GeoRanker: Distance-Aware Ranking for Worldwide Image Geolocalization

Pengyue Jia[1,2], **Seongheon Park**[2], **Song Gao**[3], **Xiangyu Zhao**[1], **Sharon Li**[2],
[1]Department of Data Science, City University of Hong Kong,
[2]Department of Computer Sciences, University of Wisconsin-Madison
[3]Department of Geography, University of Wisconsin-Madison
jia.pengyue@my.cityu.edu.hk, sharonli@cs.wisc.edu

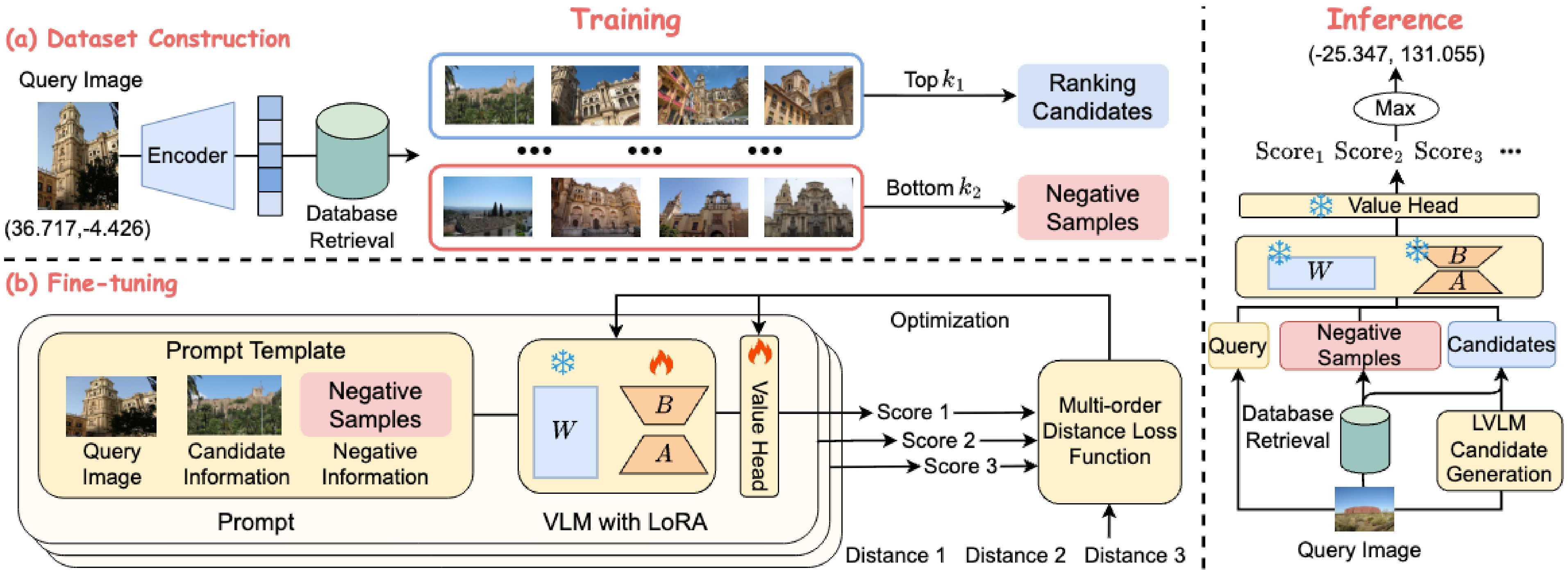
Content

01 Background & Motivation

02 Methodology

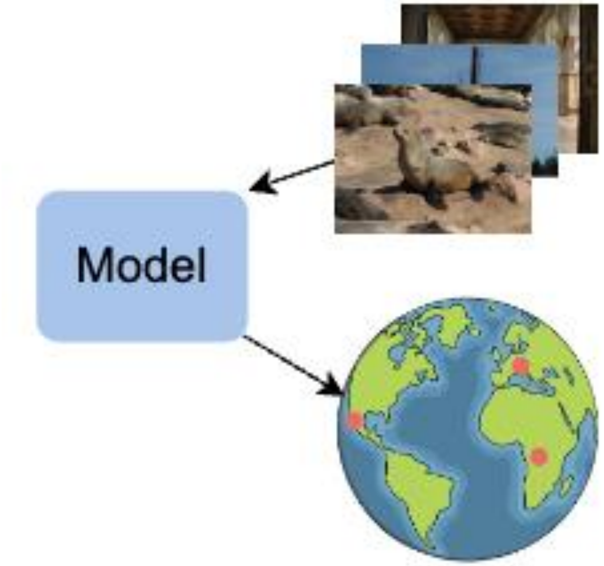
03 Experiments

04 Conclusion

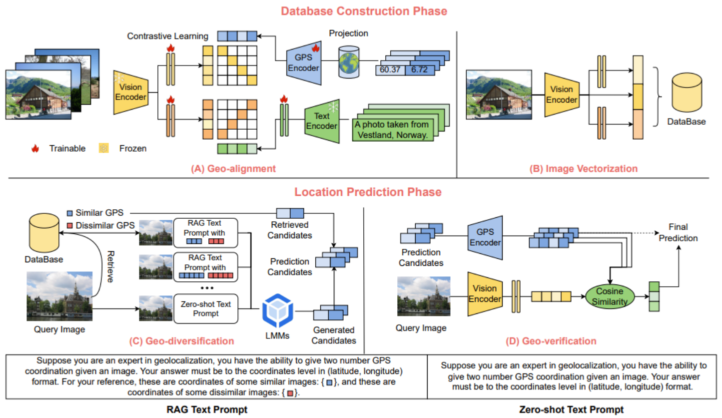


Two-stage Pipeline is Widely Adopted

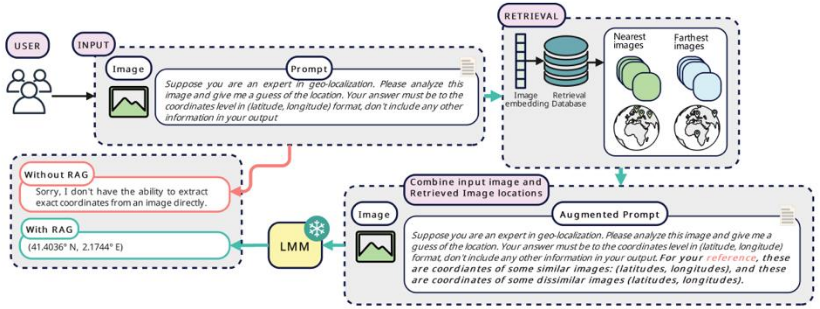
- Retrieve candidate locations from a global database
- Select the top match based on similarity scores



Worldwide Geolocalization



G3 [NeurIPS'24]

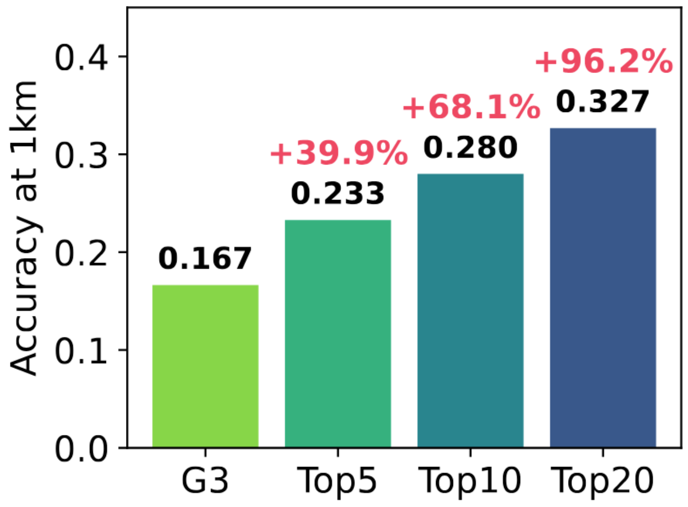


Img2Loc [SIGIR'24]

Background & Motivation

Key Observation

Better candidates often exist within the top-k, but are not selected



Root Cause

- Existing methods rely on simple **similarity heuristics** (e.g., cosine similarity of image embeddings)
- Existing training objectives primarily focus on **point-wise similarity** between individual images and locations, overlooking the rich spatial relationships among candidates.

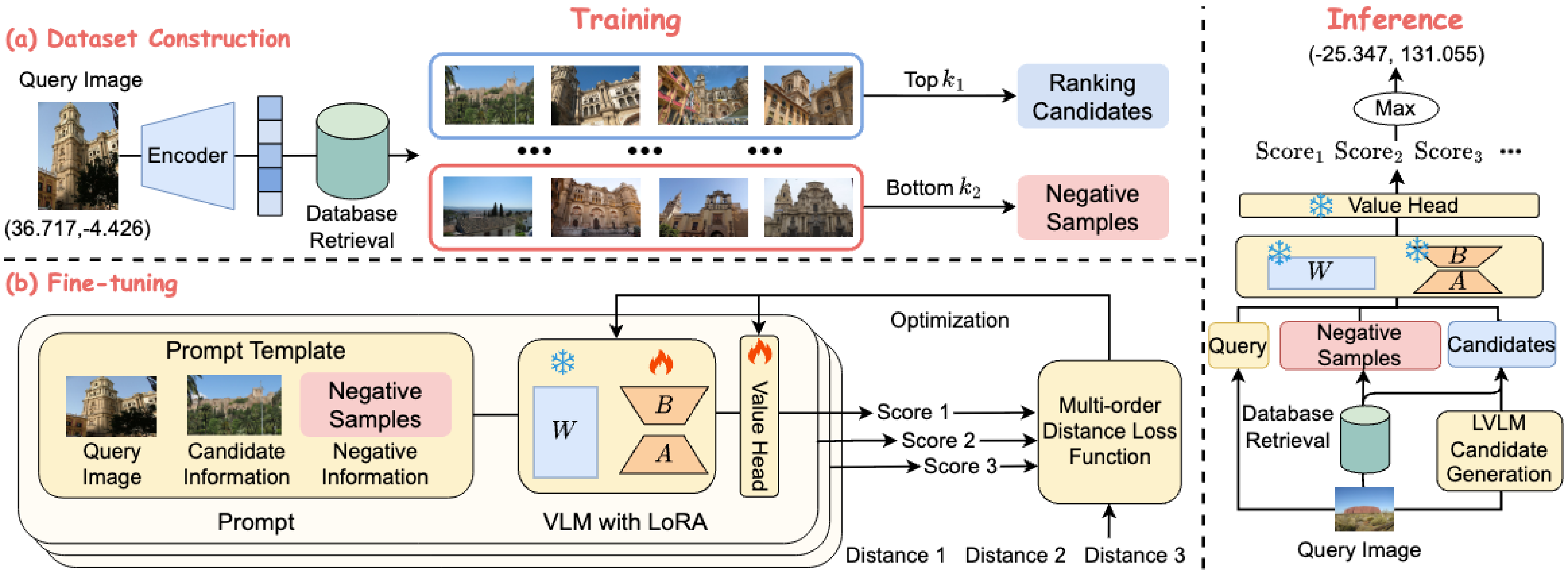
Content

01 Background & Motivation

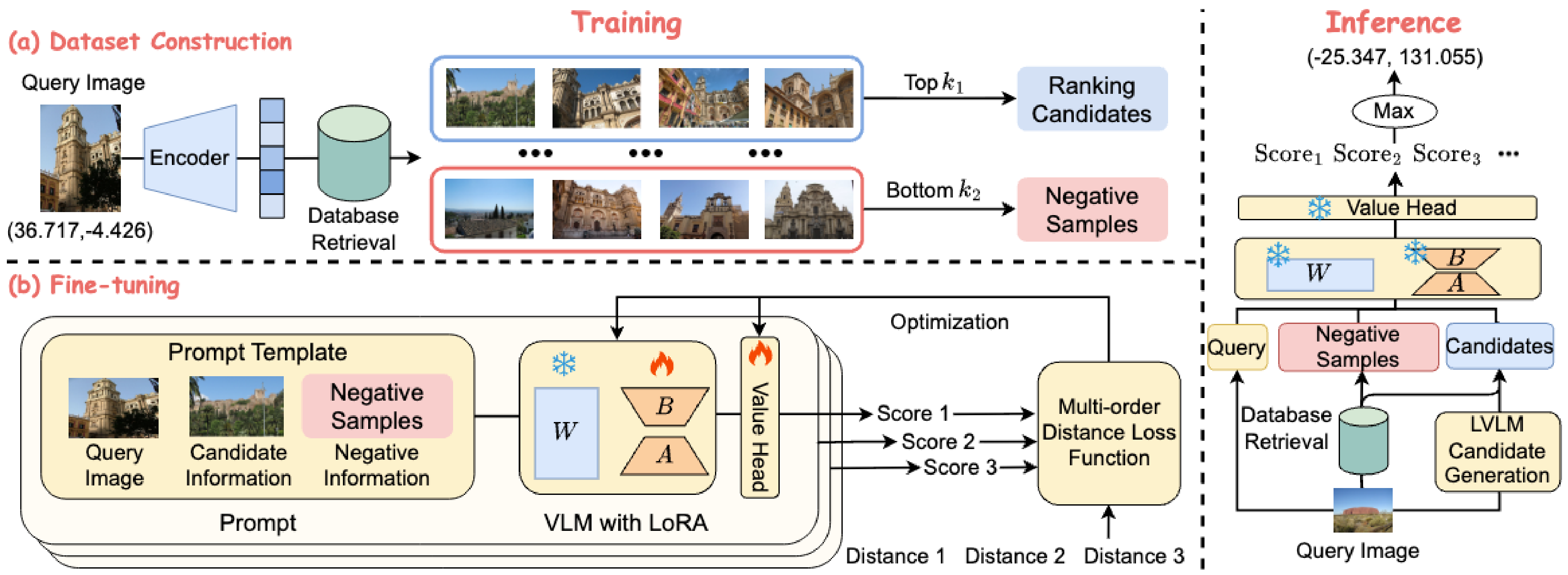
02 Methodology

03 Experiments

04 Conclusion



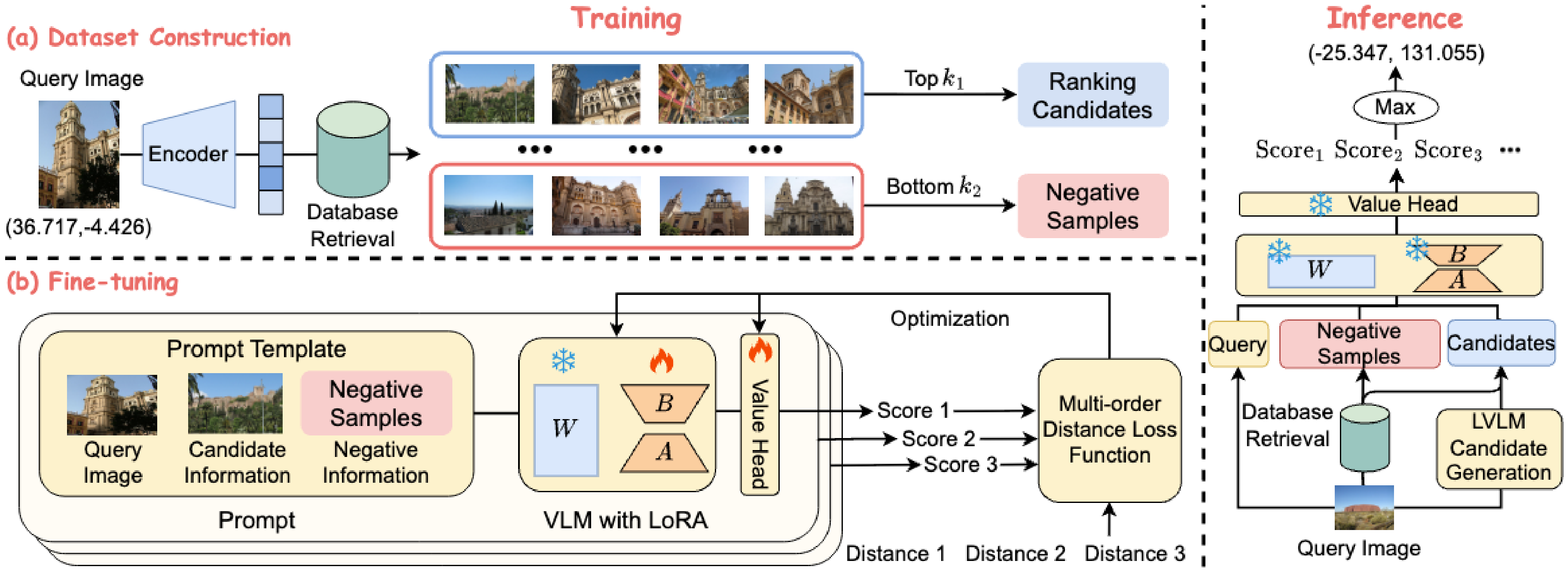
Methodology



GeoRanking Dataset Construction

- Candidate Encoding $\mathbf{v}_{c_m} = \text{concat}(\text{Encoder}_c^{\text{gps}}(c_m^{\text{gps}}), \text{Encoder}_c^{\text{text}}(c_m^{\text{text}}), \text{Encoder}^{\text{img}}(c_m^{\text{img}}))$
- Random Sampling Query Image, Query Encoding
- Retrieving Top-N Candidates $\mathbf{v}_q = \text{concat}(f_{\text{img}\to\text{gps}}(\text{Encoder}^{\text{img}}(q)), f_{\text{img}\to\text{text}}(\text{Encoder}^{\text{img}}(q)), \text{Encoder}^{\text{img}}(q))$

Methodology

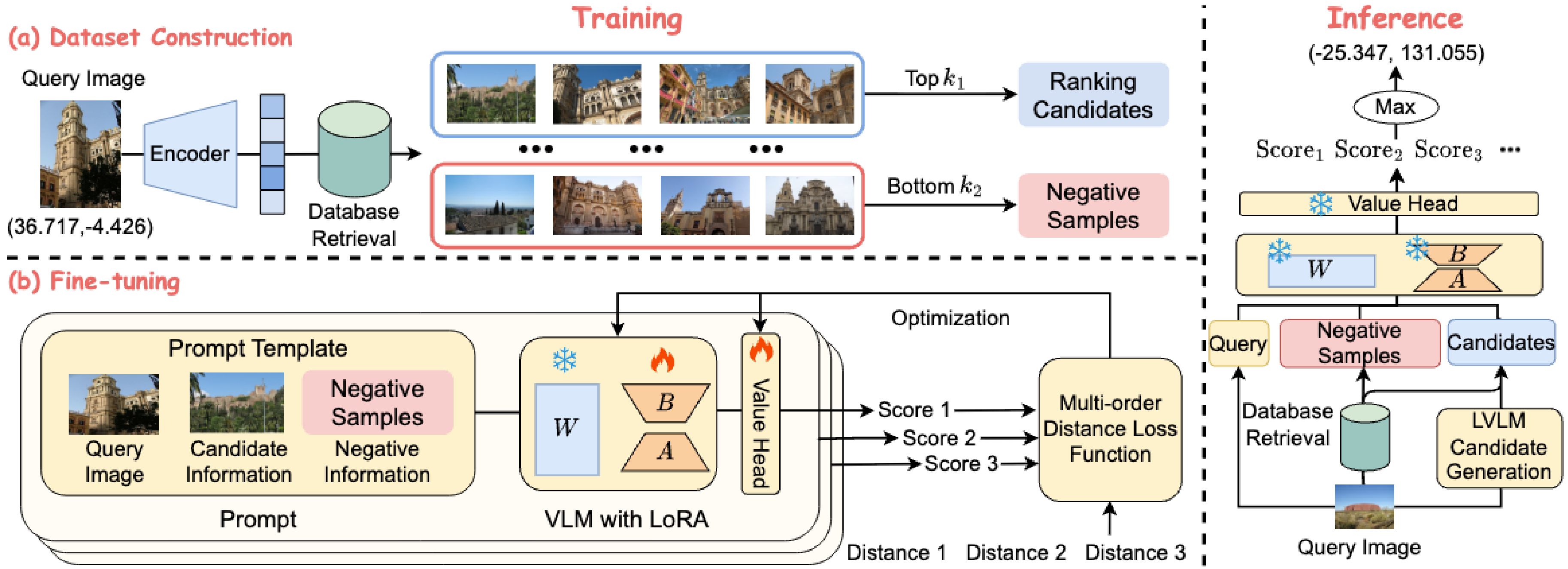


GeoRanker

- Prompt
- Model Architecture: LoRA + Value Head

{query image} How far is this place from latitude: {candidate latitude}, longitude: {candidate longitude}, {candidate textual descriptions}, {candidate image}? Negative examples: {negative information}.

$$s = \mathbf{w}^\top \mathbf{h}_{\text{final}}, \quad \text{where } \mathbf{h}_{\text{final}} = \text{LVLM}(\mathbf{x})_{[-1]}$$



GeoRanker: Optimization with Multi-Order Distance Objective

$$\mathcal{L}_{\text{PL}}^{(1)} = -\frac{1}{K^{(1)}} \sum_{i=1}^{K^{(1)}} \log \frac{\exp(s_{\pi(i)})}{\sum_{j=i}^{k_1} \exp(s_{\pi(j)})}$$

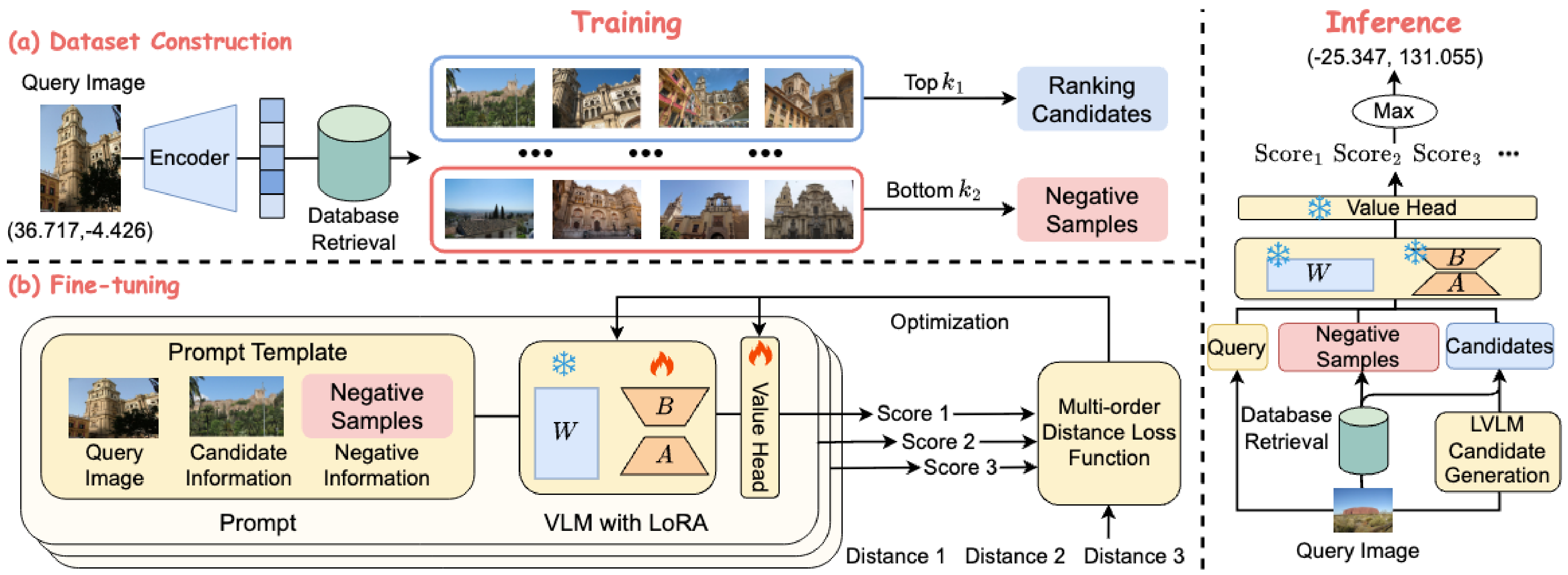
- First-order Distance Loss
- Second-order Distance Loss
- Joint Optimization

$$\Delta d_{i,j} = d_{\pi(i)} - d_{\pi(j)}, \quad \Delta s_{i,j} = s_{\pi(i)} - s_{\pi(j)}, \quad \text{for } 1 \leq i < j \leq k_1$$

$$\mathcal{L}_{\text{PL}}^{(2)} = -\frac{1}{K^{(2)}} \sum_{i=1}^{K^{(2)}} \log \frac{\exp(\Delta s_{(i)})}{\sum_{j=i}^{P} \exp(\Delta s_{(j)})}$$

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{PL}}^{(1)} + (1 - \lambda) \cdot \mathcal{L}_{\text{PL}}^{(2)}$$

Methodology



Inference

$$s_c = \mathrm{GeoRanker}(q, c), \quad \forall c \in \mathcal{C}_\mathrm{r} \cup \mathcal{C}_\mathrm{g}$$

$$\hat{c} = \arg \max_{c \in \mathcal{C}_\mathrm{r} \cup \mathcal{C}_\mathrm{g}} s_c$$

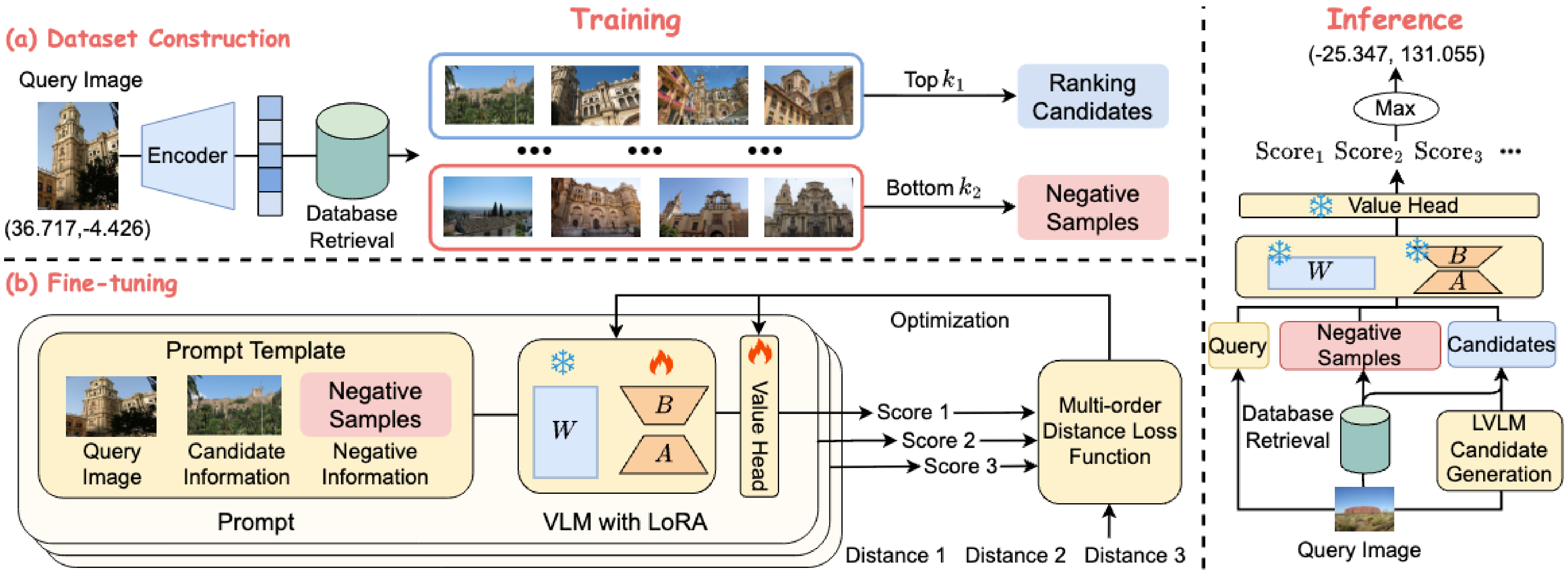
Content

01 Background & Motivation

02 Methodology

03 Experiments

04 Conclusion



Experiments

Table 1: **Main results** on IM2GPS3K and YFCC4K. For all metrics, higher is better. The best-performing results are highlighted in **bold**, while the second-best results are underlined. Δ represents the relative improvement of our method 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 [1]	ICCV'17	7.2	19.4	26.9	38.9	55.9	2.3	5.7	11	23.5	42
PlaNet [24]	ECCV'16	8.5	24.8	34.3	48.4	64.6	5.6	14.3	22.2	36.4	55.8
CPlaNet [15]	ECCV'18	10.2	26.5	34.6	48.6	64.6	7.9	14.8	21.9	36.4	55.5
ISNs [55]	ECCV'18	10.5	28	36.6	49.7	66	6.5	16.2	23.8	37.4	55
Translocator [25]	ECCV'22	11.8	31.1	46.7	58.9	80.1	8.4	18.6	27	41.1	60.4
GeoDecoder [26]	ICCV'23	12.8	33.5	45.9	61	76.1	10.3	24.4	33.9	50	68.7
GeoCLIP [8]	NeurIPS'23	14.11	34.47	50.65	69.67	83.82	9.59	19.31	32.63	55	74.69
Img2Loc [10]	SIGIR'24	15.34	39.83	53.59	69.7	82.78	19.78	30.71	41.4	58.11	74.07
PIGEON [9]	CVPR'24	11.3	36.7	53.8	72.4	85.3	10.4	23.7	40.6	62.2	77.7
G3 [14]	NeurIPS'24	16.65	40.94	55.56	71.24	84.68	23.99	35.89	46.98	64.26	78.15
GeoRanker	Ours	**18.79**	**45.05**	**61.49**	**76.31**	**89.29**	**32.94**	**43.54**	**54.32**	**69.79**	**82.45**
Rel. Improvement	Δ	↑ 12.9%	↑ 10.0%	↑ 10.7%	↑ 5.4%	↑ 4.7%	↑ 37.3%	↑ 21.3%	↑ 15.6%	↑ 8.6%	↑ 5.5%

- Superior Performance (37.3% improvement on YFCC4K in Street level)
- State-of-the-art across all datasets and metrics

Experiments

Methods	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
w/o$\mathcal{L}_{\text{PL}}^{(2)}$	18.48	44.61	60.96	75.61	88.28
w/o $\mathcal{C}_{\text{neg}}$	17.35	44.51	60.82	76.37	88.28
w/o c_m^{text}	18.02	43.91	60.19	**76.61**	88.62
w/o c_m^{img}	15.58	41.77	59.15	75.40	88.35
w/o $\mathcal{C}_g$	18.21	43.47	59.69	75.47	88.75
Ours	**18.79**	**45.05**	**61.49**	76.31	**89.29**

Ablation Study

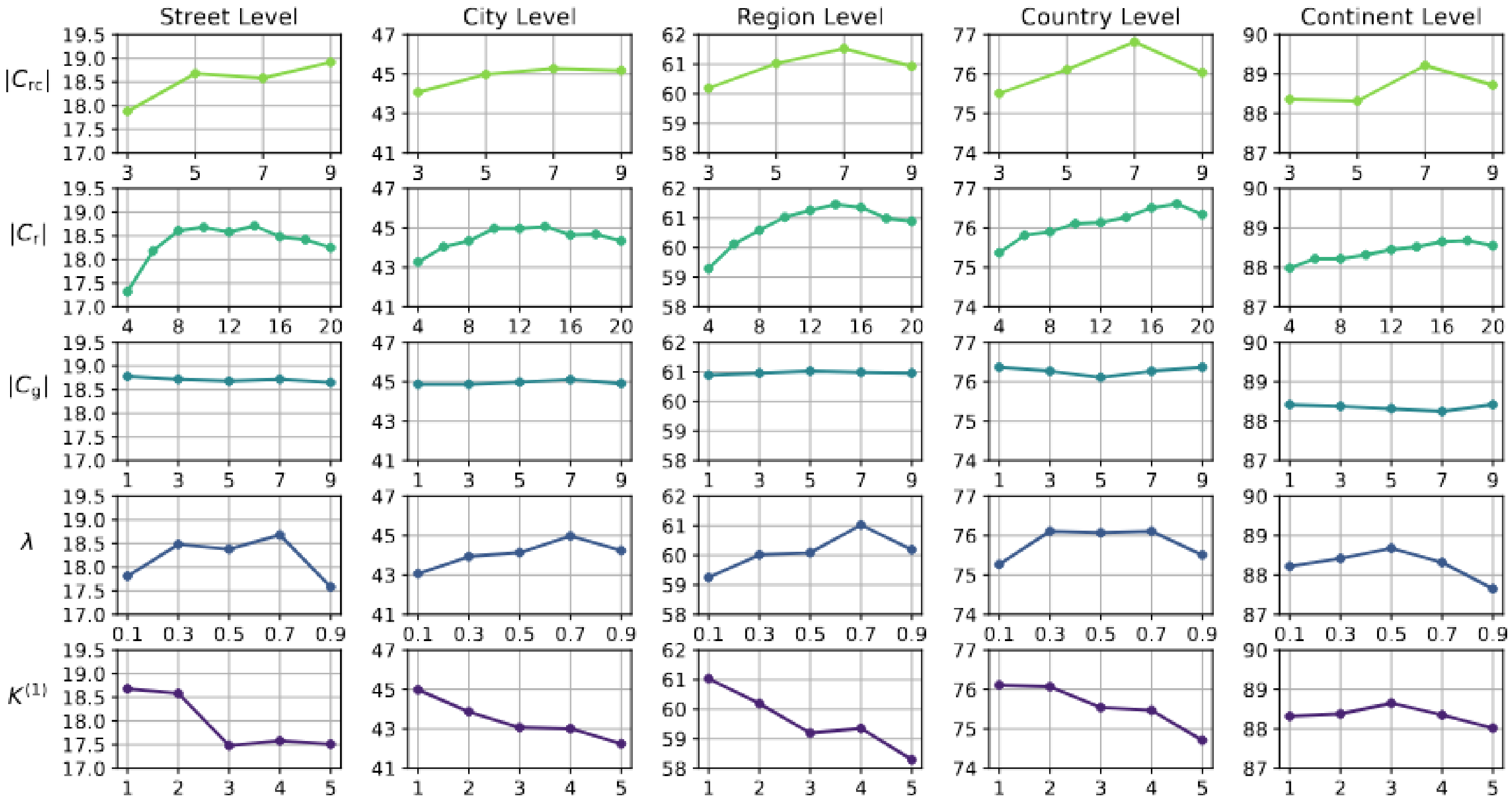
Methods	IM2GPS3K				
	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
Random	10.04	29.72	42.17	57.82	75.24
Top1	13.31	34.03	45.48	61.56	78.04
Prompting	16.62	40.21	54.55	70.07	83.24
Ours	**18.79**	**45.05**	**61.49**	**76.31**	**89.29**

Comparison with Other Ranking Baselines

- All components contribute positively
- Removing any of the modality-aware prompt components leads to performance drops
- Without generated candidates underperforms GeoRanker

- GeoRanker is superior to Random, Top-1 Selection, and Prompting baselines.

Experiments



- Impact of candidate scales in training and inference: $\mathrm{C_{rc}}$, $\mathrm{C_r}$, $\mathrm{C_g}$
- Impact of hyperparameters in multi-order distance objective: λ, $\mathrm{K}^{(1)}$

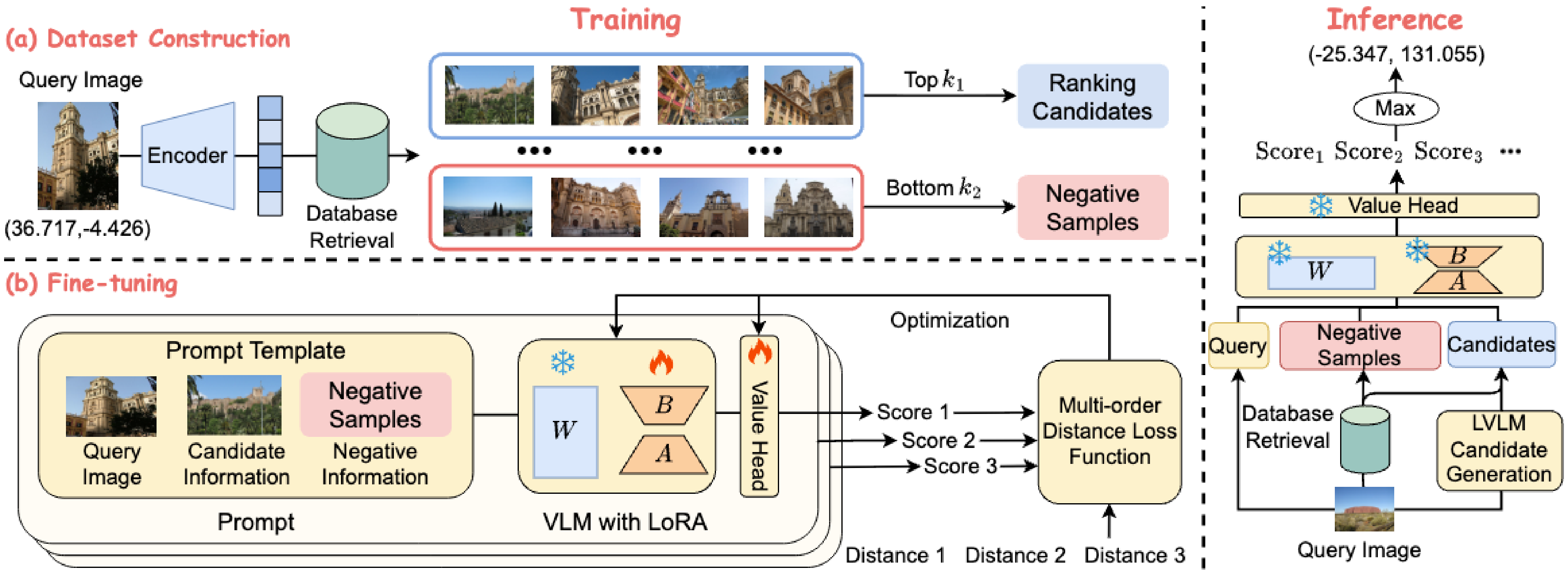
Experiments

Case Study

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

More Information on Training and Inference

Content


Training
(a) Dataset Construction
Query Image
Encoder
(36.717,-4.426)
Database Retrieval
Top k_1
Ranking Candidates
Bottom k_2
Negative Samples
(b) Fine-tuning
Prompt Template
Query Image
Candidate Information
Negative Samples
Negative Information
Prompt
W
B
A
Value Head
VLM with LoRA
Optimization
Score 1
Score 2
Score 3
Multi-order Distance Loss Function
Distance 1
Distance 2
Distance 3
Inference
(-25.347, 131.055)
Max
Score_1 Score_2 Score_3 ...
Value Head
Query
Negative Samples
Candidates
Database Retrieval
LVLM Candidate Generation
Query Image

Conclusion

- In this paper, we propose GeoRanker, a distance-aware ranking framework built upon LVLM.
- To enhance training, we introduce a novel multi-order distance loss that captures both absolute distances and relative spatial relationships among candidate locations.
- To support this framework, we construct GeoRanking, the first dataset specifically designed for spatial ranking tasks.
- Extensive experiments on IM2GPS3K and YFCC4K demonstrate the effectiveness of GeoRanker over baselines.

Code | **MP16-Pro Dataset** | **CityU AML Lab** | **Pengyue's HomePage**

Thanks

JIA Pengyue

Applied Machine Learning Lab
City University of Hong Kong
jia.pengyue@my.cityu.edu.hk

Department of Computer Sciences
University of Wisconsin - Madison
pjia7@wisc.edu