town". Clearly demonstrate how the combination of core evidence points excludes other possible regions and uniquely points to the target location. 4. Conflict resolution (skip this part if no conflicts exist): If there is any conflicting or ambiguous visual information across different Box analyses, resolve it with sufficient geographic and logical reasoning. 5. Conclusion justification: Conclude with the most likely country and city, and summarize how the synergy of all synthesized core evidence matches the known geographical characteristics of this location. </Think> <Answer> "country": "{country}", "city": "{city}" </Answer>

Figure 9: The prompt template used in BF-30K annotation.

the need to jointly consider BI and BH when analyzing landmark-induced bias in geo-localization.

Figure 10 also provides representative cases from different landmark categories, such as landform, climate, vegetation, cultural cues, architecture, and language signs. This categorization shows that landmark-induced bias can originate from a diverse range of visual cues rather than only iconic landmarks, further complementing the BI/BH-based analysis above.

C Detailed BF-30K Annotation

Here, we detail the annotation process of the BF-30K training dataset, which consists of region proposal, semantic filtering, structured reasoning annotation, and multi-dimensional validation.

Region Proposal. We employ GroundingDINO-SwinT-OGC [38] as an open-vocabulary object detector to identify candidate visual regions. To maximize the recall of geographically relevant cues, we adopt a low-threshold strategy, setting both the box confidence threshold and the text–visual alignment threshold to 0.06. This encourages the detection of a broad set of potential geographic elements. We then remove low-confidence boxes and eliminate duplicate detections using an IoU threshold of 0.75, reducing redundancy while preserving spatial coverage and semantic diversity. The corresponding prompt is shown in Figure 9.

Semantic Filtering. We further refine the candidate regions using Qwen2.5-VL-72B-Instruct [2] as a geographic semantic filter. Leveraging its visual reasoning capability, the model selects regions that contain informative geographic, cultural, or environmental cues. The prompt used in this stage is illustrated in Figure 9, where the list of candidate boxes is dynamically appended for each instance. This step yields a set of representative regions covering diverse geosemantic elements, including architectural structures, landmarks, vegetation, landforms, and cultural symbols.

Structured Reasoning Annotation. Based on the selected regions, we perform fine-grained geographic analysis and structured reasoning annotation. We use Qwen2.5-VL-72B [2] and ChatGPT-o3 [45] as reasoning models. The input consists of the full image and cropped patches corresponding to each region. Models are

required to generate reasoning outputs under a strictly defined format, including regional analysis, multi-step reasoning, and final location prediction. To ensure faithful reasoning, the prompt (Figure 9) explicitly restricts the model to rely solely on visual evidence, without access to ground-truth location labels. This process produces structured annotations that are both interpretable and logically consistent. Examples of the resulting training data are shown in Figure 11.

Multi-dimensional Validation. In this process, we employ the InternVL3-78B [76] as an independent verification model to conduct cross-model consistency checks on the reasoning annotations. The prompt used for the model is illustrated in Figure 12, whose core task is to judge whether the geolocation reasoning process is logically self-consistent and aligned with visual content based on the image, and output only "Yes" or "No" as the decision result. After automated validation, samples identified as contradictory or inconsistent are filtered out. In addition, we perform manual spot checks on the entire annotated dataset with a sampling ratio of 5% to further correct annotation errors, place name errors, logical conflicts, and hallucinations that may be missed by model-based verification.

D Training Prompt Template

To ensure input consistency between fine-tuning and testing, we design and employ a unified prompt template for vision-language model, as illustrated on the left side of Figure 13, which guarantees a fair comparison of model performance. Meanwhile, the prompt template for the Comprehensive Logical Reasoning Reward (R_{CLR}) is shown on the right side of Figure 13.

E Detailed Experimental Settings

E.1 Detailed Evaluation Benchmarks

We evaluate our method on four geo-localization benchmarks with different data sources and visual characteristics.

IM2GPS [22] is a classical geo-localization benchmark. Its standard test set contains 237 images drawn from a large Flickr-based geo-tagged image collection. Despite its relatively small size, many

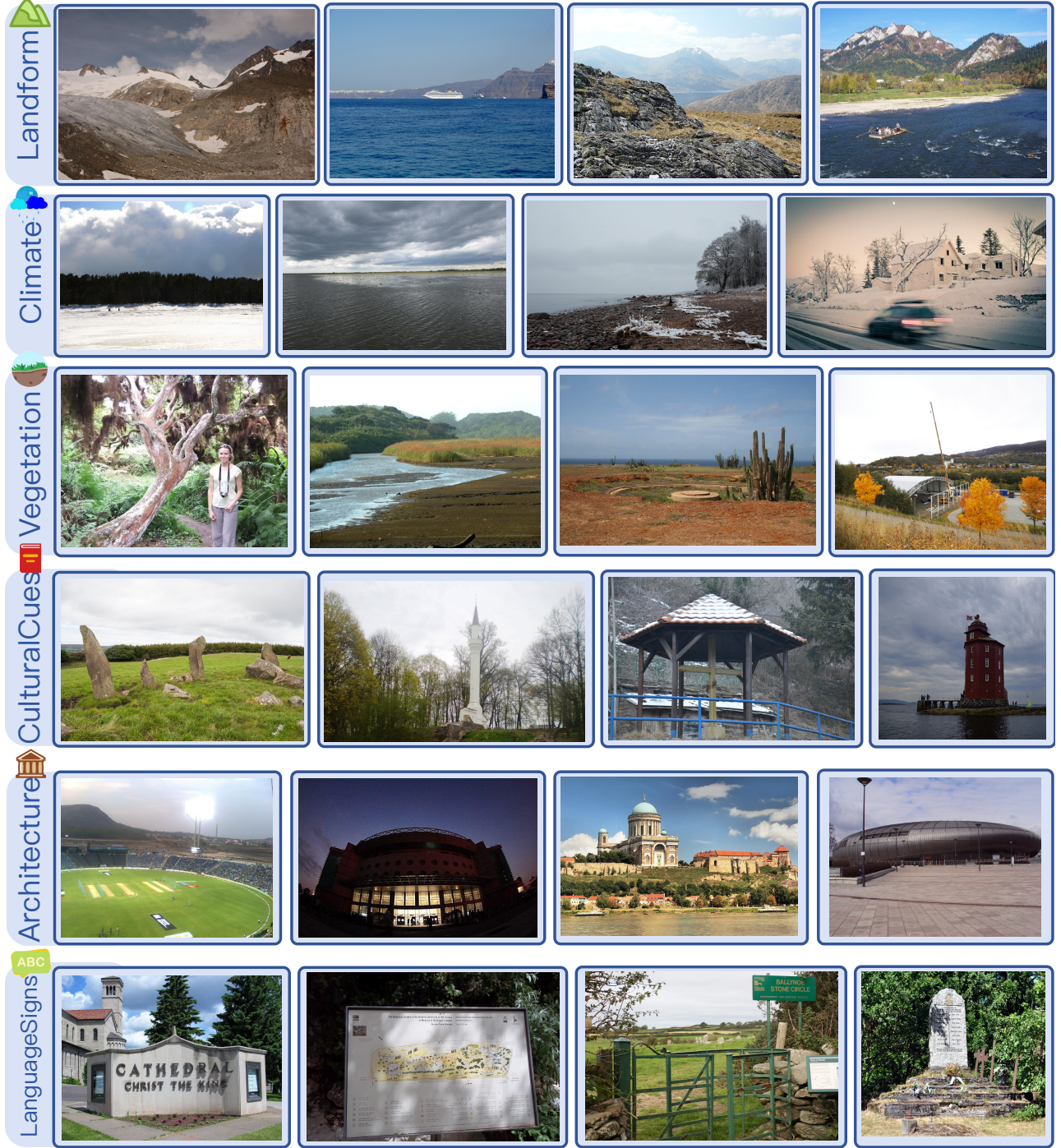


Figure 10: Representative cases of landmark categories in LandmarkBias-3K.

samples are visually distinctive and suitable for geographic reasoning, making it a widely used baseline for early evaluation.

IM2GPS3K is a larger benchmark introduced by Vo et al. [55] to enable more robust evaluation. It contains 3,000 images sampled from the original IM2GPS collection, providing a broader and more challenging test set. Compared to the standard IM2GPS split, it

exhibits greater diversity in scene types, including landmarks, urban environments, and natural landscapes.

YFCC4K [55] consists of 4,000 randomly sampled images from the YFCC-100M dataset [52], a large-scale Flickr-based multimedia collection. As a general-purpose dataset, YFCC4K is visually diverse

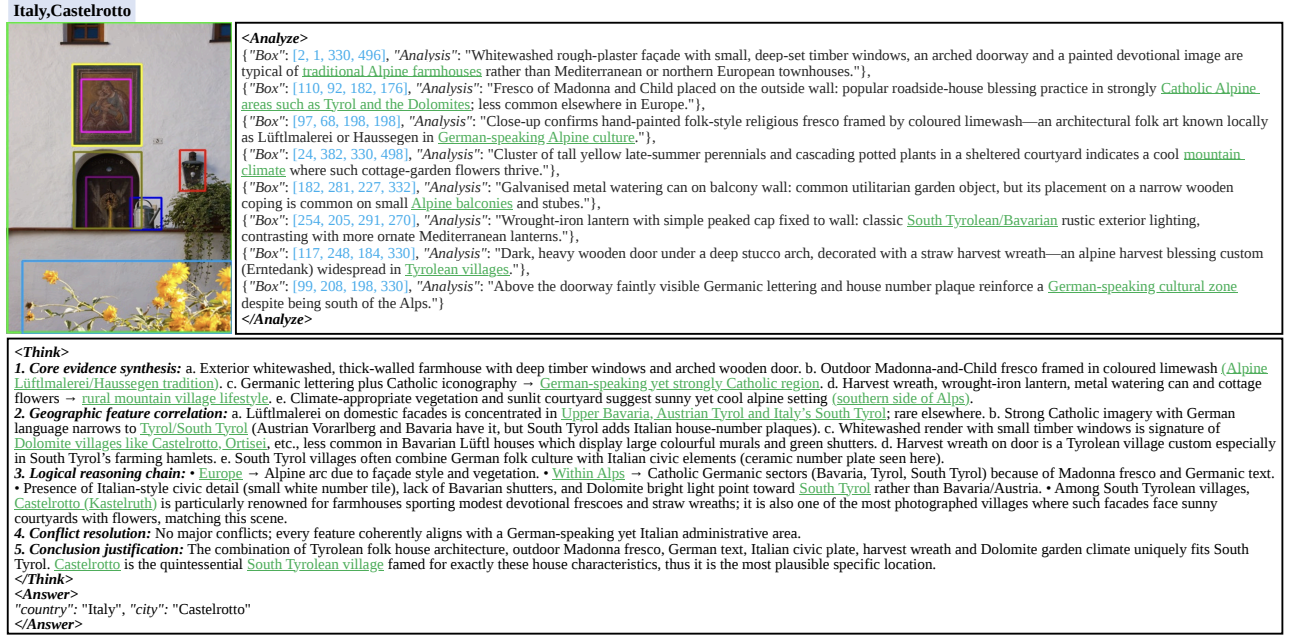


Figure 11: Visualization of an example of the training dataset.

InternVL3-78B reviewer Prompt Template

You are an expert reviewer specializing in geographic knowledge.

Given an image and its geolocation reasoning process, judge whether the reasoning logic is consistent and reasonable.

Output only Yes or No.

Figure 12: The prompt template used in the InternVL3-78B reviewer for multi-dimensional validation.

and less biased toward iconic landmarks, making it a complementary benchmark for evaluating generalization beyond landmark-centric scenarios.

Landmark-Bias-3K Benchmark is our proposed diagnostic benchmark for evaluating landmark-induced bias. It contains 3,000 images explicitly curated to assessing whether models over-rely on dominant landmark evidence.

E.2 Detailed Implementation

We conduct SFT as the first stage of model training, with GRPO-based post-training. We adopt Qwen2.5-VL-7B-Instruct as the base model and perform parameter-efficient fine-tuning using LoRA with a rank of 64. During SFT, we train the model for 1 epoch using the AdamW optimizer with an initial learning rate of 1×10^{-4} , a warmup ratio of 0.05. Training is deployed on 4× NVIDIA H100 GPUs with DeepSpeed Zero-3 for memory optimization, yielding a global batch size of 32. For GRPO-based reinforcement learning, we use a smaller learning rate of 1×10^{-5} , a warmup ratio of 0.01, and set weight decay to 0.1. We keep the LoRA configuration consistent across both training stages. We employ a cosine learning rate scheduler in both training phases. We summarize the key hyperparameters used for SFT and GRPO in our training setup in

Table 4. These settings are chosen based on standard practices for fine-tuning large vision-language models and further adjusted via preliminary ablation studies conducted on a held-out validation set. Unless stated otherwise, the same configuration is adopted across all experiments to ensure comparability and reproducibility.

Additionally, Figure 14 presents an example of the proposed $\mathcal{R}_{\text{CLR}}$, illustrating how Qwen3.5-35B-A3B evaluates a generated geographic reasoning chain. The reward model performs structured assessment across the three reasoning stages (<Analyze>, <Think>, and <Answer>), jointly considering visual grounding accuracy, logical consistency, and answer specificity. By explicitly verifying whether the reasoning is supported by localized visual evidence and whether multiple cues are coherently integrated, the model can effectively identify flawed or over-simplified reasoning patterns. Moreover, the probabilistic scoring mechanism provides a soft and stable measure of reasoning validity, enabling fine-grained optimization during reinforcement learning. This design allows $\mathcal{R}_{\text{CLR}}$ to serve as a reliable signal for guiding the model toward factually grounded and logically consistent multi-evidence reasoning.

F Extended Experiments

F.1 Analysis of Landmark Bias

Figure 16 presents another representative example of *landmark-induced bias* in image geo-localization. The scene contains several visually salient yet geographically ambiguous cues, including a medieval fortress, green-domed Orthodox architecture, and a river-front landscape, which collectively evoke a broad Eastern European visual prior. As shown, both Geo-R1 and GeoChat are attracted by these dominant but insufficient landmark cues and prematurely commit to *Moldova*, producing plausible yet incorrect hypotheses

SFT & RL Prompt Template	R _{CLR} Prompt Template
You are a geography expert. Your goal is to identify the city and country where the image is located. Please analyze in detail and think step by step. All your analyses must be based on visual evidence from the image and its boxed regions. You must provide a specific answer and infer the most plausible one if unsure. Output strictly in the following three-part format: <pre> <Analyze> [{ "Box": [x0, y0, x1, y1], "Analysis": "Detailed analysis of the association between the content of this area and the geographical location under the original image context." }, ...] </Analyze> <Think> Provide a detailed, step-by-step breakdown of your comprehensive reasoning process to infer the most likely country and city. Strictly follow these requirements for your thinking. 1. Core evidence synthesis: Synthesize and aggregate the unique visual features and geographic association clues from ALL boxes in <Analyze> (avoid repeating single-box details; focus on integrating overlapping/ complementary/ key features into a set of core evidence points). List each core evidence point with specific, consolidated details (e.g., landscape types, architectural styles, man-made facilities, vegetation, climate-related traits). 2. Geographic feature correlation: For each synthesized core evidence point, explicitly link it to the specific geographic, climatic, cultural or infrastructural characteristics of a certain region/country (explain why this feature is typical of the region, not just a general trait). 3. Logical reasoning chain: Connect the correlated geographic features step by step to narrow down the scope from "continent/hemisphere" to "country", then to "specific city/coastal town". Clearly demonstrate how the combination of core evidence points excludes other possible regions and uniquely points to the target location. 4. Conflict resolution (skip this part if no conflicts exist): If there is any conflicting or ambiguous visual information across different Box analyses, resolve it with sufficient geographic and logical reasoning. 5. Conclusion justification: Conclude with the most likely country and city, and summarize how the synergy of all synthesized core evidence matches the known geographical characteristics of this location. </Think> <Answer> "country": "", "city": "" </Answer> </pre>	Given an input image and its corresponding geographic location reasoning output, please judge whether the entire reasoning process is valid. Criteria for 'yes': - In the <Analyze> section: All described visual features are consistent with the actual content in the image. - In the <Think> section: The reasoning logic is coherent, and does not rely solely on a single geographic feature unless no other identifiable objects or visual clues exist in the image. - In the <Answer> section: The country and city fields are filled with specific, concrete place names, and do NOT contain 'Not specified', 'Unknown', or other vague expressions. Criteria for 'no': If ANY of the above three conditions is not met, output 'no'. Note: Only output 'yes' or 'no' directly, without any additional text. Do NOT judge whether the final city/country is factually correct.

Figure 13: Prompts in SFT and GRPO.

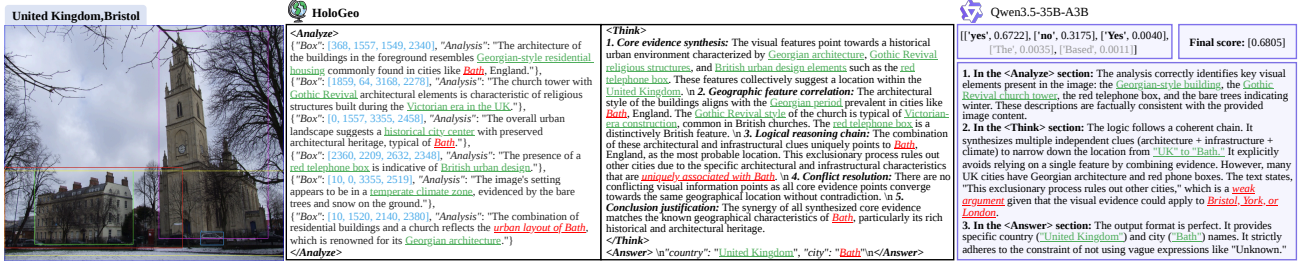
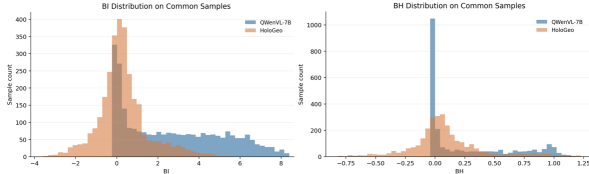
Figure 14: A case of $\mathcal{R}_{CLR}$.

Figure 15: Comparison of the BI and BH distributions between Qwen2.5-VL-7B-Instruct and HoloGeo on LandmarkBias-3K.

based on coarse landmark-level similarity rather than robust geographic evidence. In contrast, our model remains less susceptible to such dominant visual bias: instead of over-relying on a single prominent structure or stylistic cue, it integrates more holistic scene context, aggregating complementary evidence from settlement layout, architectural composition, and surrounding landscape, which leads to the correct localization in *Khotyn, Ukraine*.

F.2 BI and BH Distribution Analysis

To further assess whether HoloGeo mitigates landmark bias at the model-behavior level, we compute BI and BH for Qwen2.5-VL-7B-Instruct and HoloGeo on LandmarkBias-3K. As shown in Figure 15, compared with Qwen2.5-VL-7B-Instruct, HoloGeo exhibits BI and

BH distributions that are more concentrated in lower-value regions, with markedly shorter high-value tails. This indicates that HoloGeo reduces excessive landmark anchoring and mitigates the harmful influence of misleading landmarks on accurate localization, further validating its effectiveness in alleviating landmark bias.

F.3 More Case Examples

Figures 18 and 19 present additional successful examples of HoloGeo on LandmarkBias-3K and IM2GPS, respectively. Across these cases, the model does not rely solely on a single salient landmark or iconic structure. Instead, it integrates broader scene evidence, such as architectural style, urban or settlement layout, surrounding vegetation, traffic patterns, textual signs, and environmental context, and then combines these complementary cues to infer the correct location. This further supports our main claim that HoloGeo is able to mitigate over-reliance on dominant landmark cues and make more robust geo-localization decisions through evidence-driven reasoning. Figure 20 shows several failure cases on IM2GPS. A common pattern is that these examples come from relatively uncommon or underrepresented countries or regions, where the visual evidence is less frequently observed during training. In such cases, although the model can still capture some coarse regional characteristics, its geographic knowledge is insufficient to reliably distinguish rare locations from more familiar visually similar ones,

Table 4: The values of each hyperparameter.

Symbol	Description	Value
<i>Supervised Fine-Tuning (SFT) Configuration</i>		
l_r	Initial learning rate	1×10^{-4}
r_{lor}	LoRA rank	64
α_{lor}	LoRA alpha	128
p_{lor}	LoRA dropout	0.05
b_{train}	Per-device training batch size	1
L_{max}	Maximum sequence length	8192
r_{warm}	Warmup ratio	0.05
<i>GRPO Reinforcement Learning Configuration</i>		
$l_{r,\text{rl}}$	Initial learning rate	1×10^{-5}
G	Group size	8
t	Sampling temperature	0.7
β	KL divergence coefficient	0.001
$r_{\text{lor},\text{rl}}$	LoRA rank	64
$\alpha_{\text{lor},\text{rl}}$	LoRA alpha	128
$p_{\text{lor},\text{rl}}$	LoRA dropout	0.05
$L_{\text{max},\text{rl}}$	Maximum sequence length	8192
L_{comp}	Maximum completion length	1024
$r_{\text{warm},\text{rl}}$	Warmup ratio	0.01
λ_1	Weight of R_{geo}	1.2
λ_2	Weight of R_{box}	0.3
λ_3	Weight of R_{CLR}	0.5
λ_{fmt}	Weight of R_{format}	1.0
α	City-level weight in R_{geo}	0.8
τ_{IoU}	IoU threshold in R_{box}	0.6
<i>LLM Judge Configuration</i>		
t_{judge}	Sampling temperature	0.2
$L_{\text{max},\text{judge}}$	Maximum generation tokens	5
l_{prob}	Return log probabilities	True
k_{top}	Number of top logprobs	5

which leads to incorrect predictions. These results indicate that, beyond reducing landmark bias, further improvement also requires broader and more fine-grained geographic knowledge coverage for long-tail locations.

G Model Cross-Validation

To examine whether the proposed BH and BI signals are tied to a specific backbone, we compare the predictions of InternVL3-78B [76] and Qwen2.5-VL-72B [2] on the full evaluation set (Figure 17). Since different VLMs may produce scores on different numeric scales, we assess cross-model consistency at the level of relative trends rather than raw score values. Specifically, for each metric (BH or BI), we partition each model’s predictions into three equal-frequency bins based on its own score distribution, corresponding to low, medium, and high terciles.

Based on these tercile assignments, we compute two agreement measures. The *exact-match ratio* is the proportion of samples assigned to the same tercile by both models. The *adjacent-or-better ratio* counts samples whose assignments differ by at most one tercile, treating only low-versus-high cases as strong disagreements.

This formulation captures whether the two models preserve a consistent coarse ordering of samples despite differences in score scale. As shown in Figure 17, both metrics exhibit substantial cross-model consistency. For BH, 54.88% of samples fall into the same tercile and 83.62% into the same or an adjacent tercile. For BI, the agreement is even stronger, with 57.46% exact matches and 96.00% adjacent-or-better consistency. These values are significantly higher than a random assignment baseline, indicating that the relative ordering induced by BH and BI is largely preserved across model families. The row-normalized confusion matrices further illustrate this behavior. Each row corresponds to a tercile assigned by InternVL and each column to a tercile assigned by Qwen. Diagonal entries indicate exact agreement, near-diagonal entries reflect mild shifts, and corner entries correspond to cross-extreme disagreements. The matrices show that most discrepancies occur between neighboring bins, while extreme contradictions (low vs. high) are rare, particularly for BI.

Overall, these observations suggest that the BH and BI patterns capture stable, model-agnostic patterns of landmark-induced bias, rather than artifacts specific to a particular VLM backbone.

H Limitation

Although our proposed HoloGeo model has achieved excellent geolocalization performance on multiple public datasets, it still has certain limitations. First, since the training data mainly comes from existing datasets such as MP-16 [29] and GLDv2 [61], the data distribution is somewhat biased, resulting in poor recognition performance of the model in remote areas, underrepresented landforms, and data-sparse regions. Second, the current model still relies on supervised training with annotated datasets, which incurs certain training costs. In addition, the social impact of geolocalization models deserves equal attention. Caution should be exercised in real-world applications to guard against potential privacy leakage risks. The usage must be strictly regulated to prevent it from being used for illegal tracking, facilitating criminal activities, or violating personal privacy and spatial security.

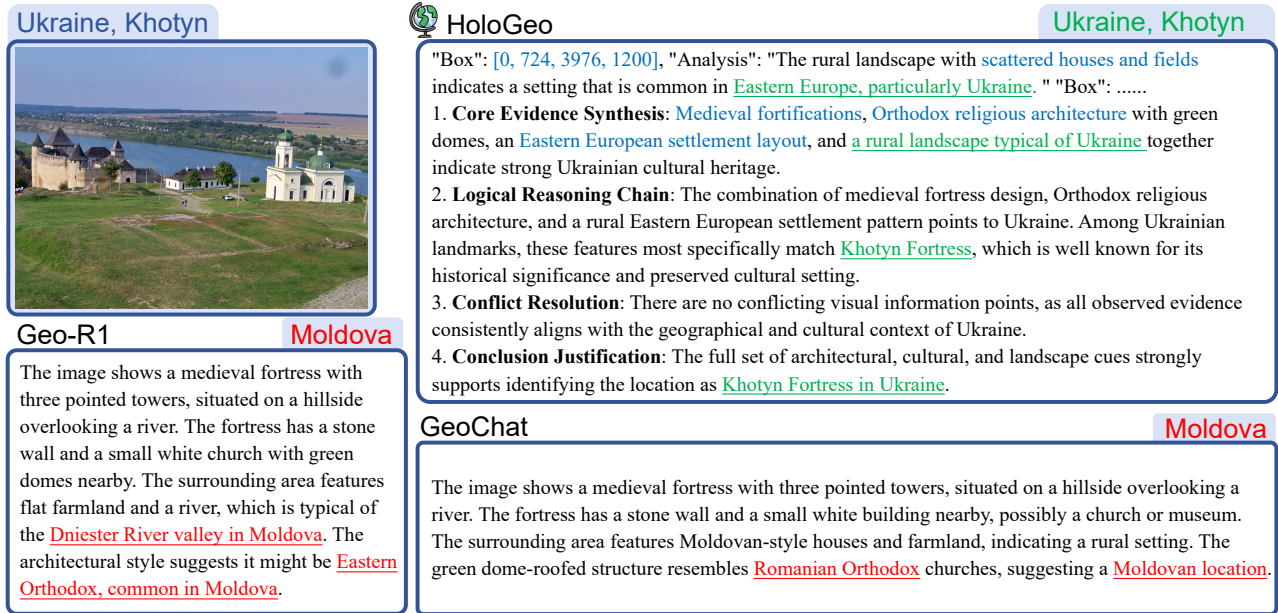


Figure 16: Case study: Reasoning comparison of three different models on the same input image. Reliable geographic information identified by the models is annotated in the text.

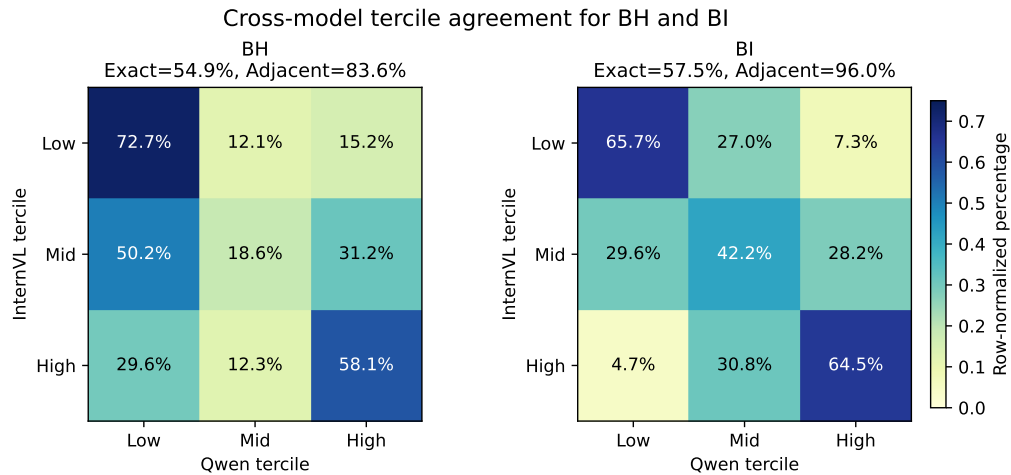
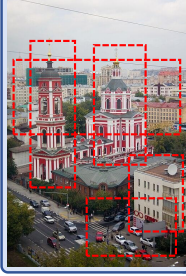


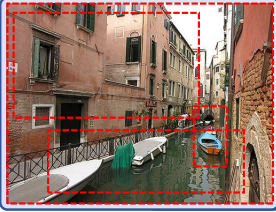
Figure 17: Cross-model validation of BH and BI based on tercile agreement between InternVL3-78B and Qwen2.5-VL-72B. The confusion matrices show that coarse relative trends are largely consistent across model families.

Russia, Moscow



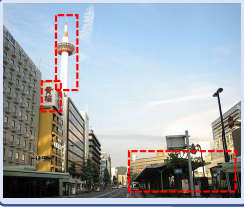
"Box": [68, 107, 169, 520], "Analysis": "The church has a red and white color scheme, onion-shaped domes, and crosses, which are characteristic of [Russian Orthodox architecture](#)." "Box": [188, 135, 414, 472], "Analysis": "This church also shows Russian Orthodox features such as onion-shaped domes and ornate detailing, reinforcing the same geographic context." "Box":
1. Core Evidence Synthesis: Architectural Style: The churches display [Russian Orthodox architectural elements](#), including red and white colors, onion-shaped domes, and crosses. Urban Context: The coexistence of historical religious structures and modern buildings indicates a contemporary urban environment. [Geographic Feature Correlation:](#) [Russian Orthodox Architecture:](#) [These features strongly suggest Russia](#). Mixed Historic and Modern Skyline: This combination is consistent with Moscow's urban landscape.
2. Logical Reasoning Chain: The church architecture points to a Russian cultural setting. The surrounding modern residential and commercial buildings indicate a large contemporary city. The combination of historic religious landmarks and modern urban development is [most consistent with Moscow, Russia](#).
3. Conflict Resolution: There are no conflicting visual cues across the boxes.
4. Conclusion Justification: The consistent [Russian Orthodox architectural style](#) and the mixed historic-modern urban setting strongly [suggest Moscow, Russia](#).

Italy, Venice



"Box": [0, 0, 367, 318], "Analysis": "The image shows a canal with historic buildings featuring brick facades, and Venetian-style details." "Box":
1. Core Evidence Synthesis: Architecture: Brick facades, small windows, and decorative [details match Venetian design](#). Canal Scene: The canal, boats, and railings indicate a [water-adapted setting](#). Historical Context: Weathered buildings reflect historical heritage. [Geographic Feature Correlation:](#) These features point to [Italy, especially Venice](#).
2. Logical Reasoning Chain: The architecture suggests a historic Italian setting. The canal and boats indicate a city centered on waterways. Together, these cues are most [consistent with Venice, Italy](#).
3. Conflict Resolution: No conflicting visual cues are present.
4. Conclusion Justification: The canal setting and [Venetian-style historic architecture](#) strongly suggest Venice, Italy.

Japan, Kyoto



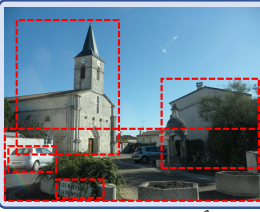
"Box": [67, 139, 138, 317], "Analysis": "The red lettering resembling Japanese characters suggests a Japanese commercial or cultural setting." "Box": [119, 19, 159, 209], "Analysis": "The [Kyoto Tower is a prominent landmark in Kyoto, Japan](#), identifiable by its distinctive design." "Box":
1. Core Evidence Synthesis: Landmark: [Kyoto Tower is a strong indicator of Kyoto](#). Architecture: The surrounding buildings reflect urban Japanese design. Infrastructure: The bus stop and signage are consistent with Japanese urban infrastructure. [Geographic Feature Correlation:](#) These features point to [Japan, specifically Kyoto](#).
2. Logical Reasoning Chain: Kyoto Tower suggests Kyoto directly. The architectural style and urban infrastructure support a Japanese city setting. Together, these cues are most [consistent with Kyoto, Japan](#).
3. Conclusion Justification: The presence of Kyoto Tower and the surrounding Japanese urban environment strongly [suggest Kyoto, Japan](#).

Ukraine, Kremenchuk



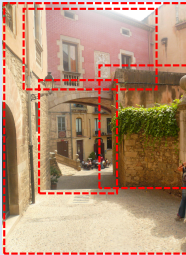
"Box": [798, 326, 1420, 2008], "Analysis": "The left tower displays the text 'СЛАВСЯ УКРАЇНО' together with the Ukrainian trident symbol, clearly indicating a connection to Ukraine." "Box": [2035, 348, 2645, 2018], "Analysis": "The right tower banner welcomes visitors to [Kremenchuk, explicitly pointing to a city in Ukraine](#)." "Box":
1. Core Evidence Synthesis: Text and Symbols: The Ukrainian-language slogan and the national trident symbol strongly indicate Ukraine. Place Name: [The banner directly mentions Kremenchuk, providing a specific city cue](#). Scene Context: The bridge entrance and passing vehicles suggest an active urban transportation route. [Geographic Feature Correlation:](#) The national symbols identify the country as [Ukraine, while the city name narrows the location to Kremenchuk](#).
2. Logical Reasoning Chain: The Ukrainian text and trident symbol establish the national context. The explicit reference to Kremenchuk identifies the city directly. The bridge setting and traffic further support that this is a real urban location rather than a symbolic display. Together, these cues are most consistent with [Kremenchuk, Ukraine](#).
3. Conclusion Justification: The combination of Ukrainian national symbols, Ukrainian-language text, and the direct mention of Kremenchuk strongly supports the conclusion that the image is from Kremenchuk, Ukraine.

France, Les Mathes



"Box": [68, 179, 1806, 2368], "Analysis": "The image shows a stone church with Gothic features and a clock tower, [consistent with France](#)." "Box": [2036, 1188, 4024, 2628], "Analysis": "The adjacent building suggests a rural town setting." "Box":
1. Core Evidence Synthesis: Architecture: [The church's Gothic style is consistent with French religious architecture](#). Place Name: The sign [Les Mathes La Palmyre Zoo' directly points to Les Mathes](#). Scene Context: The surrounding buildings and vehicles suggest a maintained rural commune. [Geographic Feature Correlation:](#) These cues point to France, specifically Les Mathes.
2. Logical Reasoning Chain: The Gothic church suggests a historic French setting. The sign provides a direct location clue. [The rural surroundings further support Les Mathes, France](#).
3. Conclusion Justification: The French-style church and the direct mention of Les Mathes strongly suggest Les Mathes, France.

Spain, Girona



"Box": [498, 0, 1867, 1288], "Analysis": "The [pink brickwork and decorative stone facade suggest a Mediterranean architectural style consistent with Catalonia, Spain](#)." "Box": [0, 0, 560, 2858], "Analysis": "The stone construction and arched entryways reflect medieval European architecture." "Box":
1. Core Evidence Synthesis: Architecture: The warm-toned brickwork, decorative stone facade, arches, and stone walls suggest a historic Mediterranean urban setting. Environment: The ivy-covered wall and outdoor dining area add to the impression of a preserved, tourist-friendly historic town. [Geographic Feature Correlation:](#) [These features are consistent with Catalan Gothic and other historic architectural styles found in northeastern Spain, especially Girona](#).
2. Logical Reasoning Chain: The building materials and facade details point to a Mediterranean historic city. The stone construction and arched entryways support a medieval European context. The outdoor seating and overall atmosphere fit a well-preserved Spanish old town. [Together, these cues are most consistent with Girona, Spain](#).
3. Conclusion Justification: The combination of Mediterranean-style architecture, medieval stone elements, and the historic urban atmosphere [strongly suggests Girona, Spain](#).

Figure 18: More correct cases of HoloGeo on LandmarkBias-3K.

India, Bangalore

HoloGeo

"Box": [0, 0, 1036, 756], "Analysis": "The image shows English signage, dense urban traffic, common [Indian vehicles](#), tropical greenery, and buildings consistent with an [Indian cityscape](#). These features suggest a busy warm-climate urban area in India." "Box":
1. Core Evidence Synthesis: English signage, urban building style, heavy traffic, locally common vehicles, and tropical vegetation together indicate a [dense city environment in India](#).
2. Logical Reasoning Chain: The English signage suggests a prominently English-using urban setting. The traffic, vehicles, and architecture point to a [large Indian city](#). The greenery and tropical trees suggest [southern India](#). Together, these cues are consistent with Bangalore.
3. Conflict Resolution: There are no major conflicting visual cues; the evidence consistently supports an [urban Indian location](#) with a warm climate.
4. Conclusion Justification: The combination of signage, traffic patterns, vehicle types, architecture, and vegetation supports the conclusion that [the image is from India](#), with Bangalore as the proposed city.
"country": "India", "city": "Bangalore"

Austria, Salzburg

HoloGeo

"Box": [0, 0, 756, 917], "Analysis": "The image shows ornate European-style buildings, German signage, a sign for 'Hotel Goldener Hirsch,' and a mountainous backdrop. These features suggest a [historic city center in Austria](#)." "Box": [368, 0, 756, 917], "Analysis": "The right side of the image shows traditional European-style buildings with ornate signage, typical of historic city centers like Salzburg." "Box":
1. Core Evidence Synthesis: The main cues are traditional European facades, German-language signage, the 'Hotel Goldener Hirsch' sign, and nearby mountains.
2. Logical Reasoning Chain: The architecture and signage point to a historic German-speaking European city. [The mountains narrow it toward Austria](#). The 'Hotel Goldener Hirsch' sign makes Salzburg a plausible match.
3. Conflict Resolution: There are no major conflicting visual cues; the architectural, linguistic, and geographic elements support the same interpretation.
4. Conclusion Justification: The combination of historic European architecture, German signage, and mountainous surroundings supports the conclusion [that the image is from Austria](#), with Salzburg as the proposed city.
"country": "Austria", "city": "Salzburg"

Bosnia and Herzegovina, Mostar

HoloGeo

"Box": [0, 12, 1036, 644], "Analysis": "The image shows a historic town with a fortress-like structure, stone houses, and a minaret. These elements suggest a [Balkan setting](#) with strong Ottoman and Islamic architectural influence." "Box":
1. Core Evidence Synthesis: The main cues are medieval-style fortifications, traditional stone houses with Balkan features, and a visible minaret indicating Islamic cultural influence.
2. Logical Reasoning Chain: The fortress and stone residential buildings point to a historic Balkan town. The minaret and Ottoman-style architecture further narrow it to [Bosnia and Herzegovina](#). Together, these features are consistent with [Mostar](#).
3. Conflict Resolution: There are no major conflicts among the visual clues; the architectural elements support the same regional interpretation.
4. Conclusion Justification: The combination of historic fortification, Ottoman-influenced buildings, and Islamic religious architecture supports the conclusion that the image is from [Bosnia and Herzegovina](#), with [Mostar](#) as the proposed city.
"country": "Bosnia and Herzegovina", "city": "Mostar"

Portugal, Porto

HoloGeo

"Box": [0, 52, 934, 667], "Analysis": "The image shows traditional rabelo boats, the Dom Luis I Bridge, and a historic riverside urban landscape. These features strongly [indicate Porto, Portugal](#)." "Box":
1. Core Evidence Synthesis: The main cues are the traditional rabelo boats, the distinctive steel arch of the Dom Luis I Bridge, and the historic cityscape along the river.
2. Logical Reasoning Chain: The boats point to the Douro River region and Porto's wine-transport heritage. The bridge is a [strong landmark cue for Porto](#). The surrounding historic urban landscape further matches the city.
3. Conflict Resolution: There are no major conflicting visual cues; the boats, bridge, and cityscape all support the same location.
4. Conclusion Justification: The combination of rabelo boats, the Dom Luis I Bridge, and the riverside historic setting supports the conclusion that the image is [from Portugal](#), with [Porto as the proposed city](#).
"country": "Portugal", "city": "Porto"

United Kingdom, London

HoloGeo

"Box": [0, 0, 756, 1036], "Analysis": "The image shows a dense urban street with modern glass buildings, multiple cars, and distinctive taxis. These features suggest a [major city in the United Kingdom](#), especially [London](#)." "Box": [109, 474, 251, 636], "Analysis": "The iconic red taxi seen in the image is a [well-known symbol of London](#)." "Box": [474, 465, 637, 640], "Analysis": "The black cab with the license plate 'LX51 CFCZ' is another hallmark of [London's transportation system](#)." "Box": [0, 0, 756, 559], "Analysis": "The overall urban environment, including the mix of historical and modern architecture, is [typical of central London](#)." "Box":
1. Core Evidence Synthesis: The main cues are the modern urban skyline, busy street traffic, and the presence of [iconic taxis associated with London](#).
2. Logical Reasoning Chain: The mix of modern architecture and dense traffic points to a major global city. The distinctive taxis strongly narrow it to London. Together, these features are [consistent with central London](#).
3. Conflict Resolution: There are no major conflicting visual cues; the transportation and cityscape elements support the same interpretation.
4. Conclusion Justification: The combination of recognizable taxis, urban traffic, and [London-style architecture](#) supports the conclusion that the image is from the United Kingdom, with London as the proposed city.
"country": "United Kingdom", "city": "London"

Russia, Saint Petersburg

HoloGeo

"Box": [168, 24, 1036, 281], "Analysis": "The image shows a city street with neoclassical-style buildings, public transportation, pedestrians, and street vending. These elements suggest a historic European-style urban [setting consistent with Russia](#)." "Box":
1. Core Evidence Synthesis: The main cues are the neoclassical architecture, cohesive historic urban layout, presence of buses, and everyday street activity.
2. Logical Reasoning Chain: The architecture and city layout point to a historic European-style city. The building style is consistent with [Russian urban centers](#), and the overall scene supports a [large Russian city](#). Together, these features are consistent with Saint Petersburg.
3. Conflict Resolution: There are no major conflicting visual cues; the architecture, transport, and street scene support the same interpretation.
4. Conclusion Justification: The combination of neoclassical buildings, urban layout, and street activity supports the conclusion that [the image is from Russia](#), with Saint Petersburg as the proposed city.
"country": "Russia", "city": "Saint Petersburg"

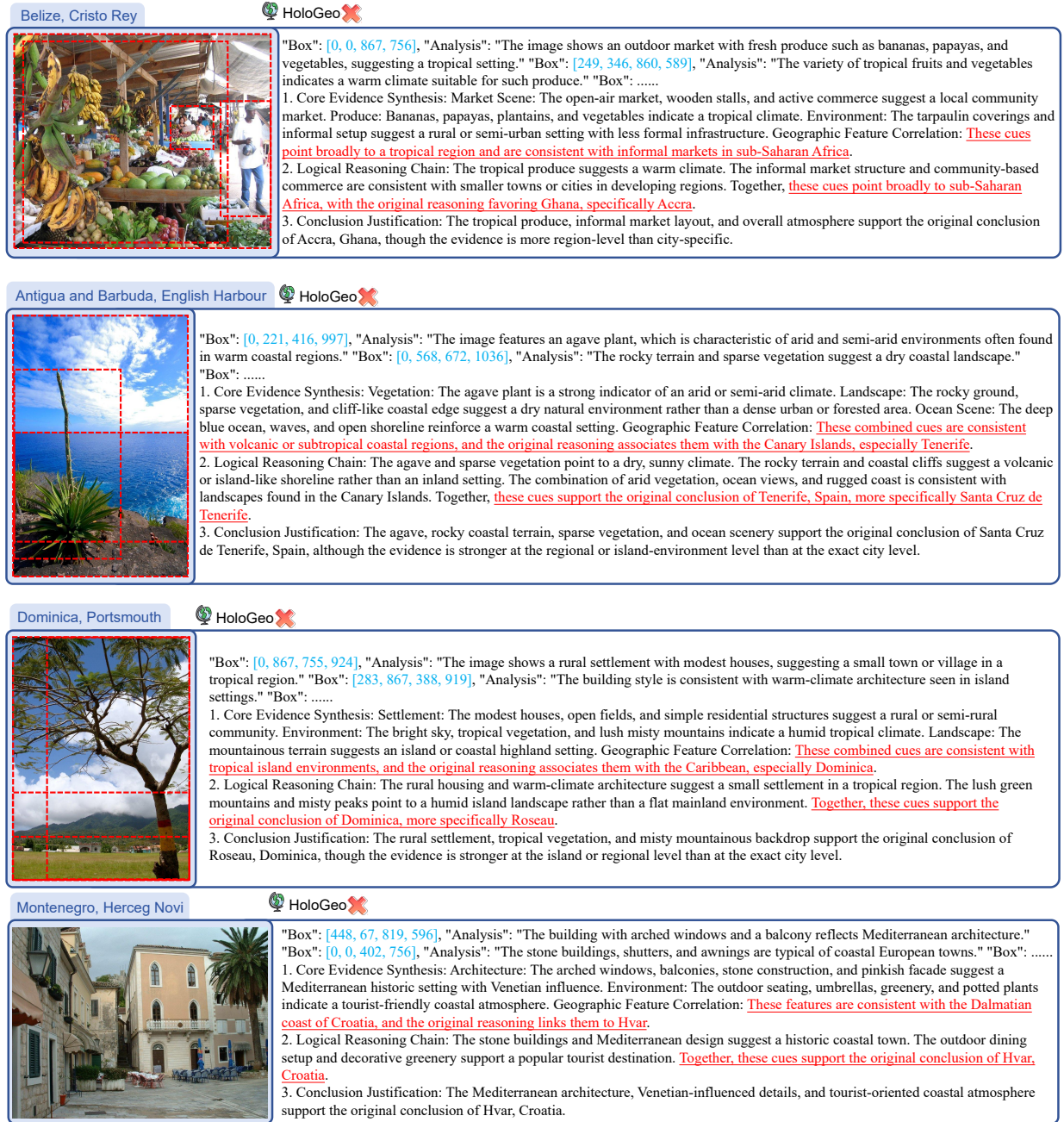


Figure 20: Some failure cases of HoloGeo on IM2GPS.