

# Technical Appendix

525	<b>Table of Contents</b>	
526	<b>A Prompts</b>	<b>15</b>
527	<b>B GeoRanking Data Entries</b>	<b>15</b>
528	<b>C More Information on Training and Inference</b>	<b>16</b>
529	<b>D Baseline Method Details</b>	<b>16</b>
530	<b>E Query Images with Different Error Thresholds</b>	<b>17</b>
531	<b>F Complete experimental results on ablation study</b>	<b>18</b>
532	<b>G Hyperparameter Analysis with All Geographic Levels</b>	<b>18</b>
533	<b>H Complete Experimental Results on Backbone Model Scale</b>	<b>19</b>
534	<b>I Limitations</b>	<b>19</b>





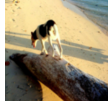





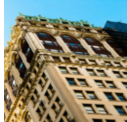









Query Image	Candidates			
 (48.306, 11.907)	 (48.311, 11.918) St. Paul, Landkreis Erding, Germany	 (51.941, 13.897) Lübben (Spreewald), Dahme- Spreewald, Germany	 (48.306, 11.907) St. Paul, Landkreis Erding, Germany	
 (9.751, 118.725)	 (12.444, -83.432) Nicaragua	 (8.740, 76.712) Varkala, Kerala, India	 (10.589, 124.313) San Francisco, Cebu, Philippines	
 (37.782, -122.409)	 (40.763, -73.973) New York, New York, United States	 (40.759, -73.984) New York, New York, United States	 (37.789, -122.401) San Francisco, California, United States	
 (52.597, -117.804)	 (44.574, -110.518) Wyoming, United States	 (63.576, 14.589) Jämtland County, Sweden	 (52.396, -116.076) Clearwater County, Alberta, Canada	
 (45.929, 8.662)	 (46.009, 9.282) Lecco, Lombardy, Italy	 (46.011, 9.282) Lecco, Lombardy, Italy	 (45.995, 9.263) Como, Lombardy, Italy	

Figure 8: Examples of GeoRanking Data Entries.

## A Prompts

**Prompting for generating candidates  $\mathcal{C}_g$ .** Following previous work [14], we use the following prompt template for generating candidates:

{query image} Suppose you are an expert in geolocalization. You have the ability to give two number GPS coordinates given an image. Please give me the location of the given image. Your answer should be in the following JSON format without any other information: {"latitude": float, "longitude": float}.

## B GeoRanking Data Entries

Figure 8 illustrates example entries from the GeoRanking dataset. Specifically, each query image is associated with 20 candidates, and each candidate contains GPS coordinates, textual descriptions, and image data. In total, GeoRanking includes 100K samples and 2 million query–candidate pairs. To the best of our knowledge, GeoRanking is the first dataset specifically designed for modeling distance-aware ranking between geographic entities. We release the dataset publicly and hope it will foster future research in areas such as GeoAI, information retrieval, and vision-language modeling.

Table 4: More Details on Training and Inference.

Parameter	Setting
GPU	NVIDIA L40S * 4
Training Time	16 hours / epoch
Total params	8,298,256,896
Trainable params	6,881,280 (0.083%)
Dataset Samples	100K
Batch Size	4
Batch Size per Device	1
Training GPU Memory Consumption	30 GB / GPU
VLM Backbone	Huggingface Qwen2-VL-7b-Instruct
Deepspeed	Stage 2

## C More Information on Training and Inference

In this section, we provide additional details regarding the training and inference setup. Table 4 summarizes the key hyperparameters used during these phases. Most experiments were conducted on four NVIDIA L40S GPUs. We also performed tests on two NVIDIA H200 GPUs, where training took approximately 7.5 hours per epoch with a batch size of 4, consuming around 90 GB of GPU memory per device with the gradient checkpointing off.

## D Baseline Method Details

In this section, we will give introductions to the baselines:

- **[L]kNN**,  $\sigma = 4$  [1]. kNN first retrieves the top- $k$  nearest neighbor images and aggregates their coordinates to form the final prediction. As the  $k$  decreases, the aggregation process becomes more focused. When  $k$  equals 1, the method turns to the NN.
- **PlaNet** [24]. PlaNet is the first work to formulate the worldwide geolocalization task as a classification problem. It partitions the Earth’s surface into a large number of geographical cells and trains a convolutional neural network to predict the correct cell for each image. Unlike previous approaches that primarily rely on landmark recognition or approximate matching with global image descriptors, PlaNet effectively integrates multiple visible cues within the image to enhance localization accuracy.
- **CPlaNet** [15]. CPlaNet follows PlaNet and propose combinatorial partitioning, which generates fine-grained output class by intersecting larger partitions.
- **ISNs** [55]. ISNs enhance the input image information by extracting additional scene context features, such as indoor, natural, or urban environments, alongside the original image content. By incorporating these richer contextual cues, ISNs achieves improved localization performance.
- **Translocator** [25]. Translocator designs a dual-branch transformer framework that simultaneously ingests the original image and its semantic segmentation map. This architecture enables the extraction of fine-grained spatial cues and the construction of more robust feature representations for geolocalization.
- **GeoDecoder** [26]. GeoDecoder identifies that earlier methods insufficiently leverage hierarchical spatial information. It addresses this by proposing a cross-attention mechanism that explicitly captures relationships across heterogeneous features, enhancing the model’s ability to interpret complex location-dependent features.
- **GeoCLIP** [8]. GeoCLIP extends the CLIP architecture by introducing a GPS encoder, aligning geographic coordinates with image and GPS embeddings. This enhancement enables more effective modeling of worldwide geolocalization tasks by incorporating spatial information directly into the learned feature space.
- **Img2Loc** [10]. Img2Loc advances geolocalization by integrating a RAG pipeline. It first retrieves visually similar candidates, then formulates a prompt incorporating these candidates’ coordinates, guiding a vision-language model to generate a final prediction.














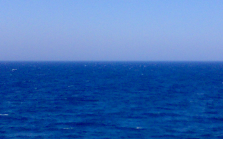

Error $\leq$	Images		
1KM			
25KM			
200KM			
750KM			
2500KM			

Figure 9: Example query images fall in different error thresholds.

- **PIGEON** [9]. PIGEON introduces an innovative framework that combines semantic geocell partitioning, multi-task contrastive pretraining, and a novel loss function. By clustering candidate locations semantically and refining predictions through targeted retrieval, PIGEON significantly boosts localization accuracy.
- **G3** [14]. G3 proposes a three-stage framework comprising Geo-alignment, Geo-diversification, and Geo-verification. Geo-alignment aligns GPS coordinates, textual descriptions, and visual data into a unified multi-modal representation to strengthen retrieval capabilities. Subsequently, Geo-diversification and Geo-verification are integrated within a RAG framework to robustly generate and select candidate geolocations.

## E Query Images with Different Error Thresholds

Figure 9 presents example query images under different error thresholds (1km, 25km, 200km, 750km, and 2500km). We observe that images with errors within 1km often contain distinctive location cues, such as landmark buildings, which facilitate accurate geolocalization. This is partly because retrieval candidates are more likely to retrieve visually similar images from the database due to the popularity of such locations. Additionally, generated candidates tend to produce more reliable predictions in these cases, as the locations are well-represented in the world knowledge embedded

Table 5: Complete ablation study on IM2GPS3K and YFCC4K.

Methods	IM2GPS3K					YFCC4K				
	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
w/o $\mathcal{L}_{PL}^{(2)}$	18.48	44.61	60.96	75.61	88.28	31.97	43.12	53.53	69.03	81.19
w/o $C_{neg}$	17.35	44.51	60.82	76.37	88.28	31.57	43.06	53.62	69.09	81.67
w/o $c_{img}^{text}$	18.02	43.91	60.19	<b>76.61</b>	88.62	31.70	43.06	<u>54.03</u>	69.42	82.07
w/o $c_{img}^{img}$	15.58	41.77	59.15	75.40	88.35	15.81	27.86	41.31	61.39	77.66
w/o $C_g$	18.21	43.47	59.69	75.47	<u>88.75</u>	32.60	43.03	53.43	<u>69.77</u>	<b>82.71</b>
Ours	<b>18.79</b>	<b>45.05</b>	<b>61.49</b>	76.31	<b>89.29</b>	<b>32.94</b>	<b>43.54</b>	<b>54.32</b>	<b>69.79</b>	<u>82.45</u>

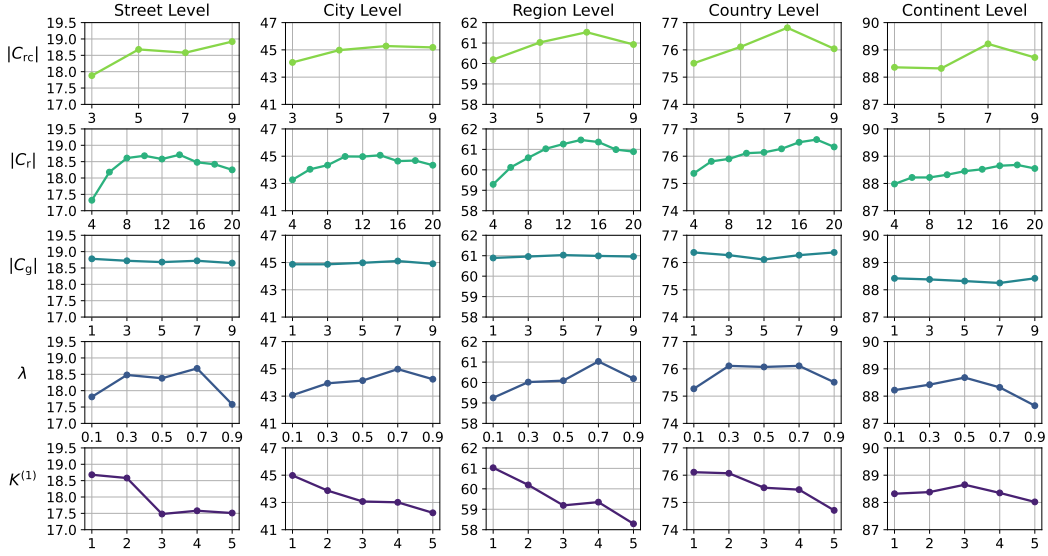


Figure 10: Hyperparameter analysis with all geographic levels on IM2GPS3K.

599 in large vision-language models. In contrast, query images with large geolocation errors (e.g.,  
600 2500km) typically lack informative visual cues—such as images depicting open oceans or vast  
601 grasslands—making it extremely challenging to infer their true locations. In such cases, neither  
602 retrieval nor generation is likely to yield useful candidates.

## 603 F Complete experimental results on ablation study

604 Table 5 presents the complete ablation results on both IM2GPS3K and YFCC4K. Consistent with the  
605 findings discussed in the main text, we observe that each component in our framework contributes  
606 positively to overall performance. Moreover, different types of contextual information incorporated  
607 into the prompt—such as visual cues, textual descriptions, and negative examples—all help improve  
608 both model training and inference. Finally, generated candidates are shown to complement retrieval-  
609 based candidates effectively. This is particularly beneficial for rare or long-tail query images, where  
610 retrieval candidates alone may fail to provide sufficient clues for accurate geolocation.

## 611 G Hyperparameter Analysis with All Geographic Levels

612 Figure 10 shows the impact of different hyperparameters on GeoRanker across all geographic levels.  
613 As observed, the trends of each hyperparameter remain largely consistent across levels, highlighting  
614 the stability and robustness of our model under varying localization granularities.

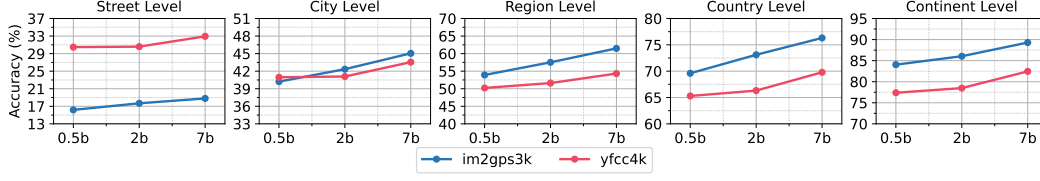


Figure 11: Impact of Backbone Scale across All Levels.

## H Complete Experimental Results on Backbone Model Scale

Figure 11 shows the effect of backbone model size across all geographic levels. Consistent performance improvements are observed on both IM2GPS3K and YFCC4K datasets as the backbone scales from 0.5B to 7B parameters, further confirming GeoRanker’s scalability and compatibility with stronger LVLM.

## I Limitations

Our method achieves notable improvements in geolocalization accuracy over existing baselines. In addition, it demonstrates superior time efficiency compared to LVLM prompting methods, and its data efficiency allows strong performance even with relatively limited supervision. However, compared to direct embedding-based retrieval approaches, GeoRanker introduces an additional ranking stage, which leads to increased computational overhead during inference. One solution is to analyze the retrieval results: if the top-k candidates are already geographically concentrated, the ranking step can be skipped without significant loss in accuracy, thereby reducing the overall inference time. In addition, GeoRanker supports parallel scoring of candidates during large-scale deployment, which can significantly improve runtime and computational efficiency.