GeoRouter: Dynamic Paradigm Routing for Worldwide Image Geolocalization

Pengyue Jia[1], **Derong Xu**[1], **Yingyi Zhang**[1], **Xiaopeng Li**[1], **Wenlin Zhang**[1]
Yi Wen[1], **Liu Lan**[2], **Yuanshao Zhu**[1], **Xuetao Wei**[3], **Wanyu Wang**[1], **Xiangyu Zhao**[1*]
[1]Department of Data Science, City University of Hong Kong, [2]Didi International Business Group,
[3]Computer Science and Engineering, Southern University of Science and Technology,
jia.pengyue@my.cityu.edu.hk,xianzhao@cityu.edu.hk

Abstract

Worldwide image geolocalization aims to predict precise GPS coordinates for images captured anywhere on Earth, which is challenging due to the large visual and geographic diversity. Recent methods mainly follow two paradigms: retrieval-based approaches that match queries against a reference database, and generation-based approaches that directly predict coordinates using Large Vision-Language Models (LVLMs). However, we observe distinct error profiles between them: retrieval excels at fine-grained instance matching, while generation offers robust semantic reasoning. This complementary heterogeneity suggests that no single paradigm is universally superior. To harness this potential, we propose **GeoRouter**, a dynamic routing framework that adaptively assigns each query to the optimal paradigm. GeoRouter leverages an LVLM backbone to analyze visual content and provide routing decisions. To optimize GeoRouter, we introduce a distance-aware preference objective that converts the distance gap between paradigms into a continuous supervision signal, explicitly reflecting relative performance differences. Furthermore, we construct **GeoRouting**, the first large-scale dataset tailored for training routing policies with independent paradigm predictions. Extensive experiments on IM2GPS3K and YFCC4K demonstrate that GeoRouter consistently outperforms state-of-the-art baselines across all distance thresholds. Our dataset, code, and checkpoints are publicly available[1] to facilitate future research.

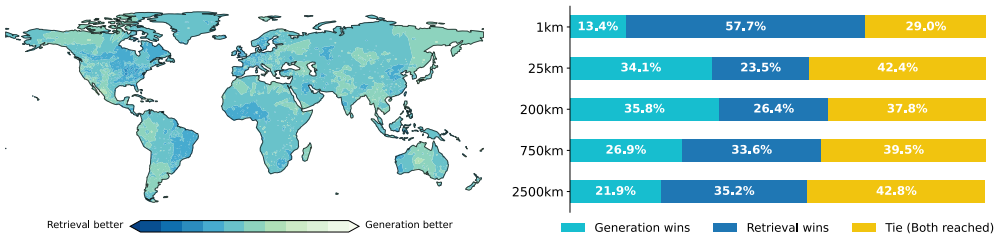
1 Introduction

Worldwide image geolocalization [1, 2, 3] aims to determine the precise geographic coordinates of an image captured anywhere on Earth [4]. In contrast to restricted localization tasks that limit the space to specific cities or known landmarks [5, 6, 7, 8, 9], this global scale approach supports a broader range of open-world applications, including crime tracking [10, 11, 12, 13, 14], environmental monitoring [15], and autonomous navigation [16]. However, achieving high global accuracy remains challenging due to immense environmental diversity. Models must distinguish visually similar but geographically distant scenes across varied climates, vegetation, and cultures [17, 18, 19, 20].

Recent state-of-the-art methods generally adhere to two paradigms: (1) retrieval-based methods [21, 22, 23], which predict coordinates by matching query images against a reference database, and (2) generation-based methods [24, 25, 26, 27], which utilize Large Vision-Language Models (LVLMs) to directly predict locations [28, 29], invoke external tools [30, 31, 32], or apply semantic reasoning for spatial inference [33, 34, 35, 36, 37, 38]. To systematically compare the capabilities of these paradigms, we conduct an empirical study on the MP16 dataset [39], deploying GeoRanker [23] and

*Corresponding author.
[1]https://github.com/Applied-Machine-Learning-Lab/NeurIPS2026_GeoRouter

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Spatial Performance Heterogeneity

(b) Win-Tie Distribution across Distance Scales

Figure 1: Paradigm Complementarity in Worldwide Image Geolocalization. (b) groups queries into mutually exclusive Oracle-error intervals and excludes those with both errors above 2500km.

Gemini-2.5-pro [40] as representative models for retrieval and generation, respectively. As illustrated in Figure 1(a), our analysis reveals a substantial spatial performance heterogeneity, where the better-performing paradigm varies across queries and neither paradigm dominates across geographic regions. Furthermore, the quantitative breakdown in Figure 1(b) reveals two key patterns. First, retrieval-based methods achieve higher success rates at fine-grained (1km) and coarse (750km, 2500km) thresholds, while generation-based methods perform better at intermediate levels (25km, 200km). Second, at any given threshold, both paradigms each win on a substantial fraction of queries, confirming that neither method consistently dominates. These observations indicate that the two paradigms exhibit complementary error profiles rather than simply overlapping in capability. Despite this, existing research [41, 42, 43, 44] primarily focuses on the isolated optimization of individual paradigms, overlooking the potential for holistic integration. This observation motivates a critical research question: *How can we dynamically route each query to the most suitable paradigm to break the performance ceiling of single-model systems?*

To address this challenge, we introduce GeoRouter, a unified framework designed to tackle the novel task of geolocalization routing. Specifically, GeoRouter leverages an LVLM as its backbone to exploit advanced visual understanding and spatial reasoning capabilities. By analyzing the semantic content of the given contextual information, the model predicts a scalar routing score that adaptively assigns the query to either the retrieval-based or the generation-based paradigm. To robustly optimize this decision mechanism, we propose the Distance-Aware Preference Optimization (DisPO) objective. Unlike standard binary classification, which treats routing as a discrete choice, DisPO transforms the geodesic distance disparity between paradigm predictions into a continuous supervision signal. This enables the model to discern not only which paradigm provides a more accurate prediction but also the magnitude of the performance gain, ensuring precise and calibrated routing decisions.

To facilitate the training of GeoRouter, we construct GeoRouting, the first large-scale dataset designed for the geolocalization routing task. Unlike standard geolocalization datasets that only provide ground-truth coordinates, GeoRouting includes performance comparisons between paradigms. This rich supervision allows the model to learn the comparative advantage of each paradigm across diverse queries, rather than focusing on direct coordinate prediction. Consequently, GeoRouting can serve as a foundational benchmark to standardize and advance the study of dynamic model selection in image worldwide geolocalization. We conduct extensive experiments on two well-established benchmarks, IM2GPS3K [45] and YFCC4K [46], to validate the effectiveness of our approach. The results demonstrate that GeoRouter achieves high routing accuracy and substantially improves overall geolocalization performance. Furthermore, we perform comprehensive ablation studies and hyperparameter analyses to verify the necessity of each component and to assess the impact of different configurations on the final results. Our contributions can be summarized as follows:

1. We systematically reveal the performance complementarity between retrieval and generation paradigms and formulate the novel routing task to break single-model performance ceilings.

2. We propose GeoRouter, a VLM-based framework utilizing DisPO to enable continuous confidence modeling for robust and precise dynamic routing.

3. We release GeoRouting, the first large-scale routing dataset for worldwide image geolocalization, establishing a standard for this emerging research direction.

4. Extensive experiments on IM2GPS3K and YFCC4K demonstrate that GeoRouter achieves comprehensive improvements over baselines, proving the efficacy of the dynamic routing paradigm.

2 Related Work

2.1 Image Geolocalization

Image geolocalization [47, 48] aims to predict the precise geographic coordinates of an image captured anywhere on Earth, and it is an important interdisciplinary task at the intersection of computer vision [49, 50] and geographic artificial intelligence [51, 52, 53, 54, 55, 56, 57]. Early approaches [58, 59, 60, 61, 62] formulated this problem as a *classification task* by dividing the Earth into discrete geographic grids. These models are trained to predict the specific grid cell containing the query image, outputting its center coordinate as the final prediction. However, this strategy suffers from inherent errors, as the actual location can be far from the grid center even when the classification is correct. More recent methodologies generally diverge into two main paradigms: retrieval-based methods and generation-based methods. (1) *Retrieval-based methods* [63, 64, 65, 66, 67, 68, 69] treat geolocalization as a similarity search problem. They maintain a reference database of geotagged images [43, 23, 70, 71, 72] or GPS entries [22] and identify the most similar candidates for a given query image. These approaches often achieve high precision for previously seen landmarks, but they may struggle with unseen views or visually similar yet geographically distant regions. (2) *Generation-based methods* [24, 25, 26, 27] rely on LVLMs to directly infer geographic locations. This paradigm has recently evolved into agentic frameworks, where models utilize internal semantic reasoning [33, 34, 35, 36] or invoke external search tools [31] to deduce coordinates from visual clues like climate, architecture, and vegetation. Nevertheless, these models can suffer from spatial hallucinations and struggle to achieve fine-grained spatial precision [73]. GeoRouter diverges from these existing categories by proposing a new framework that dynamically routes each query to the most suitable paradigm, thereby breaking the performance ceiling of single-model systems.

2.2 Routing and Dynamic Model Selection

The concept of dynamic model selection is grounded in the algorithm selection problem, which posits that no single algorithm is universally optimal for all problem instances [74, 75]. Instead, distinct algorithms exhibit performance complementarity, where their relative effectiveness varies based on the specific characteristics of the input. Recently, this methodology has gained significant traction in the domain of Large Language Models (LLMs). Known as LLM routing, the objective is to dynamically direct user queries to the most suitable component—typically selecting between a capability-rich proprietary model and a cost-effective open-source counterpart—to optimize resource consumption without compromising response quality [76]. Existing literature primarily categorizes these approaches into four streams: (1) *Similarity-based Routing Methods* [77, 78] route queries by clustering vector embeddings to identify semantically similar historical interactions. (2) *Supervised Routing Methods* [79, 80, 81] train standalone classifiers on query-performance profiles to explicitly predict the optimal model for specific domains. (3) *Reinforcement Learning-based Routing Methods* [82, 83] formulate routing as a sequential decision-making process, learning optimal policies through trial-and-error feedback [84, 85]. (4) *Generative Routing Methods* [86, 87] leverage LLM reasoning capabilities to directly generate routing decisions via instruction tuning or in-context learning. GeoRouter is the first framework to dynamically route between retrieval and generation paradigms, synergizing their complementary strengths to achieve superior geolocalization accuracy.

3 Methodology

In this section, we present the details of GeoRouter, our dynamic routing framework for worldwide image geolocalization. Figure 2 provides an overview of the framework, which consists of two stages: training and inference. The training stage begins with dataset construction in Section 3.1, where we build the GeoRouting dataset to support routing supervision. It then proceeds to the model architecture in Section 3.2 and the optimization in Section 3.3. During the inference stage (Section 3.4), GeoRouter analyzes the input image and predicts a routing score to select the most suitable paradigm for each query.

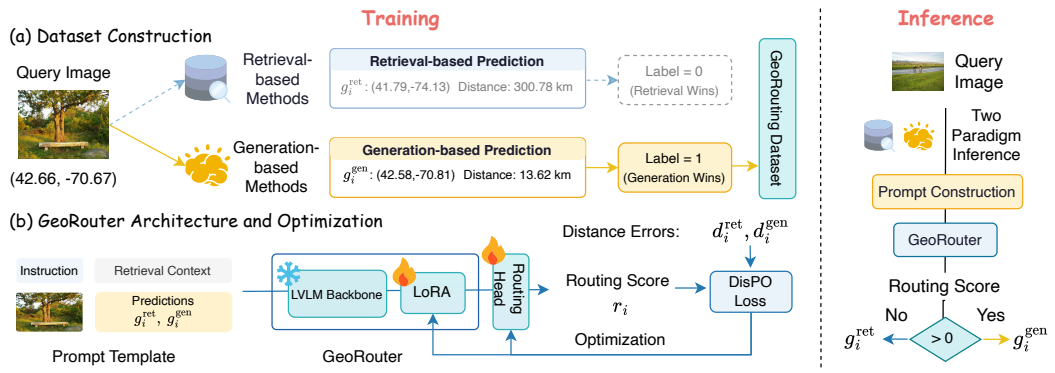


Figure 2: Overview of the dynamic routing framework – GeoRouter.

3.1 Dataset Construction

Training GeoRouter requires supervision that reflects the relative performance of the two paradigms on each query. We construct this supervision based on existing geotagged image databases (e.g., MP16 [39]), from which we sample a set of query images. Each sampled query consists of an image I_i and its ground-truth geographic coordinate g_i. For each query I_i, we generate spatial predictions using two distinct paradigms: a retrieval-based method $\mathcal{R}$ and a generation-based method $\mathcal{G}$. The retrieval method extracts a set of similar candidates $\mathcal{C}_i = \{c_i^1, \ldots, c_i^K\} = \mathcal{R}(I_i)$, where each candidate c_i^k denotes a database entry associated with an image $I_{c_i^k}$ and its geographic coordinate $\text{gps}_{c_i^k}$. The location of the top-ranked candidate c_i^1 is taken as the retrieval prediction:$g_i^{\text{ret}} = \text{gps}_{c_i^1}$. Concurrently, the generation method employs an LVLM to directly infer a coordinate prediction: $g_i^{\text{gen}} = \mathcal{G}(I_i)$. We calculate the geolocalization error for both methods against the ground-truth coordinate as $d_i^{\text{ret}} = \mathcal{D}(g_i^{\text{ret}}, g_i)$ and $d_i^{\text{gen}} = \mathcal{D}(g_i^{\text{gen}}, g_i)$, where $\mathcal{D}(\cdot,\cdot)$ represents the geographic distance function. The binary routing label is defined as $y_i = \mathbf{1}[d_i^{\text{gen}} < d_i^{\text{ret}}]$, where $y_i = 1$ indicates generation is preferred and $y_i = 0$ indicates retrieval is preferred. Consequently, we format each training instance in the GeoRouting dataset as the tuple $(I_i, g_i^{\text{ret}}, g_i^{\text{gen}}, d_i^{\text{ret}}, d_i^{\text{gen}}, \mathcal{C}_i, y_i)$ to provide comprehensive supervision for the routing model.

3.2 GeoRouter

Figure 2(b) illustrates the overall architecture of GeoRouter. Existing methods typically follow a single paradigm, either relying entirely on retrieval-based matching [43, 22, 23] or solely on generation-based reasoning [34], which overlooks their complementary strengths and limits overall performance. To address this limitation, GeoRouter introduces a dynamic routing mechanism that formulates paradigm selection as a learnable decision problem. Instead of directly predicting geographic coordinates, GeoRouter evaluates the query image and estimates which paradigm is more reliable for the given input.

Prompt Template. To enable the LVLM to perform paradigm selection, the query image and auxiliary contextual information are assembled into a unified prompt template. The prompt explicitly frames routing as a decision task and presents predictions from two paradigms and candidate information as reference inputs. The prompt template is shown as follows:

> Task: Decide whether to use generation or retrieval for this image's geolocalization. Query: {Query Image I_i} Generation-based Prediction: {Generated GPS Coordinate g_i^{gen}} Retrieval-based Prediction: {Top-1 Retrieved GPS Coordinate g_i^{ret}}{Top-1 Retrieved Image $I_{c_i^1}$} Other Retrieved Candidate Coordinates: {Top-K Retrieved GPS List (excluding top-1) $\{\text{gps}_{c_i^k}\}_{k=2}^{K}$}

Formally, the assembled prompt is denoted as $\mathbf{x}_i = \mathcal{P}(I_i, g_i^{\text{gen}}, g_i^{\text{ret}}, \mathcal{C}_i)$, where $\mathcal{P}(\cdot)$ represents the prompt construction function. This structured design allows the model to jointly consider visual content, predicted coordinates, and retrieval context when estimating the relative reliability of the

two paradigms. It is worth noting that variants of this template under different input settings (e.g., using only the query image) are also evaluated, with a detailed analysis provided in Section 4.3.

Architecture. The constructed prompt input $\mathbf{x}_i$ is fed into a pretrained LVLM backbone to jointly encode visual and textual information. To adapt the model to the routing task in a parameter-efficient manner, Low-Rank Adaptation (LoRA) modules [88] are injected into selected transformer layers during training, while the majority of backbone parameters remain frozen.

We use the hidden state of the last token in the final layer as the joint representation of the input: Let $\mathbf{u}_i = \text{LVLM}(\mathbf{x}_i)_{[-1]} \in \mathbb{R}^m$, where m is the hidden dimension. A linear routing head then projects $\mathbf{u}_i$ to a scalar value by $r_i = \boldsymbol{\theta}^\top \mathbf{u}_i$, where $\boldsymbol{\theta} \in \mathbb{R}^m$ is the trainable parameter of the routing head. The resulting scalar $r_i \in \mathbb{R}$ indicates the model's preference between the retrieval-based and generation-based paradigms.

3.3 Optimization

To effectively train GeoRouter, relying on standard binary classification with hard labels ignores the continuous nature of geographic errors. Therefore, we introduce a distance-aware optimization method that converts the geographic distance disparity between the two paradigms into continuous soft labels. This strategy allows the model to learn not only which paradigm performs better but also the magnitude of the performance gap, leading to more calibrated and robust routing decisions.

Distance-Aware Preference Optimization. Given a training instance $\mathbf{x}_i$, let d_i^{ret} and d_i^{gen} denote the geographic prediction errors for the retrieval and generation paradigms, respectively. Instead of using the discrete binary label, we compute a continuous preference score. First, we calculate the logarithmic difference between the two errors to capture their relative scale:

$$\Delta_i = \log(d_i^{\text{ret}} + \epsilon) - \log(d_i^{\text{gen}} + \epsilon), \tag{1}$$

where ϵ is a small constant included to ensure numerical stability. Next, we apply a sigmoid function to transform this unbounded difference into a soft probability label $p_i \in (0, 1)$:

$$p_i = \sigma(\alpha \cdot \Delta_i) = \frac{1}{1 + \exp(-\alpha \cdot \Delta_i)}, \tag{2}$$

where α is a scaling hyperparameter that controls the steepness of the soft label assignment. Under this formulation, when the generation method is significantly more accurate than the retrieval method ($d_i^{\text{gen}} \ll d_i^{\text{ret}}$), Δ_i becomes a large positive value, pushing p_i toward 1. Conversely, when the retrieval method is superior, Δ_i becomes negative, and p_i approaches 0. According to Section 3.2, r_i denotes the scalar logit predicted by the routing head of GeoRouter for the i-th query. We optimize the model using a weighted binary cross-entropy loss function computed over the soft labels: $\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[p_i \log(\sigma(r_i)) + (1 - p_i)\log(1 - \sigma(r_i))]$, where N is the batch size and $\sigma(r_i)$ represents the predicted probability of selecting the generation-based paradigm. By directly minimizing this distance-aware objective, the framework explicitly aligns the model's continuous confidence with the actual geographic performance gap between the two paradigms.

3.4 Inference

During the inference stage, the primary objective is to dynamically assign each query to the most suitable paradigm to maximize accuracy. Given a new query image I_i, the framework first collects its associated contextual information, including the retrieval-based prediction g_i^{ret}, the generation-based prediction g_i^{gen}, and the retrieved candidate set $\mathcal{C}_i$, to construct the input prompt $\mathbf{x}_i$. GeoRouter then processes this prompt to compute a continuous routing score by $r_i = \text{GeoRouter}(\mathbf{x}_i)$ and assigns the final prediction as $\hat{g}_i = g_i^{\text{gen}}$ if $r_i > 0$, and $\hat{g}_i = g_i^{\text{ret}}$ otherwise. The ultimate geographic prediction for a set of queries is the aggregation of these individually routed coordinates, ensuring that the system leverages the complementary strengths of both paradigms.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. To build the GeoRouting dataset for training, we adopt the MP16-Pro database [26] as our source of reference images and geographic coordinates. For testing, we

Table 1: **Main results** on the IM2GPS3K and YFCC4K benchmarks. Higher values indicate better performance. The best and second-best results are highlighted in **bold** and underlined, respectively. Δ denotes the relative improvement of GeoRouter over the best baseline.

Methods		IM2GPS3K					YFCC4K				
		Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
[L]kNN, sigma=4 [48]	ICCV'17	7.2	19.4	26.9	38.9	55.9	2.3	5.7	11.0	23.5	42.0
PlaNet [58]	ECCV'16	8.5	24.8	34.3	48.4	64.6	5.6	14.3	22.2	36.4	55.8
CPlaNet [59]	ECCV'18	10.2	26.5	34.6	48.6	64.6	7.9	14.8	21.9	36.4	55.5
ISNs [60]	ECCV'18	10.5	28.0	36.6	49.7	66.0	6.5	16.2	23.8	37.4	55.0
Translocator [61]	ECCV'22	11.8	31.1	46.7	58.9	80.1	8.4	18.6	27.0	41.1	60.4
GeoDecoder [62]	ICCV'23	12.8	33.5	45.9	61.0	76.1	10.3	24.4	33.9	50.0	68.7
GeoCLIP [22]	NeurIPS'23	14.11	34.47	50.65	69.67	83.82	9.59	19.31	32.63	55.00	74.69
Img2Loc [25]	SIGIR'24	15.34	39.83	53.59	69.70	82.78	19.78	30.71	41.40	58.11	74.07
PIGEON [43]	CVPR'24	11.30	36.70	53.80	72.40	85.30	10.4	23.70	40.60	62.20	77.70
G3 [26]	NeurIPS'24	16.65	40.94	55.56	71.24	84.68	23.99	35.89	46.98	64.26	78.15
GeoToken [27]	ICDM'25	16.80	39.60	53.80	70.80	85.00	24.30	35.30	46.60	64.20	78.60
GeoRanker [23]	NeurIPS'25	18.79	45.05	61.49	76.31	89.29	32.94	43.54	54.32	69.79	82.45
GeoBayes [37]	AAAI'26	6.30	34.70	53.60	73.70	85.90	4.90	16.10	30.90	55.80	75.40
Geo-R [28]	AAAI'26	18.10	41.53	58.31	75.33	86.42	10.47	22.67	40.04	60.83	75.84
GeoRouter	Ours	**20.82**	**50.48**	**65.73**	**80.35**	**90.66**	**32.98**	**46.01**	**57.52**	**72.02**	**83.02**
Rel. Improvement	Δ	↑ 10.80%	↑ 12.05%	↑ 6.90%	↑ 5.29%	↑ 1.53%	↑ 0.12%	↑ 5.67%	↑ 5.89%	↑ 3.20%	↑ 0.69%

align our evaluation with standard geolocalization literature [22, 23, 25] and evaluate GeoRouter on two widely used benchmarks: IM2GPS3K [45] and YFCC4K [46]. To quantify prediction accuracy, we compute the geodesic distance between the predicted locations and the ground-truth coordinates. The overall performance is reported as the percentage of queries successfully localized within five predefined distance thresholds: 1km, 25km, 200km, 750km, and 2500km. The routing accuracy is also analyzed in Section 4.5.

GeoRouting Dataset. To support the training and evaluation of routing models in worldwide image geolocalization, we construct the GeoRouting dataset. Specifically, we sample **100K instances** from the MP16-Pro database to serve as the queries. For each query, we generate predictions using both the retrieval-based and generation-based paradigms. We then compute the distance between each prediction and the ground-truth coordinate to determine a binary routing label indicating the more suitable paradigm. Consequently, each complete data sample contains the query image data, ground-truth GPS, retrieved candidate context, predictions from both paradigms, their respective distance errors, and the binary routing label. By publicly releasing this dataset, we aim to provide a standard benchmark that advances research in geolocalization routing and multimodal decision-making tasks. Detailed examples of the data entries are provided in **Appendix A** for reference.

Implementation Details. To represent the generation-based paradigm, we select Gemini-2.5-Flash[2] due to its strong performance on existing image geolocalization benchmarks [2, 89]. For the retrieval-based paradigm, we adopt the current state-of-the-art retrieval-based method, GeoRanker [23], utilizing its default configurations without generated candidates. We implement the LVLM backbone of GeoRouter using Qwen2-VL-7B-Instruct[3]. During training, we apply LoRA to efficiently fine-tune the backbone, specifically targeting the `q_proj`, `k_proj`, and `v_proj` modules. We configure LoRA with a rank of 16, a scaling factor of 32, and a dropout rate of 0.05. GeoRouter is optimized using the AdamW optimizer [90] with a learning rate of 1e-4 and a batch size of 24, and it is trained for 3 epochs. For the distance-aware preference optimization, we set the numerical stability constant ϵ to 1e-6 and the scaling hyperparameter α to 1.6. More details are given in **Appendix B**.

Baselines. To evaluate the effectiveness of our approach, we conduct comparative experiments with 14 baseline methods: [L]kNN, sigma=4 [48], PlaNet [59], CPlaNet [59], ISNs [60], Translocator [61], GeoDecoder [62], GeoCLIP [22], Img2Loc [25], PIGEON [43], G3 [26], GeoToken [27], GeoRanker [23], GeoBayes [37], and Geo-R [28]. Detailed descriptions of baselines are provided in **Appendix C**.

4.2 Main Results

To validate the effectiveness of the proposed routing framework, we compare GeoRouter against state-of-the-art geolocalization methods on the IM2GPS3K and YFCC4K benchmarks. As shown in Table 1, the experimental results reveal the following main observations. GeoRouter consistently

[2] `https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash`
[3] `https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct`

achieves the highest accuracy across all distance thresholds on both datasets. Specifically, GeoRouter demonstrates its most notable advantages at the city-level spatial scale, outperforming the best baselines by relative margins of 12.05% and 5.67% at the 25km threshold on the IM2GPS3K and YFCC4K datasets, respectively. The results confirm the limitations of static models. Relying exclusively on either retrieval matching (e.g., GeoCLIP, GeoRanker) or semantic generation (e.g., Img2Loc, G3) restricts the overall spatial accuracy. By explicitly evaluating the confidence gap and assigning the query to the optimal paradigm, GeoRouter effectively overcomes this performance ceiling. Additionally, **Appendix D** provides qualitative examples categorized by distance thresholds to illustrate the distribution of queries at varying levels of geolocalization accuracy.

4.3 Ablation Study

To systematically evaluate the contribution of each component in GeoRouter, we conduct an ablation study on both datasets. We define the following variants of our approach: (1) **w/o** DisPO: Our method optimized using a standard Binary Cross-Entropy (BCE) loss with discrete hard labels. (2) **w/o** $\mathcal{C}$: Our method where the auxiliary retrieved candidate information is removed from the input prompt. (3) **w/o** $\mathcal{C}$ & g^{ret}: Our method where both the candidate set and the retrieval-based prediction are removed. (4) **w/o** g^{gen}: Our method where the generation-based prediction is excluded from the prompt. (5) **w/o** Context: Our method where all paradigm predictions and candidate information are removed; the routing decision is based solely on the query image. The quantitative results on IM2GPS3K are presented in Table 2 and the results for YFCC4K are provided in **Appendix E**. From the results we can draw several key observations. (1) The full GeoRouter framework achieves the highest average accuracy, confirming that every proposed component contributes positively to the final performance. (2) Replacing the DisPO with a standard BCE loss (w/o DisPO) leads to a performance drop, particularly at the intermediate and coarse spatial levels (e.g., 25km and 200km). This demonstrates that discrete binary labels fail to capture the geographic magnitude of prediction errors. By utilizing distance-aware optimization, the model learns a more calibrated confidence score, which is essential for accurate paradigm selection. (3) Removing contextual information from the prompt consistently decreases performance. Excluding the candidate set (w/o $\mathcal{C}$) causes a minor drop in accuracy, which indicates that the retrieved candidates provide helpful context for the model to learn routing. Removing either the retrieval contexts (w/o $\mathcal{C}$ & g^{ret}) or the generation prediction (w/o g^{gen}) leads to larger accuracy reductions. Finally, the variant that relies only on the query image (w/o Context) yields the lowest overall performance. These results demonstrate that providing the predicted coordinates allows the model to compare the paradigms directly, rather than making routing decisions without references.

Table 2: Ablation study on IM2GPS3K.

Methods	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km	Average
w/o DisPO	**20.85**	49.38	64.26	79.05	89.76	60.66
w/o $\mathcal{C}$	20.69	49.18	64.60	79.75	90.09	60.86
w/o $\mathcal{C}$ & g^{ret}	19.22	46.11	61.76	77.11	89.29	58.69
w/o g^{gen}	20.22	48.31	64.06	78.75	89.92	60.25
w/o Context	19.29	45.85	61.46	77.41	89.16	58.63
Ours	20.82	**50.48**	**65.73**	**80.35**	**90.66**	**61.60**

4.4 Effect of the Scaling Hyperparameter α

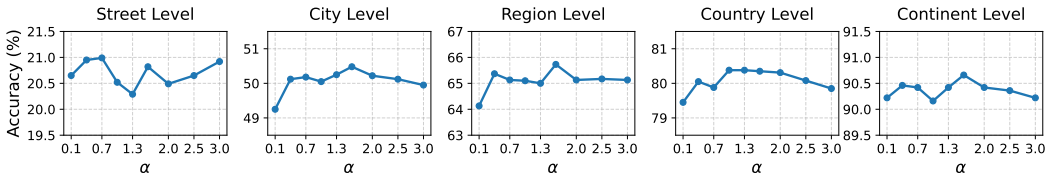


Figure 3: Effect of the hyperparameter α on geolocalization accuracy. The subfigures present the results evaluated on the IM2GPS3K dataset across distance thresholds ranging from 1km to 2500km.

To understand the impact of the scaling hyperparameter α within our Distance-Aware objective, we conduct an experiment by systematically varying its value from 0.1 to 3.0. The hyperparameter α controls the steepness of the soft probability labels generated from the geographic distance disparity between the retrieval and generation paradigms. We present the geolocalization accuracy across all five distance thresholds on the IM2GPS3K dataset in Figure 3. The complete evaluation results for both the IM2GPS3K and YFCC4K datasets are provided in **Appendix F** for reference.

Table 3: Comparison of geolocalization and routing accuracy across different distance thresholds. The best performance among the realistic (non-Oracle) methods is highlighted in **bold**, while the Oracle results are presented in *italics* as the theoretical upper bound.

Dataset	Method	Geolocalization Accuracy (%)						Routing Accuracy (%)					
		1km	25km	200km	750km	2500km	Average	1km	25km	200km	750km	2500km	Average
IM2GPS3K	Pure Retrieval	18.42	43.14	59.29	75.31	88.36	56.90	57.28	48.44	48.48	47.57	48.37	50.03
	Pure Generation	17.55	47.91	63.30	78.85	88.59	59.24	42.72	51.56	51.52	52.43	51.63	49.97
	GeoRouter	**20.82**	**50.48**	**65.73**	**80.35**	**90.66**	**61.61**	**65.52**	**63.36**	**62.09**	**61.61**	**61.30**	**62.78**
	Oracle	*24.29*	*54.55*	*71.37*	*85.09*	*93.89*	*65.84*	*100.00*	*100.00*	*100.00*	*100.00*	*100.00*	*100.00*
YFCC4K	Pure Retrieval	32.47	42.70	53.13	69.95	82.67	56.18	85.23	73.34	67.51	65.47	64.44	71.20
	Pure Generation	10.43	26.85	41.93	59.30	73.28	42.36	14.77	26.66	32.49	34.53	35.56	28.80
	GeoRouter	**32.89**	**46.01**	**57.52**	**72.02**	**83.02**	**58.29**	**86.90**	**80.96**	**77.44**	**72.83**	**70.92**	**77.81**
	Oracle	*35.67*	*49.78*	*62.43*	*77.89*	*89.09*	*62.97*	*100.00*	*100.00*	*100.00*	*100.00*	*100.00*	*100.00*

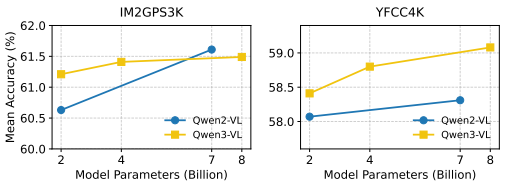


Figure 4: Effect of backbone model scale on geolocalization performance. The evaluation compares Mean Accuracy across different parameter sizes of the Qwen2-VL and Qwen3-VL families.

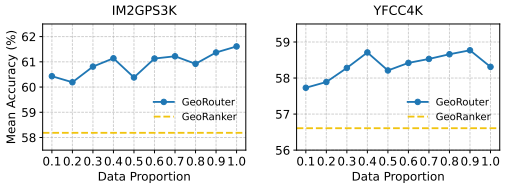


Figure 5: Data efficiency analysis. The figures illustrate the Mean Accuracy achieved when training GeoRouter with varying proportions of the available dataset.

From the visual trends in Figure 3, we draw several key insights regarding the model behavior. First, setting α to a very low value (e.g., 0.1) consistently yields the lowest accuracy across most spatial thresholds. A small α produces an overly flat probability distribution, which prevents the model from effectively capturing the magnitude of the performance gap between the two paradigms. Second, as α increases, the geolocalization performance generally improves, reaching an optimal peak around $\alpha = 1.6$. Finally, when α is set too high (e.g., 2.5 or 3.0), the accuracy begins to gradually decline. An excessively large α causes the sigmoid transformation to approximate a step function, which effectively reverts the continuous soft labels back into discrete hard binary labels. This degradation highlights the necessity of continuous distance modeling and validates the core design of DisPO.

4.5 Effectiveness of Dynamic Routing

To evaluate the effectiveness of the proposed routing framework, we analyze the routing accuracy and the resulting geolocalization performance against static single-paradigm policies (Pure Retrieval and Pure Generation) and a theoretical upper bound (Oracle). As shown in Table 3, we can make three primary observations. First, GeoRouter consistently outperforms both pure baselines across all metrics, proving its effectiveness in identifying the optimal paradigm per query. Second, comparing the two datasets highlights the robustness of GeoRouter across different error distributions. On the IM2GPS3K dataset, the performance of the pure retrieval and pure generation methods is relatively balanced. In contrast, on the YFCC4K dataset, the pure retrieval method exhibits an advantage over the generation method. Even under this imbalanced scenario, GeoRouter successfully adapts to the underlying data distribution and maintains its performance superiority, demonstrating strong robustness. Third, while GeoRouter successfully exceeds the performance limits of single models, a clear gap remains compared to the Oracle bound (e.g., 61.61% versus 65.84% average accuracy on IM2GPS3K). This confirms the routing task's potential and highlights policy optimization as a valuable future direction. We also investigated the transferability of GeoRouter in **Appendix G**.

4.6 Effect of Backbone Model Scale

To investigate how the parameter scale of the LVLM backbone influences the performance, we evaluate GeoRouter using varying sizes of the Qwen2-VL and Qwen3-VL model families. Figure 4 illustrates the mean accuracy on the IM2GPS3K and YFCC4K datasets across different model scales. From these results, we draw two primary conclusions. First, increasing the number of model parameters consistently improves routing performance. For both the Qwen2-VL and Qwen3-VL families, the mean accuracy exhibits a clear upward trend as the parameter count scales from 2

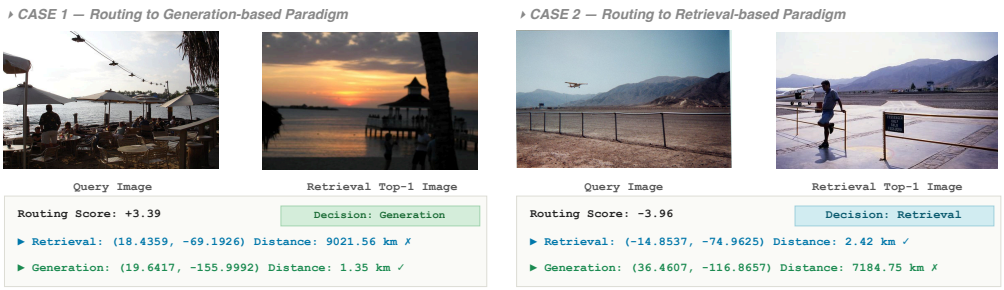


Figure 6: Case studies illustrating GeoRouter's routing decisions across two paradigms.

billion to 7 or 8 billion. Second, advancements in the foundational model architecture provide immediate performance benefits. At equivalent or similar parameter scales (e.g., comparing the 2-billion parameter versions), the newer Qwen3-VL models consistently outperform the older Qwen2-VL models on both datasets. This confirms that the GeoRouter framework can easily scale and benefit from ongoing improvements in state-of-the-art LVLM architectures. The detailed results for each level are provided in **Appendix H**.

4.7 Data Efficiency Analysis

To investigate the impact of training data volume on the learned routing policy, we evaluate GeoRouter using varying proportions of the available dataset. As illustrated in Figure 5, we draw two primary conclusions from the results. First, the mean geolocalization accuracy exhibits a general upward trend as the data proportion increases from 0.1 to 1.0. This indicates that exposure to more training samples steadily improves the routing decisions. Second, the framework demonstrates high data efficiency. Even when restricted to only 10% of the training data (a proportion of 0.1), GeoRouter consistently surpasses the SOTA baseline GeoRanker on both the IM2GPS3K and YFCC4K datasets. This confirms that the model effectively learns the comparative routing mechanism from a very small fraction of the data, while continuing to benefit from the broader geographic diversity provided by the complete dataset. For detailed experimental results at each distance level and analysis on inference overhead, please refer to the **Appendix I** and **Appendix J**.

4.8 Case Study

To qualitatively illustrate the routing behavior of GeoRouter, we analyze two representative examples shown in Figure 6. **Case 1** presents a coastal dining scene containing generic visual elements. The retrieval method matches a visually similar candidate from a distant location (9021.56 km away), as such scenes offer insufficient discriminative features for precise instance matching. In contrast, the generation-based method correctly infers the geographic context from semantic cues and produces an accurate prediction (1.35 km). GeoRouter assigns a positive routing score ($r = +3.39$) and correctly selects the generation paradigm. **Case 2** presents an older photograph of an arid airstrip, where limited image quality and the absence of salient semantic cues make direct coordinate generation unreliable (7184.75 km). However, the distinctive mountain silhouette in the background provides a strong visual fingerprint that enables the retrieval method to identify a geographically accurate candidate (2.42 km). GeoRouter assigns a negative routing score ($r = -3.96$) and correctly selects the retrieval paradigm.

5 Conclusion

In this paper, we propose GeoRouter, a dynamic routing framework that adaptively assigns each query to the most suitable paradigm for worldwide image geolocalization. To support this framework, we introduce Distance-Aware Preference Optimization (DisPO) and construct GeoRouting, the first large-scale dataset designed for geolocalization routing. Extensive experiments on IM2GPS3K and YFCC4K demonstrate that GeoRouter consistently outperforms state-of-the-art baselines.

Acknowledgements

This research was partially supported by National Natural Science Foundation of China (No.62502404), Hong Kong Research Grants Council (Research Impact Fund No.R1015-23, Collaborative Research Fund No.C1043-24GF, RGC Research Fellow Scheme No.RFS2627-1S03, General Research Fund No.11218325, No.11212926), Institute of Digital Medicine of City University of Hong Kong (No.9229503), and Didi (CCF-DiDi GAIA Collaborative Research Funds).

References

[1] Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Visual and object geo-localization: A comprehensive survey. *arXiv preprint arXiv:2112.15202*, 2021.

[2] Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models. *arXiv preprint arXiv:2508.01608*, 2025.

[3] Pengyue Jia, Song Gao, Sharon Li, and Xiangyu Zhao. Worldwide image geolocalization in the foundation model era: A survey. 2026.

[4] Pengyue Jia, Yingyi Zhang, Xiangyu Zhao, and Sharon Li. Geoarena: Evaluating open-world geographic reasoning in large vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20886–20901, 2026.

[5] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017.

[6] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European conference on computer vision*, pages 726–743. Springer, 2020.

[7] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 12105–12115, 2021.

[8] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5384, 2022.

[9] Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, and Bingyi Cao. Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11036–11046, 2023.

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

[11] Hao Li, Fabian Deuser, Wenping Yin, Xuanshu Luo, Paul Walther, Gengchen Mai, Wei Huang, and Martin Werner. Cross-view geolocalization and disaster mapping with street-view and vhr satellite imagery: A case study of hurricane ian. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220:841–854, 2025.

[12] VijayaKumar Kadha, Sambit Bakshi, and Santos Kumar Das. Unravelling digital forgeries: A systematic survey on image manipulation detection and localization. *ACM Computing Surveys*, 57(12):1–36, 2025.

[13] Xiangyu Zhao, Wenqi Fan, Hui Liu, and Jiliang Tang. Multi-type urban crime prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4388–4396, 2022.

[14] Xiangyu Zhao and Jiliang Tang. Exploring transfer learning for crime prediction. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 1158–1159. IEEE, 2017.

[15] Zhiqiang Wang, Dejia Xu, Rana Muhammad Shahroz Khan, Yanbin Lin, Zhiwen Fan, and Xingquan Zhu. Llmgeo: Benchmarking large language models on image geolocation in-the-wild. *arXiv preprint arXiv:2405.20363*, 2024.

[16] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013.

[17] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, et al. Openstreetview-5m: The many roads to global visual geolocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21967–21977, 2024.

[18] Neel Jay, Hieu Minh Nguyen, Trung Dung Hoang, and Jacob Haimes. Evaluating precise geolocation inference capabilities of vision language models. *arXiv preprint arXiv:2502.14412*, 2025.

[19] Nilay Pande, Sahiti Yerramilli, Jayant Sravan Tamarapalli, and Rynaa Grover. Geochain: Multimodal chain-of-thought for geographic reasoning. In *The First Workshop on Multimodal Knowledge and Language Modeling*.

[20] Mohit Talreja, Joshua Diao, Jim Thannikary James, Radu Casapu, Tejas Santanam, Ethan Mendes, Alan Ritter, Wei Xu, and James Hays. Georc: A benchmark for geolocation reasoning chains. *arXiv preprint arXiv:2601.21278*, 2026.

[21] Lukas Haas, Silas Alberti, and Michal Skreta. Learning generalized zero-shot learners for open-domain image geolocalization. *arXiv preprint arXiv:2302.00275*, 2023.

[22] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023.

[23] Pengyue Jia, Seongheon Park, Song Gao, Xiangyu Zhao, and Sharon Li. Georanker: Distance-aware ranking for worldwide image geolocalization. *arXiv preprint arXiv:2505.13731*, 2025.

[24] Jian Fang, Siyi Qian, and Shaohui Liu. Geomr: Integrating image geographic features and human reasoning knowledge for image geolocalization. *Knowledge-Based Systems*, page 115391, 2026.

[25] Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval*, pages 2749–2754, 2024.

[26] Pengyue Jia, Yiding Liu, Xiaopeng Li, Xiangyu Zhao, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, and Dawei Yin. G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. *Advances in Neural Information Processing Systems*, 37:53198–53221, 2024.

[27] Narges Ghasemi, Amir Ziashahabi, Salman Avestimehr, and Cyrus Shahabi. Geotoken: Hierarchical geolocalization of images via next token prediction. *arXiv preprint arXiv:2511.01082*, 2025.

[28] Biao Wu, Meng Fang, Ling Chen, Ke Xu, Tao Cheng, and Jun Wang. Vision-language reasoning for geolocalization: A reinforcement learning approach. *arXiv preprint arXiv:2601.00388*, 2026.

[29] Bo Yu, Fengze Yang, Yiming Liu, Chao Wang, Xuewen Luo, Taozhe Li, Ruimin Ke, Xiaofan Zhou, and Chenxi Liu. Locatability-guided adaptive reasoning for image geo-localization with vision-language models. *arXiv preprint arXiv:2603.13628*, 2026.

[30] Yikun Wang, Zuyan Liu, Ziyi Wang, Han Hu, Pengfei Liu, and Yongming Rao. Geovista: Web-augmented agentic visual reasoning for geolocalization. *arXiv preprint arXiv:2511.15705*, 2025.

[31] Yuxiang Ji, Yong Wang, Ziyu Ma, Yiming Hu, Hailang Huang, Xuecai Hu, Guanhua Chen, Liaoni Wu, and Xiangxiang Chu. Thinking with map: Reinforced parallel map-augmented agent for geolocalization. *arXiv preprint arXiv:2601.05432*, 2026.

[32] Yushuo Zheng, Huiyu Duan, Zicheng Zhang, Xiaohong Liu, and Xiongkuo Min. Learning to wander: Improving the global image geolocation ability of lmms via actionable reasoning, 2026.

[33] Ling Li, Yu Ye, Yao Zhou, Bingchuan Jiang, and Wei Zeng. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. *arXiv preprint arXiv:2406.18572*, 2024.

[34] Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. *arXiv preprint arXiv:2506.14674*, 2025.

[35] Chun Wang, Xiaojun Ye, Xiaoran Pan, Zihao Pan, Haofan Wang, and Yiren Song. Gre suite: Geo-localization inference via fine-tuned vision-language models and enhanced reasoning chains. *arXiv preprint arXiv:2505.18700*, 2025.

[36] Modi Jin, Yiming Zhang, Boyuan Sun, Dingwen Zhang, MingMing Cheng, and Qibin Hou. Geoagent: Learning to geolocate everywhere with reinforced geographic characteristics, 2026.

[37] Weimin Shi, Xiang Li, Kaige Li, Junhao Fang, Qiang Zhou, Qichuan Geng, and Zhong Zhou. Geobayes: Probabilistic image geo-localization inference via sequential bayesian updating. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8997–9005, 2026.

[38] Yiyang Su and Xiaoming Liu. Interpretable perception and reasoning for audiovisual geolocation. *arXiv preprint arXiv:2603.05708*, 2026.

[39] Jaeyoung Choi, Claudia Hauff, Olivier Van Laere, and Bart Thomee. The placing task at mediaeval 2016. In Guillaume Gravier, Claire-Hélène Demarty, Hervé Bredin, Bogdan Ionescu, Christina Boididou, Emmanuel Dellandréa, Jaeyoung Choi, Michael Riegler, Richard F. E. Sutcliffe, Igor Szöke, Gareth J. F. Jones, and Martha A. Larson, editors, *Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016*, volume 1739 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[40] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[41] Nicolas Dufour, Vicky Kalogeiton, David Picard, and Loic Landrieu. Around the world in 80 timesteps: A generative approach to global visual geolocation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23016–23026, 2025.

[42] Zhangyu Wang, Zeping Liu, Jielu Zhang, Zhongliang Zhou, Qian Cao, Nemin Wu, Lan Mu, Yang Song, Yiqun Xie, Ni Lao, et al. Locdiff: Identifying locations on earth by diffusing in the hilbert space. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

[44] Zhiyang Dou, Zipeng Wang, Xumeng Han, Guorong Li, Zhipei Huang, and Zhenjun Han. Gaga: Towards interactive global geolocation assistant. *arXiv preprint arXiv:2412.08907*, 2024.

[45] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[46] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[47] Anindya Sarkar, Srikumar Sastry, Aleksis Pirinen, Chongjie Zhang, Nathan Jacobs, and Yevgeniy Vorobeychik. Gomaa-geo: Goal modality agnostic active geo-localization. *Advances in Neural Information Processing Systems*, 37:104934–104964, 2024.

[48] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 2621–2630, 2017.

[49] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018(1):7068349, 2018.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[51] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Transactions on Spatial Algorithms and Systems*, 10(2):1–46, 2024.

[52] Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. Geoai: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, 2020.

[53] Jiahao Wen, Hang Yu, and Zhedong Zheng. Weatherprompt: Multi-modality representation learning for all-weather drone visual geo-localization. *Advances in Neural Information Processing Systems*, 38:32974–32994, 2026.

[54] Yuanshao Zhu, James Yu, Xiangyu Zhao, Xun Zhou, Liang Han, Xuetao Wei, and Yuxuan Liang. Unitraj: Learning a universal trajectory foundation model from billion-scale worldwide traces. *Advances in Neural Information Processing Systems*, 38:42258–42295, 2026.

[55] Jiawei Cheng, Jingyuan Wang, Yichuan Zhang, Jiahao Ji, Yuanshao Zhu, Zhibo Zhang, and Xiangyu Zhao. Poi-enhancer: An llm-based semantic enhancement framework for poi representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 11509–11517, 2025.

[56] Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Yu. Synmob: Creating high-fidelity synthetic gps trajectory dataset for urban mobility analysis. *Advances in Neural Information Processing Systems*, 36:22961–22977, 2023.

[57] Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi Wang, and Zitao Liu. Promptst: Prompt-enhanced spatio-temporal multi-attribute prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3195–3205, 2023.

[58] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European conference on computer vision*, pages 37–55. Springer, 2016.

[59] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018.

[60] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–579, 2018.

[61] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. In *European Conference on Computer Vision*, pages 196–215. Springer, 2022.

[62] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23182–23190, 2023.

[63] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[64] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019.

[65] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021.

[66] Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing*, 31:3780–3792, 2022.

[67] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022.

[68] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 3480–3488, 2023.

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

[70] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3616, 2017.

[71] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020.

[72] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19370–19380, 2023.

[73] Furong Jia, Ling Dai, Wenjin Deng, Fan Zhang, Chen Hu, Daxin Jiang, and Yu Liu. Spotagent: Grounding visual geo-localization in large vision-language models through agentic reasoning. *arXiv preprint arXiv:2602.09463*, 2026.

[74] Lars Kotthoff. Algorithm selection for combinatorial search problems: A survey. In *Data mining and constraint programming: Foundations of a cross-disciplinary approach*, pages 149–190. Springer, 2016.

[75] Pascal Kerschke, Holger H Hoos, Frank Neumann, and Heike Trautmann. Automated algorithm selection: Survey and perspectives. *Evolutionary computation*, 27(1):3–45, 2019.

[76] Clovis Varangot-Reille, Christophe Bouvard, Antoine Gourru, Mathieu Ciancone, Marion Schaeffer, and François Jacquenet. Doing more with less: A survey on routing strategies for resource optimisation in large language model-based systems. *arXiv preprint arXiv:2502.00409*, 2025.

[77] Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Exploring the benefits of training expert language models over instruction tuning. In *International Conference on Machine Learning*, pages 14702–14729. PMLR, 2023.

[78] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

[79] Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. Embedllm: Learning compact representations of large language models. *arXiv preprint arXiv:2410.02223*, 2024.

[80] Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks VS Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*, 2024.

[81] Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*, 2024.

[82] Dimitrios Sikeridis, Dennis Ramdass, and Pranay Pareek. Pickllm: Context-aware rl-assisted large language model routing. In *International Workshop on AI for Transportation*, pages 227–239. Springer, 2025.

[83] Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. Routing to the expert: Efficient reward-guided ensemble of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1964–1974, 2024.

[84] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pages 95–103, 2018.

[85] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1040–1048, 2018.

[86] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*, 2024.

[87] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems*, 37:126544–126565, 2024.

[88] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[89] Pengyue Jia, Yingyi Zhang, Xiangyu Zhao, and Sharon Li. Geoarena: An open platform for benchmarking large vision-language models on worldwide image geolocalization. *arXiv preprint arXiv:2509.04334*, 2025.

[90] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[91] Derong Xu, Xinhang Li, Ziheng Zhang, Zhenxi Lin, Zhihong Zhu, Zhi Zheng, Xian Wu, Xiangyu Zhao, Tong Xu, and Enhong Chen. Harnessing large language models for knowledge graph question answering via adaptive multi-aspect retrieval-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25570–25578, 2025.

Technical Appendix

Table of Contents

A GeoRouting Data Entries 17

B More Details on Experimental Settings 17

C Detailed Introduction to Baselines 17

D Qualitative Examples Across Distance Thresholds 19

E Complete Experimental Results on Ablation Study 19

F Complete Experimental Results on Hyperparameter Analysis 20

G Transferability Analysis 21

H Complete Experimental Results on Backbone Model Scale 21

I Data Efficiency Analysis with All Geographic Levels 22

J Inference Overhead Analysis 23

K Limitations 23

L Comparison with Non-Learned and Vision-Free Routing Baselines 23

M End-to-End Inference Cost 24

N Transferability across Retrieval Models 24

O Failure Mode Analysis 25

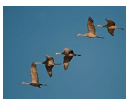
Query Image	Predictions	Candidates			Label
 (27.017, -108.938)	Retrieval-based Prediction (20.935, -101.430), Distance: 1018km Generation-based Prediction (20.676, -103.347), Distance: 903km	 (20.935, -101.430)	 (19.050, -98.209)	 (19.307, -98.241)	1
 (30.751, -98.454)	Retrieval-based Prediction (30.751, -98.454), Distance: 0km Generation-based Prediction (33.897, -96.579), Distance: 390km	 (30.751, -98.454)	 (30.703, -96.475)	 (33.016, -94.757)	0
 (41.090, -80.202)	Retrieval-based Prediction (40.742, -98.546), Distance: 1542km Generation-based Prediction (41.25, -96.00), Distance: 1324km	 (40.742, -98.546)	 (41.121, -80.196)	 (41.086, -80.159)	1
 (41.899, 12.472)	Retrieval-based Prediction (41.898, 12.473), Distance: 0.17km Generation-based Prediction (41.899, 12.473), Distance: 0.06km	 (41.898, 12.473)	 (41.899, 12.473)	 (41.897, 12.473)	1
 (36.196, -112.053)	Retrieval-based Prediction (36.217, -112.061), Distance: 2.3km Generation-based Prediction (36.198, -112.096), Distance: 3.9km	 (36.217, -112.061)	 (36.013, -111.866)	 (25.236, -100.433)	0

Figure 7: Representative examples of GeoRouting dataset entries. Each row shows a query image with its ground-truth GPS coordinates, predictions from the retrieval-based and generation-based paradigms with their respective distance errors, retrieved candidate images with their GPS coordinates, and the binary routing label (1 = generation-based paradigm is more accurate; 0 = retrieval-based paradigm is more accurate).

A GeoRouting Data Entries

Each entry in the GeoRouting dataset consists of a query image, its ground-truth GPS coordinates, retrieved candidate images with their coordinates, predictions from both the retrieval-based and generation-based paradigms, the distance error of each prediction, and a binary routing label. The label indicates which paradigm produces the more accurate prediction: label 1 denotes that the generation-based paradigm is more suitable, while label 0 denotes that the retrieval-based paradigm is more suitable. For display purposes, Figure 7 shows only three of the ten candidates per entry.

B More Details on Experimental Settings

Table 4 provides a comprehensive summary of the hardware and training configuration used in our experiments. All experiments are conducted on a node equipped with 8 NVIDIA H800 GPUs. Each training epoch takes approximately 50 minutes, and the full 3-epoch training run is completed in under 3 hours, demonstrating the computational efficiency of our approach.

C Detailed Introduction to Baselines

In this section, we provide introductions to all baseline methods compared in our experiments.

Table 4: Detailed experimental configurations.

Parameter	Value
Batch Size per Device	3
Batch Size	24
Dataset Size	100K
DeepSpeed Stage	2
GPU	NVIDIA H800 $\times$ 8
Training Time	50 min / epoch
Total Parameters	8,298,264,064
Trainable Parameters	6,884,864 (0.08\%)
Training GPU Memory	Total 485 GiB, Avg 60.5 GiB / Device
LVLM Backbone	Qwen2-VL-7B-Instruct

- **[L]kNN**, $\sigma = 4$ [48]. kNN retrieves the top-k most similar images from a reference database and aggregates their GPS coordinates to produce a final prediction. The parameter σ controls the spatial bandwidth of the aggregation kernel; as k decreases, the prediction becomes more focused, and at $k = 1$ the method reduces to nearest-neighbor retrieval.
- **PlaNet** [58]. PlaNet is the first method to cast worldwide image geolocalization as a classification problem. It partitions the Earth's surface into geographic cells and trains a convolutional neural network to assign each query image to its corresponding cell. The center coordinate of the predicted cell is used as the final location estimate.
- **CPlaNet** [59]. CPlaNet extends PlaNet by introducing a combinatorial partitioning strategy that derives fine-grained output classes through the intersection of multiple coarser geographic partitions. This hierarchical decomposition allows the model to generate more precise location predictions compared to using a single flat partitioning scheme.
- **ISNs** [60]. ISNs augment the standard image input with auxiliary scene features that capture high-level environmental context, such as whether a scene is indoor, natural, or urban. Incorporating these additional signals alongside the original image content enables the model to produce richer representations, leading to improved geolocalization accuracy across diverse environments.
- **Translocator** [61]. Translocator adopts a dual-branch transformer architecture that processes both the original query image and its corresponding semantic segmentation map in parallel. By jointly exploiting appearance and structural information, the model learns more discriminative location-aware features and achieves stronger performance.
- **GeoDecoder** [62]. GeoDecoder identifies that prior classification-based methods do not fully exploit the hierarchical nature of geographic information. It addresses this limitation by introducing a cross-attention mechanism that models relationships across features at different spatial granularities, improving the model's capacity to reason about location from complex scene content.
- **GeoCLIP** [22]. GeoCLIP adapts the CLIP contrastive learning framework for geolocalization by introducing a dedicated GPS encoder that projects geographic coordinates into a shared embedding space alongside image features. This alignment allows the model to directly associate visual content with geographic locations, enabling effective worldwide geolocalization without relying on a fixed classification grid.
- **Img2Loc** [25]. Img2Loc integrates a retrieval-augmented generation (RAG) [91] pipeline into geolocalization. Given a query image, the system first retrieves a set of visually similar reference images and collects their GPS coordinates. These coordinates are then incorporated into a prompt for a large vision-language model, which generates the final location prediction by reasoning over both the visual content and the retrieved geographic context.
- **PIGEON** [43]. PIGEON introduces a framework that combines semantically guided geographic cell partitioning, multi-task contrastive pretraining, and a specialized loss function. By grouping candidate locations according to semantic similarity and subsequently applying targeted retrieval to refine predictions, PIGEON achieves notably improved geolocalization accuracy over prior classification and retrieval methods.
- **G3** [26]. G3 is a three-stage RAG-based framework composed of Geo-alignment, Geo-diversification, and Geo-verification. The Geo-alignment stage jointly trains image, GPS, and text encoders to learn location-aware multimodal representations for retrieval. The subsequent stages

Table 5: Ablation study on IM2GPS3K and YFCC4K.

Methods	IM2GPS3K						YFCC4K					
	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km	Average	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km	Average
w/o DisPO	**20.85**	49.38	64.26	79.05	89.76	60.66	33.05	45.75	57.08	71.65	82.80	58.06
w/o $\mathcal{C}$	20.69	49.18	64.60	79.75	90.09	60.86	**33.58**	45.55	56.42	71.94	**83.53**	58.20
w/o $\mathcal{C}$ \& g^{ret}	19.22	46.11	61.76	77.11	89.29	58.69	29.83	41.34	52.05	68.67	81.50	54.67
w/o g^{gen}	20.22	48.31	64.06	78.75	89.92	60.25	33.20	45.08	55.62	71.25	82.94	57.61
w/o Context	19.29	45.85	61.46	77.41	89.16	58.63	30.14	41.60	52.98	69.91	82.34	55.39
Ours	20.82	**50.48**	**65.73**	**80.35**	**90.66**	**61.60**	32.98	**46.01**	**57.52**	**72.02**	83.02	**58.31**

leverage an LVLM to generate and verify a diverse set of candidate GPS coordinates, with the final prediction selected by comparing candidate similarity against the query using the learned representations.

- **GeoToken** [27]. GeoToken reformulates geolocalization as a coarse-to-fine autoregressive token prediction task, inspired by the way humans progressively narrow down a location from a broad region to a specific address. The model uses Google S2 cells as a hierarchical spatial grid and autoregressively predicts location tokens at increasing levels of geographic resolution, conditioned on both visual input and previously predicted tokens.
- **GeoRanker** [23]. GeoRanker addresses a key limitation of existing two-stage retrieval pipelines, where candidate selection relies on simplistic point-wise similarity heuristics that fail to capture spatial relationships among candidates. GeoRanker employs an LVLM to jointly encode query–candidate pairs and predict geographic proximity. A multi-order distance loss is introduced to supervise both absolute and relative distance rankings, enabling the model to reason about the structured spatial arrangement of candidates and select the most geographically accurate prediction.
- **GeoBayes** [37]. GeoBayes is a training-free framework that formulates geo-localization as a Maximum a Posteriori (MAP) estimation problem. A state memory mechanism propagates hypotheses, evidence, and inference context across geographic hierarchy levels, enabling coarse-to-fine geolocalization.
- **Geo-R** [28]. Geo-R is a retrieval-free framework that introduces the Chain of Region paradigm, which decomposes geolocalization into a hierarchical reasoning process. The model is further optimized via reinforcement learning using GRPO with a composite reward combining Haversine distance-based spatial accuracy and output format consistency, enabling spatially grounded refinement of coordinate predictions.

D Qualitative Examples Across Distance Thresholds

Figure 8 shows representative query images grouped by their geolocalization error across five distance thresholds. Images that are localized within 1km typically contain strong and unambiguous visual identifiers, such as well-known architectural structures or iconic landmarks. These scenes benefit both paradigms: the retrieval method can readily find visually similar database entries for such frequently photographed locations, while the generation method can leverage the world knowledge embedded in LVLMs to reason about them directly. As the error threshold increases to 25km and 200km, the query images tend to feature more generic scenes, such as local churches, rural buildings, or regional landscapes, which provide weaker discriminative cues. At these intermediate levels, accurate geolocalization depends more heavily on subtle contextual signals such as architectural style, vegetation type, or topographic features, making paradigm selection by GeoRouter particularly valuable. At the coarsest thresholds (750km and 2500km), the query images frequently depict scenes with little to no location-specific information, such as open fields, featureless seascapes, or generic natural environments. In these cases, neither the retrieval nor the generation paradigm can reliably infer the true location, and the large errors reflect the fundamental ambiguity inherent in such visually uninformative inputs.

E Complete Experimental Results on Ablation Study

Table 5 presents the complete ablation results across both IM2GPS3K and YFCC4K datasets. The trends observed on YFCC4K are consistent with those reported on IM2GPS3K in the main paper,

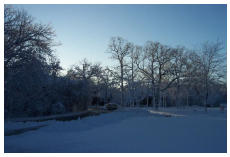
Error $\leq$	Images		
1KM			
25KM			
200KM			
750KM			
2500KM			

Figure 8: Qualitative examples of GeoRouter predictions across different geolocalization accuracy levels

confirming that each proposed component delivers a positive contribution regardless of the dataset. Regarding the composition of the input prompt, the results highlight a clear hierarchy among the contextual components. The two paradigm predictions, g^{gen} and g^{ret}, are the most critical inputs. This indicates that explicitly providing the predicted coordinates from both paradigms enables direct comparison and is fundamental to accurate routing decisions. By contrast, removing only the retrieved candidate set (**w/o** $\mathcal{C}$) results in a comparatively minor performance degradation, suggesting that while the candidate context offers supplementary spatial cues that help the model assess retrieval reliability, the paradigm predictions themselves remain the primary basis for routing.

F Complete Experimental Results on Hyperparameter Analysis

Figure 9 presents the complete hyperparameter sensitivity results for α on both IM2GPS3K and YFCC4K. On both datasets, setting α to an excessively large value (e.g., 2.5 or 3.0) leads to a consistent decline in accuracy across most distance thresholds. This behavior is particularly evident at the city and region levels (25km and 200km), where calibrated soft labels are most critical for distinguishing between the two paradigms. As discussed in the main text, an overly large α collapses

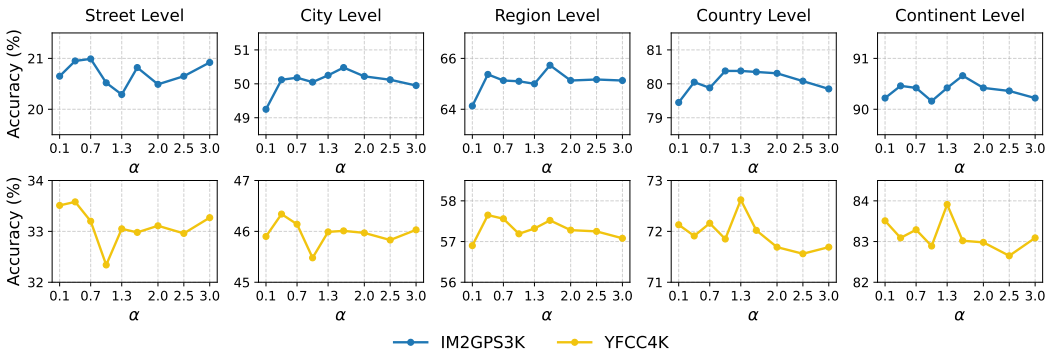


Figure 9: Effect of the hyperparameter α on geolocalization accuracy across all five distance thresholds on both IM2GPS3K (blue) and YFCC4K (yellow). Each column corresponds to a distance threshold from Street level (1km) to Continent level (2500km).

Table 6: Transferability of the GeoRouter across different generative LVLMs on the IM2GPS3K dataset. All metrics are reported as geolocalization accuracy (%). The Δ Mean column indicates the performance gain achieved by integrating our routing mechanism.

Base Generative Model	Method	Distance Thresholds						Δ Mean
		1km	25km	200km	750km	2500km	Mean	
Gemini 2.5 Flash	Base Model	17.55	47.91	63.30	78.85	88.59	59.24	+4.00%
	+ GeoRouter	**20.82**	**50.48**	**65.73**	**80.35**	**90.66**	**61.61**	
Gemini 2.5 Flash (Thinking)	Base Model	18.59	48.88	62.96	78.61	88.99	59.61	+3.37%
	+ GeoRouter	**21.15**	**50.98**	**64.83**	**80.35**	**90.79**	**61.62**	
Gemini 2.0 Flash	Base Model	15.42	45.38	61.23	78.14	87.92	57.62	+4.84%
	+ GeoRouter	**19.45**	**48.45**	**63.73**	**79.88**	**90.56**	**60.41**	
GPT-5 Mini	Base Model	17.95	44.84	58.66	74.21	86.49	56.43	+4.64%
	+ GeoRouter	**20.95**	**47.78**	**61.53**	**76.58**	**88.39**	**59.05**	

the continuous soft labels into near-binary supervision, undermining the distance-aware modeling that DisPO is designed to provide.

G Transferability Analysis

To evaluate whether the learned routing policy generalizes to different generative models, we conduct a transferability analysis. Specifically, we train GeoRouter using the predictions generated by Gemini 2.5 Flash. During inference, we directly apply this frozen routing model to evaluate predictions from various unseen LVLMs without any additional fine-tuning. The results on the IM2GPS3K dataset, presented in Table 6, yield several key conclusions. First, GeoRouter consistently improves the geolocalization accuracy of every evaluated LVLM across all distance thresholds. This demonstrates that even when the generation-based method is replaced during inference, the routing mechanism stably enhances the overall system performance. Second, the successful transfer across different model families and parameter scales indicates that GeoRouter learns a generalized routing capability rather than overfitting to the specific error distribution of the training LVLM.

H Complete Experimental Results on Backbone Model Scale

Table 7 reports the detailed geolocalization accuracy for each backbone model configuration on both IM2GPS3K and YFCC4K, complementing the mean accuracy summary presented in the main text. The results are consistent with the trends described in the main text. Within the Qwen2-VL family, scaling from 2B to 7B parameters yields improvements across all distance thresholds on both datasets. Within the Qwen3-VL family, scaling from 2B to 8B produces steady improvements, and the gains are more evenly distributed across thresholds compared to Qwen2-VL family, suggesting that the architectural improvements in Qwen3-VL lead to more balanced routing capability across

Table 7: Performance scaling across different sizes of the routing backbone model. The evaluation is conducted using Qwen2-VL and Qwen3-VL families on the IM2GPS3K and YFCC4K datasets. All results are presented as geolocalization accuracy (%).

Backbone Model	IM2GPS3K						YFCC4K					
	1km	25km	200km	750km	2500km	Mean	1km	25km	200km	750km	2500km	Mean
Qwen2-VL-2B	20.35	49.02	64.36	79.35	90.06	60.63	33.49	45.50	56.97	71.38	83.00	58.07
Qwen2-VL-7B	20.82	50.48	65.73	80.35	90.66	61.61	32.98	46.01	57.52	72.02	83.02	58.31
Qwen3-VL-2B	20.95	49.68	64.80	79.91	90.72	61.21	33.55	45.90	57.30	71.89	83.40	58.41
Qwen3-VL-4B	21.02	49.98	65.37	79.98	90.72	61.41	33.82	46.45	57.76	72.53	83.44	58.80
Qwen3-VL-8B	20.99	50.18	65.33	80.28	90.66	61.49	33.71	46.45	57.80	73.06	84.37	59.08

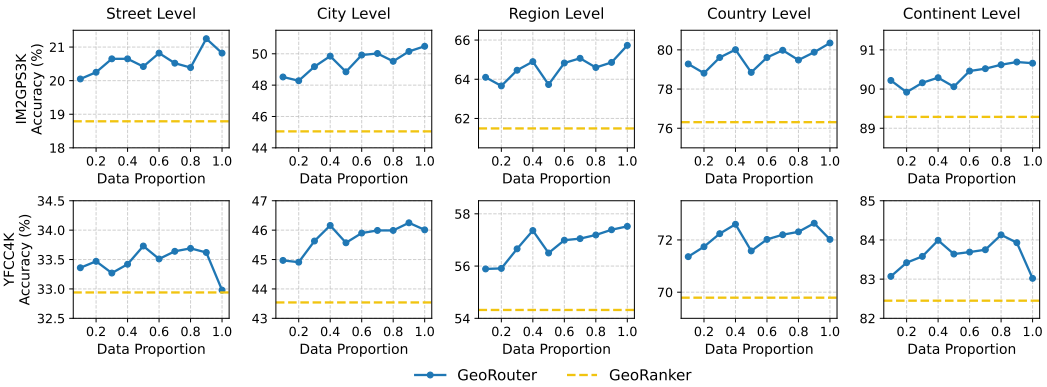


Figure 10: Detailed data efficiency analysis across five geographic levels on the IM2GPS3K (top row) and YFCC4K (bottom row) datasets. The solid blue line represents the accuracy of GeoRouter trained on varying proportions of the dataset, while the dashed yellow line indicates the performance of the state-of-the-art baseline GeoRanker.

different spatial granularities. Notably, even the smallest configuration, Qwen2-VL-2B, consistently outperforms the strongest single-paradigm baseline GeoRanker, which further confirms that the routing framework is effective regardless of backbone capacity.

I Data Efficiency Analysis with All Geographic Levels

In the main text, we present the mean geolocalization accuracy of GeoRouter under varying volumes of training data. Figure 10 provides a comprehensive breakdown of these data efficiency results across all five geographic levels for both the IM2GPS3K and YFCC4K datasets. The detailed evaluation yields two consistent observations that align with the aggregated results. First, the advantage of GeoRouter over the state-of-the-art baseline, GeoRanker, is robust across all spatial scales. Even when trained on only 10% of the available dataset, GeoRouter outperforms GeoRanker at every individual distance threshold on both datasets. Second, as the proportion of training data increases, the geolocalization accuracy generally improves across the various spatial levels. While minor fluctuations occur at specific granularities (for example, the Street level on the YFCC4K dataset), the overall positive trend confirms that the dynamic routing mechanism stably benefits from increased geographic diversity in the training data.

Table 8: Inference overhead of GeoRouter on the IM2GPS3K and YFCC4K datasets.

GPU	Batch Size	IM2GPS3K		YFCC4K	
		GPU Memory Consumption	Inference Speed	GPU Memory Consumption	Inference Speed
NVIDIA H800 × 1	16	Avg 65 GiB	0.1054s per figure	Avg 65 GiB	0.0551s per figure

Table 9: Comparison with non-learned routing heuristics and vision-free classifiers on the YFCC4K dataset. All metrics are reported as geolocalization accuracy (%). The best results are highlighted in **bold**.

Type	Method	1km	25km	200km	750km	2500km	Mean
Single Paradigm	Pure Retrieval	32.47	42.70	53.13	69.95	82.67	56.18
	Pure Generation	10.43	26.85	41.93	59.30	73.28	42.36
Non-Learned	Retrieval confidence	31.64	43.28	54.85	68.78	79.03	55.52
	Neighbor distance	23.99	36.35	48.96	65.61	77.60	50.50
	Candidate-cluster median distance	23.81	35.60	47.16	63.76	76.61	49.39
	Unconditional spherical midpoint	9.08	22.20	39.68	58.66	75.15	40.96
Vision-Free	Logistic regression	27.18	38.49	50.44	68.03	81.08	53.05
	XGBoost	24.45	37.83	51.19	68.61	81.22	52.66
Learned LVLM	GeoRouter	**32.98**	**46.01**	**57.52**	**72.02**	**83.02**	**58.31**

J Inference Overhead Analysis

In this section, we evaluate the computational overhead introduced by the routing component during the inference stage. Table 8 details the GPU memory consumption and inference speed of GeoRouter. The evaluation is conducted on a single NVIDIA H800 GPU using a batch size of 16.

K Limitations

While GeoRouter demonstrates strong performance, several directions remain worth further exploration. First, the current hard-routing scheme may be suboptimal in borderline cases where both paradigms perform similarly; a soft ensemble extension that weighs predictions from both paradigms proportionally to their routing confidence is a natural next step. Second, the current binary routing formulation does not directly generalize to settings with more than two candidate paradigms, though the DisPO objective can be naturally extended via pairwise comparisons to accommodate a broader model selection space; as new paradigms emerge, the existing dataset construction pipeline can be readily applied to collect multi-paradigm performance comparisons and train the corresponding router with minimal additional effort.

L Comparison with Non-Learned and Vision-Free Routing Baselines

Since GeoRouter selects between two existing paradigms, a natural question is whether a simpler rule could make the same selection without a learned LVLM router. To answer this question, we compare GeoRouter with two groups of lightweight alternatives on YFCC4K.

The first group contains four non-learned heuristics. Retrieval confidence routes by the similarity score of the top-1 retrieved candidate. Neighbor distance routes by the geodesic distance between the GPS coordinates of the top-1 and top-2 retrieved candidates. Candidate-cluster median distance computes all pairwise geodesic distances among the top-10 candidate coordinates and routes by their median. Unconditional spherical midpoint does not route, and instead outputs the spherical midpoint of the two paradigm predictions. The second group contains two vision-free classifiers, logistic regression and XGBoost, which are trained on the same training split as GeoRouter. Both classifiers take three scalar features as input, namely the distance between the generation and retrieval predictions, the distance between the top-1 and top-2 retrieved candidates, and the median pairwise distance among the top-k retrieved candidates.

Table 9 reports the results, from which we draw two observations. First, none of the non-learned heuristics outperforms Pure Retrieval, the stronger of the two paradigms on YFCC4K. Retrieval confidence is the strongest heuristic, which indicates that the retrieval score carries a meaningful signal, but this signal alone is not enough to select the better paradigm for each query. Second, both vision-free classifiers fall between the two paradigms and below GeoRouter. The scalar features do carry real signal, but they only summarize the internal geometry of the retrieval pipeline and are blind to the content of the query image. Since the query in image geolocalization is an image, discarding it

Table 10: End-to-end inference latency and mean geolocalization accuracy on the IM2GPS3K and YFCC4K datasets. The latency of GeoRouter is computed as the maximum latency of the two parallel branches plus the routing latency.

Dataset	System	Mean Acc. (%)	Single-Sample Latency	Batch-16 Latency
IM2GPS3K	Retrieval-only	56.90	1.573s (FAISS 0.048 + Ranking 1.525)	27.087s (FAISS 0.355 + Ranking 26.733)
	Generation-only	59.24	1.600s	25.600s
	GeoRouter	**61.61**	1.763s (max(1.573, 1.600) + 0.163)	29.642s (max(27.087, 25.600) + 2.555)
YFCC4K	Retrieval-only	56.18	0.767s (FAISS 0.040 + Ranking 0.727)	12.587s (FAISS 0.228 + Ranking 12.359)
	Generation-only	42.36	1.600s	25.600s
	GeoRouter	**58.31**	1.689s (max(0.767, 1.600) + 0.089)	26.888s (max(12.587, 25.600) + 1.288)

removes the evidence needed to judge which paradigm is more reliable, and GeoRouter, which keeps the image in the loop, achieves the best result at every threshold.

M End-to-End Inference Cost

GeoRouter requires the predictions of both paradigms before making a routing decision. We therefore measure the end-to-end latency of every component and report it together with the mean geolocalization accuracy in Table 10. All measurements are conducted on a single NVIDIA H800 GPU, and the generation-based paradigm is accessed through the Gemini 2.5 Flash API. The reported latency includes both data preparation and model inference. Since the retrieval and generation branches are independent of each other, they can run in parallel, so the end-to-end latency of GeoRouter is the latency of the slower branch plus that of the routing pass.

From Table 10, we draw three observations. First, the routing overhead is small. The router performs a single forward pass and reads out a scalar from a linear head, so it adds only 0.163s and 0.089s per sample on IM2GPS3K and YFCC4K, respectively. Second, this modest overhead brings a substantial accuracy gain. In exchange for an approximately 5%–10% latency increase, GeoRouter improves the mean accuracy over the stronger single paradigm by 2.37 points on IM2GPS3K (61.61 vs. 59.24) and 2.13 points on YFCC4K (58.31 vs. 56.18). The stronger paradigm also flips between the two datasets, so a practitioner who picks the single paradigm that wins on one benchmark incurs a large loss on the other, whereas GeoRouter is the best system on both without any per-dataset tuning. Third, the dominant cost is inherited rather than introduced. Over 90% of the end-to-end latency comes from the two base paradigms, both of which are standard components already required by prior retrieval-augmented geolocalization systems, such as Img2Loc [25] and G3 [26], which likewise run retrieval before invoking an LVLM.

N Transferability across Retrieval Models

To examine whether the observed complementarity and the learned routing policy are specific to GeoRanker, we replace GeoRanker with GeoCLIP [22] as the retrieval branch and evaluate on YFCC4K. We consider two settings, and the generation branch remains Gemini 2.5 Flash in both. In the frozen transfer setting, the router trained on GeoRanker is applied without any retraining, and only the retrieval prediction and the retrieved candidates are replaced with those of GeoCLIP. In the retrained setting, we rebuild the GeoRouting training data from the retrieval results of GeoCLIP and retrain the router. Table 11 leads to two observations. First, the complementarity is not specific to GeoRanker. Even with a substantially weaker retrieval branch, routing still exceeds both components in mean accuracy. Second, the frozen router transfers without any retraining, which indicates that it has learned a paradigm-level decision rather than a policy tied to the specific error distribution of GeoRanker. Retraining on GeoCLIP-derived data aligns the training and inference distributions and further improves the mean accuracy from 43.71 to 45.63. Together with Appendix G, these results indicate that the gain comes from the routing mechanism rather than from a specific pair of underlying models.

Table 11: Transferability of GeoRouter when the retrieval branch is replaced with GeoCLIP on the YFCC4K dataset. All metrics are reported as geolocalization accuracy (%). The best results are highlighted in **bold**.

Method	1km	25km	200km	750km	2500km	Mean
GeoCLIP-only	9.59	19.31	32.63	55.00	74.69	38.24
Generation-only	10.43	26.85	41.93	59.30	73.28	42.36
GeoRouter (Frozen transfer)	12.35	26.06	41.12	61.11	77.93	43.71
GeoRouter (Retrained)	**12.81**	**28.59**	**44.05**	**63.45**	**79.28**	**45.63**

Table 12: Composition of routing errors on the IM2GPS3K and YFCC4K datasets. Each entry reports the number of routing errors in each group and its proportion among all routing errors.

Dataset	Both $<$ 1km	Both $>$ 2500km	Middle: Gen. Better	Middle: Ret. Better
IM2GPS3K (1,174 errors)	147 (12.52%)	85 (7.24%)	275 (23.42%)	667 (56.81%)
YFCC4K (1,398 errors)	90 (6.44%)	223 (15.95%)	327 (23.39%)	758 (54.22%)

O Failure Mode Analysis

To understand when GeoRouter makes wrong decisions, we collect every query on which GeoRouter selects the paradigm with the larger geodesic error. We partition these routing errors by the joint error profile of the two paradigms into four groups, namely both accurate (both errors below 1km), both poor (both errors above 2500km), middle region with generation better, and middle region with retrieval better, where the middle region covers all remaining routing errors.

Table 12 leads to two observations. First, 19.76% (IM2GPS3K) and 22.39% (YFCC4K) of the routing errors are benign. When both paradigms land within 1km or both exceed 2500km, the routing decision is wrong by label, but the resulting geolocalization outcome is essentially unaffected. Second, the consequential errors are clearly asymmetric. Roughly 70% of them on both datasets are cases where retrieval is better but generation is selected. The dominant failure mode of GeoRouter is therefore over-selecting the generation-based paradigm, namely trusting a fluent but hallucinated coordinate over a correct retrieval match.

We further examine a representative example from YFCC4K (`10888967684.jpg`) with the ground truth $(35.66, 139.76)$. The top-1 retrieval prediction $(35.65, 139.70)$ is accurate to 6.14km, whereas the generation-based paradigm hallucinates $(34.05, -118.24)$, an error of 8830.38km. However, the remaining retrieved candidates, $(37.78, -122.41)$, $(38.72, -121.16)$, $(52.18, -1.70)$, and $(53.37, -2.58)$, are scattered across North America and the UK, and none of them supports the top-1 hypothesis. As shown in Table 13, the router reads this dispersion as evidence that retrieval is unreliable, assigns a positive routing score of $+0.55$, and selects generation. Removing only the candidate list from the same prompt flips the score to -0.20 and recovers the correct decision.

Table 13: Routing decisions on the YFCC4K example `10888967684.jpg` with and without the retrieved candidates in the prompt.

Input Setting	Routing Score	Decision	Resulting Error
Full prompt (with candidates)	$+0.55$	Generation	8830.38km
Same prompt, candidates removed	-0.20	Retrieval	6.14km

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state all contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the limitations section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the model architecture, training details and hyperparameters, dataset construction procedure, and an anonymized public repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper states in the abstract that "our dataset, code, and checkpoints are publicly available," with an anonymized repository link provided. The GeoRouting dataset construction is described in detail in Section 3.1 and Appendix A.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to the implementation details section and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please refer to the experiments section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to the Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: this work does not involve data collection that would raise ethical concerns and is submitted anonymously in compliance with NeurIPS guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the introduction section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites all datasets and models used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Please refer to GeoRouting Dataset paragraph and Appendix A.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.