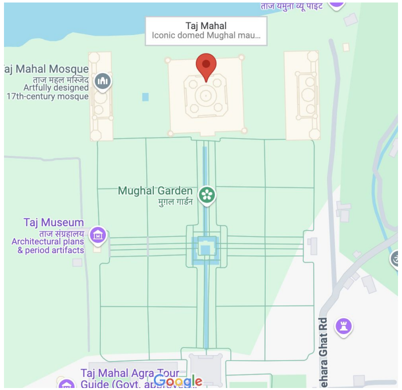
CLIP-based Image Geolocation using Hierarchical Feature Learning and RAG

Akshay Raman (ar8692)
Aman Gupta (ag9960)
Prithviraj Murthy (pkm5789)
Satyanarayana Chillale (sc9960)
Srikanth Balakrishna (sb9558)

Problem Statement

- Image Geolocation: Find the precise location of an image taken anywhere on Earth.
- Challenges:
  - Diversity of images. Need large datasets and models.
  - How do you predict at a global scale? Standard classification/regression techniques are infeasible/inaccurate.

Image

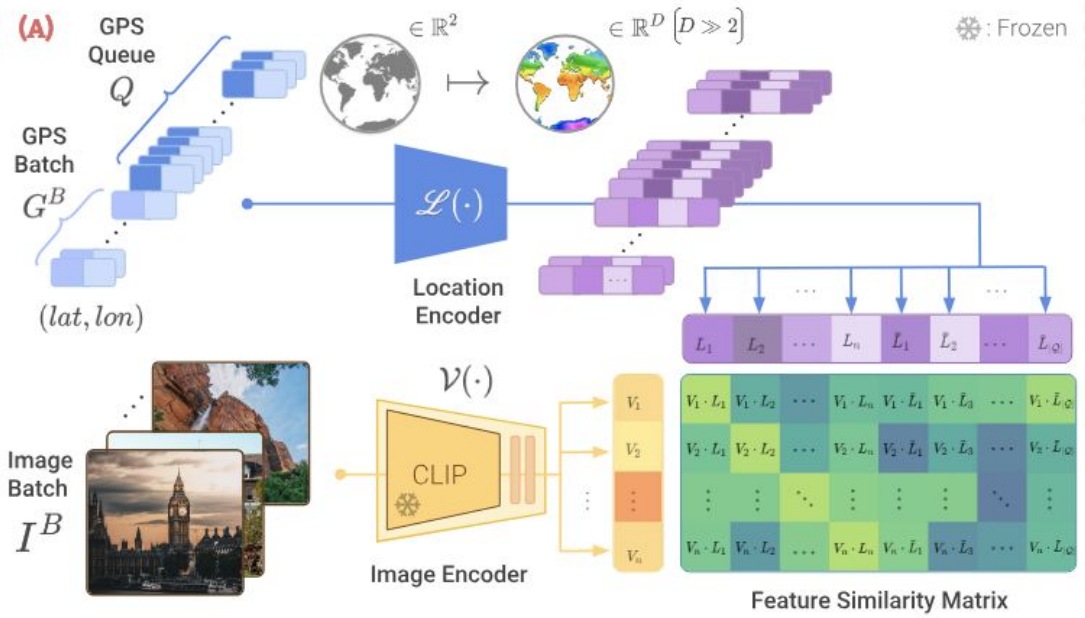


GPS Coordinates [LAT, LON]

Contributions

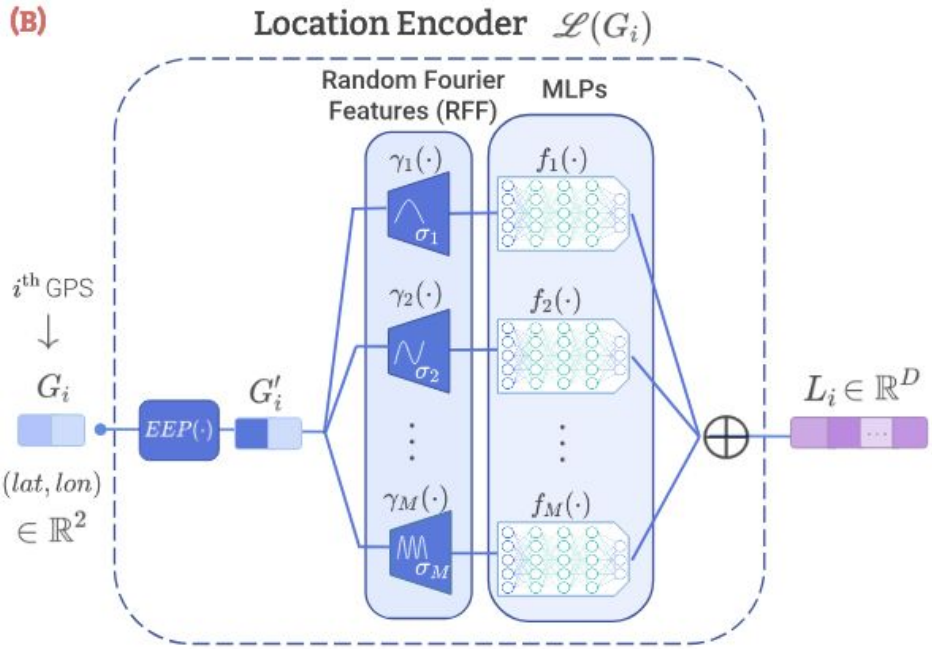
1. Trained a CLIP-based image geolocation model on the MediaEval-16 Dataset (4M+ images).
2. Designed a novel inference approach based on hierarchical feature clustering which achieves comparable performance while being **~100x more efficient** than previous methods.
3. Conducted RAG-based text inference using LLMs.

Architecture Diagram

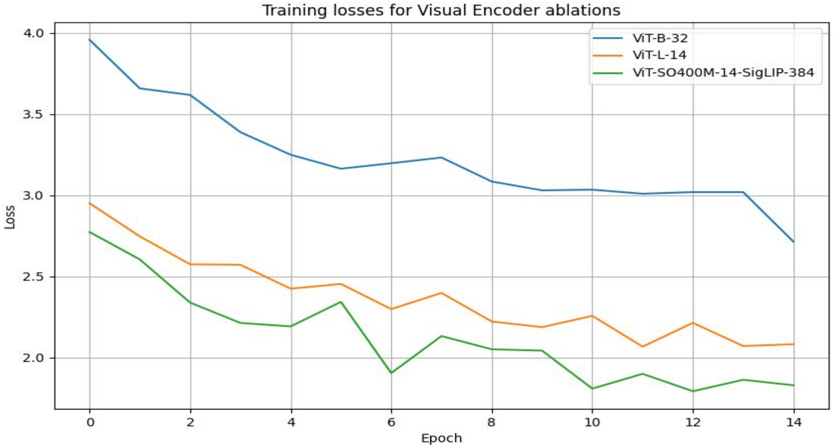


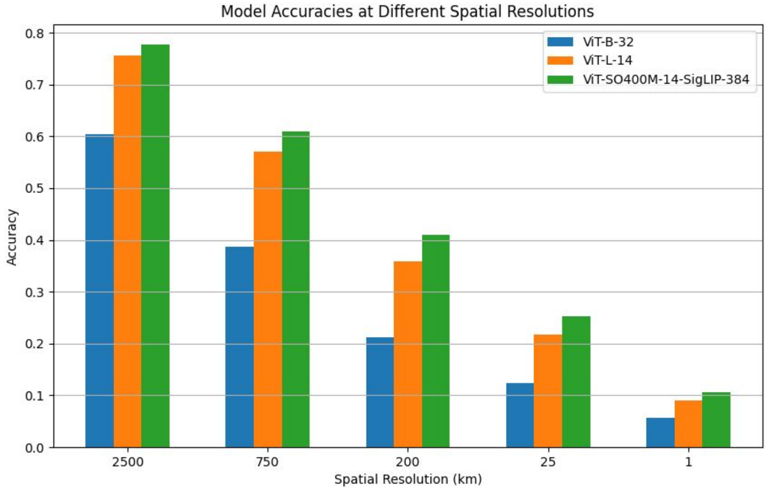
Vivanco Cepeda, Vicente, Gaurav Kumar Nayak, and Mubarak Shah. "Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization." Advances in Neural Information Processing Systems 36 (2024).

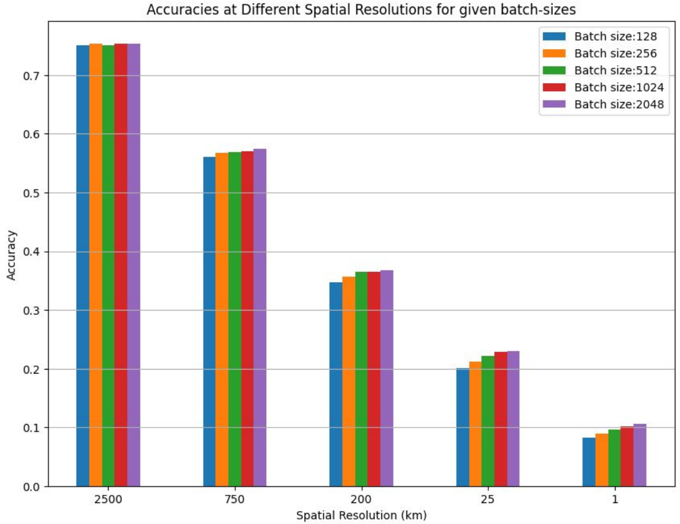
Location Encoder - Deep Dive



Model Training



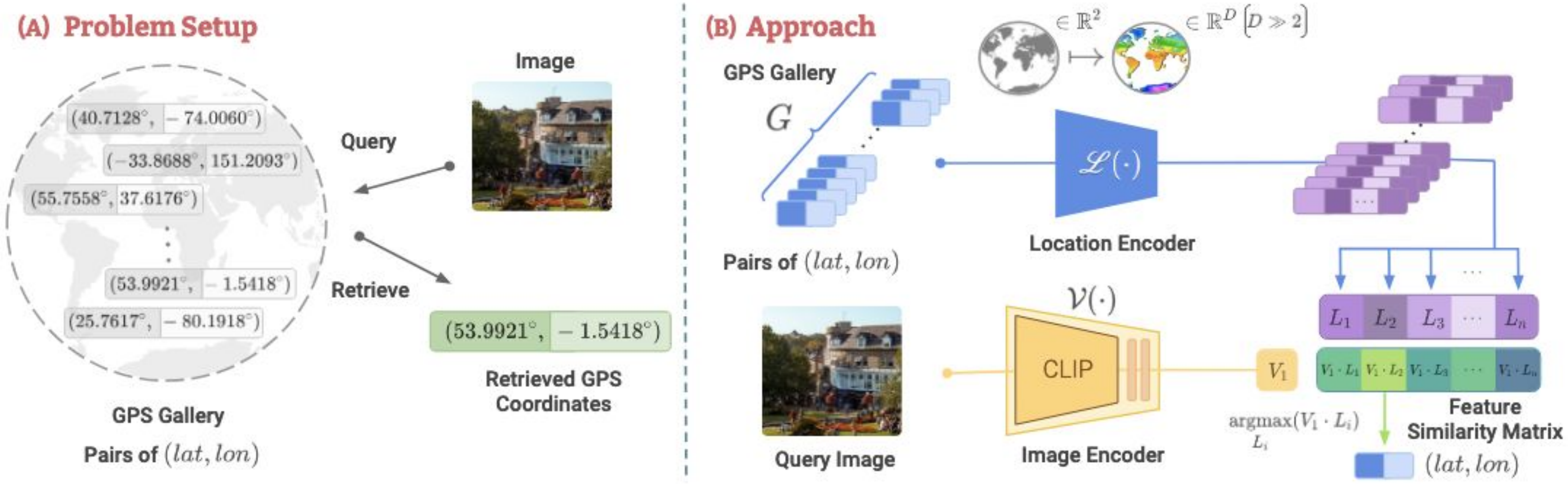




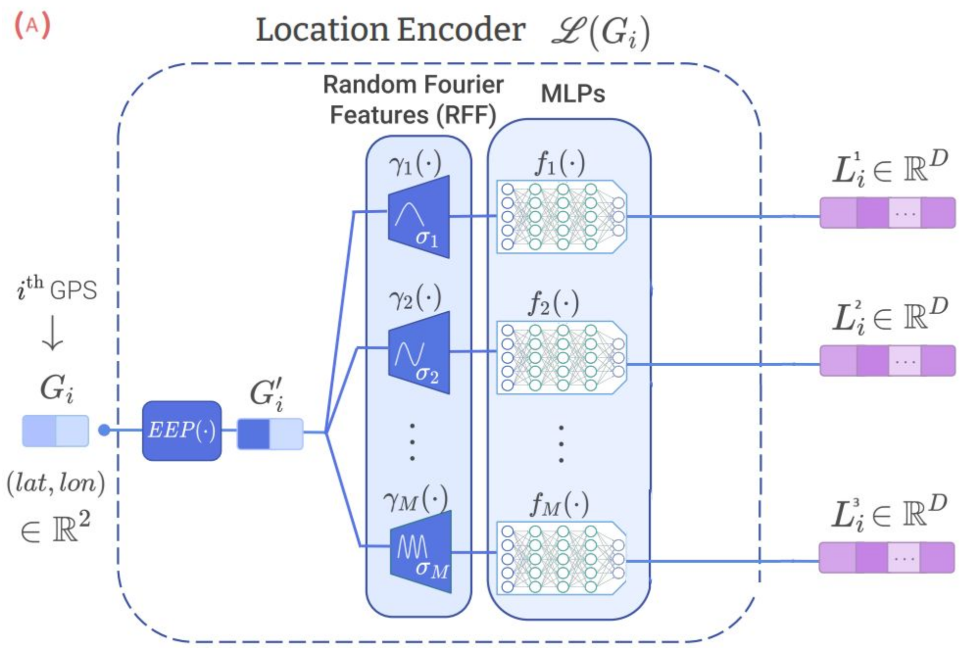
Inference Methods

1. Original (from GeoCLIP paper)
2. Hierarchical Feature Clustering
3. RAG-based Method

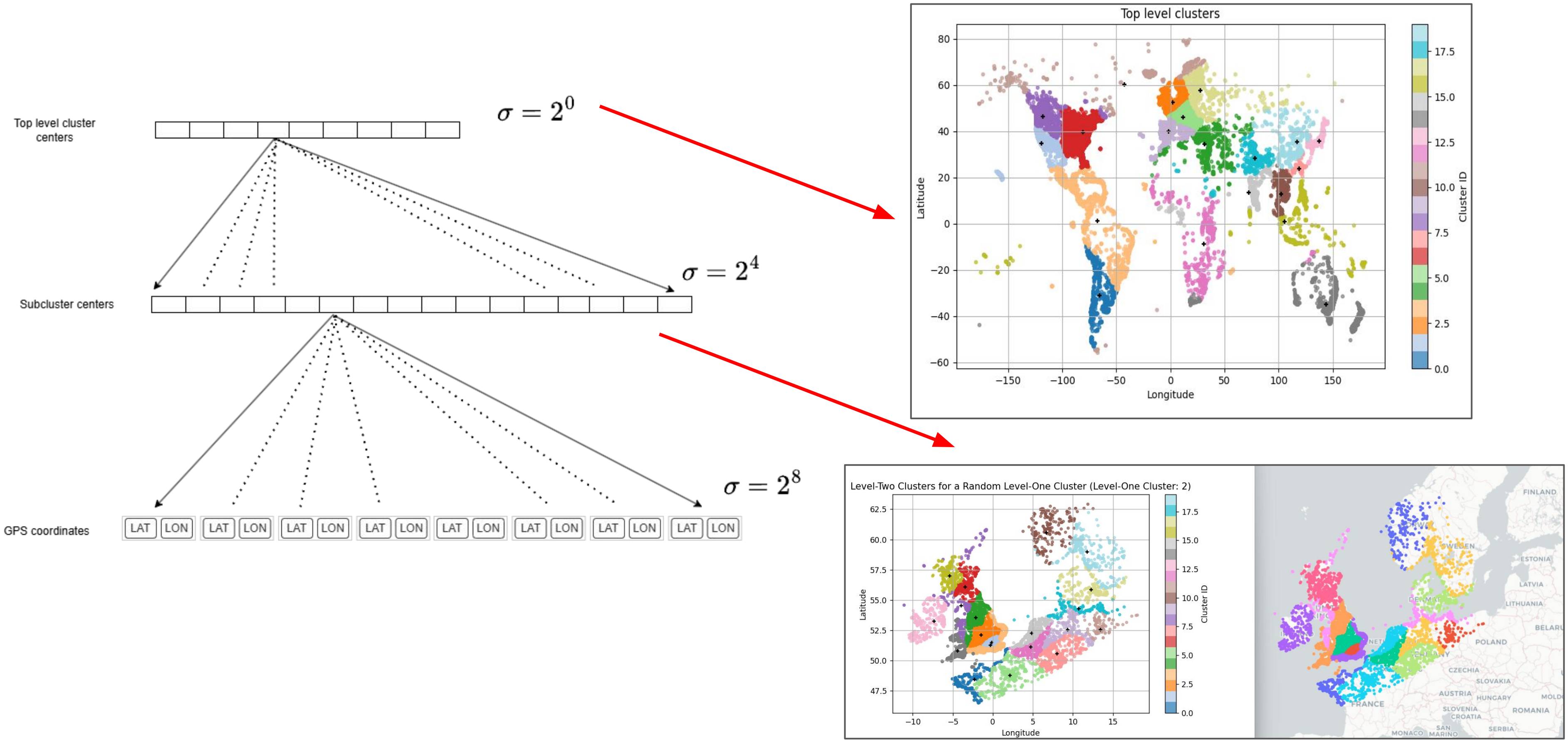
Original Inference Approach



Hierarchical Features Clustering



- We produce 3 RFF encodings using 3 different sigma values that determine the encodings fed into each trained MLP capsule to capture features at different granularities.
- Instead of aggregating these features into a single embedding, We utilize each embedding separately to perform clustering at different levels/global distance scales.



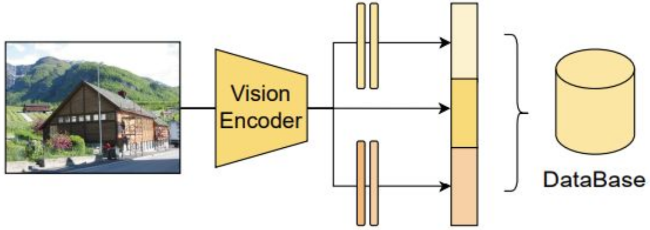
Top level cluster centers
$\sigma = 2^0$
Subcluster centers
$\sigma = 2^4$
GPS coordinates
$\sigma = 2^8$
LAT LON LAT LON LAT LON LAT LON LAT LON LAT LON LAT LON LAT LON
Top level clusters
Latitude
Longitude
Cluster ID
80 60 40 20 0 −20 −40 −60
−150 −100 −50 0 50 100 150
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5
Level-Two Clusters for a Random Level-One Cluster (Level-One Cluster: 2)
Latitude
Longitude
Cluster ID
62.5 60.0 57.5 55.0 52.5 50.0 47.5
−10 −5 0 5 10 15
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5
FINLAND
SWEDEN
ESTONIA
LATVIA
LITHUANIA
BELARUS
POLAND
CZECHIA
SLOVAKIA
AUSTRIA
HUNGARY
SLOVENIA
CROATIA
ROMANIA
SERBIA
FRANCE
MONACO
SAN MARINO

Cluster Sizes Comparison

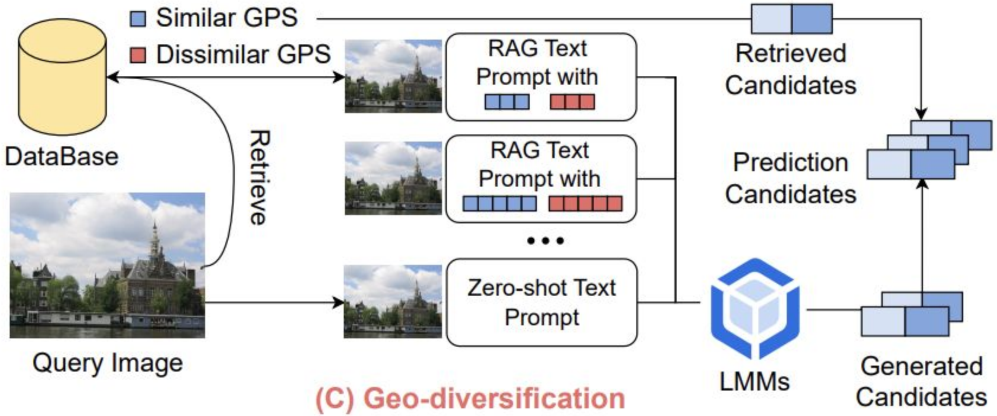
GPS Gallery Tree	**Accuracy at 2500km**	**Accuracy at 750km**	**Accuracy at 200km**	**Accuracy at 1km**	**% of Coordinates Considered**
100000 (original)	0.753	0.568	0.351	0.084	100%
200 (one-level)	0.618	0.390	0.188	0.047	0.7%
800 (one-level)	**0.710**	**0.481**	**0.253**	**0.049**	**1%**
1000 (one-level)	0.701	0.480	0.248	0.050	2%
20, 100 (two-level)	0.676	0.430	0.200	0.034	0.2%
100, 20 (two-level)	0.636	0.398	0.177	0.029	0.2%
200, 10 (two-level)	**0.676**	**0.429**	**0.200**	**0.033**	**0.3%**

Inference: RAG with LMM

- We also incorporated **text embeddings** for country, state, and city information and trained the model again.
- During inference, we utilize **multiple RAG prompts** with LLMs (GPT-4o and LLaMA3-LLAVA-Next-8B) and select the best response as the final output.

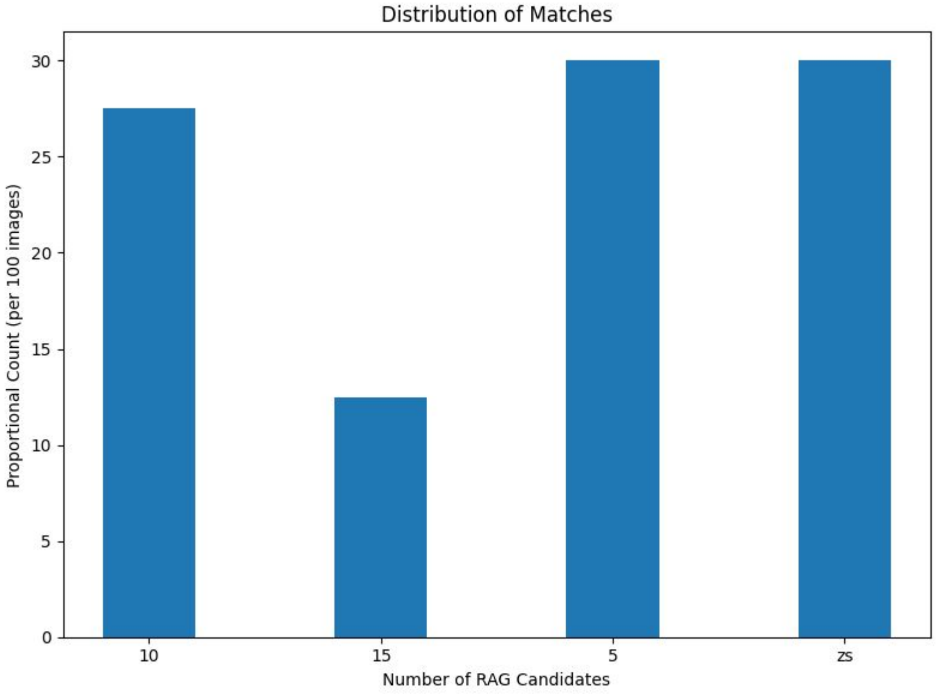


(B) Image Vectorization



(C) Geo-diversification

RAG comparisons



Comparisons

Test Dataset: IM2GPS3k

Methods	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
PlaNet [22]	8.5	24.8	34.3	48.4	64.6
GeoCLIP [29]	14.11	34.47	50.65	69.67	83.82
Hierarchical Clustering	9.2	30.26	40.46	67.85	79.57
RAG	15.01	32.53	60.06	72.5	85.08

Samples

Images	Prediction	Geodesic
 Grand Chavalard	Lat: 45.96096 Long: 6.94477	26km
 Sydney Tower	Lat: -33.8708476 Long: 151.2073203	1km

Future Work

- Hierarchical Feature Clustering
  - Exploring beam search algorithms to improve accuracy.
  - Scaling to large GPS gallery sizes (1M+).
- Use fine grained text information like neighborhood and county.

Thank You