NEURAL INFORMATION
PROCESSING SYSTEMS

# Recognition through Reasoning: Reinforcing Image Geo-localization with Large Vision-Language Models

Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, Jiaheng Wei*

**NeurIPS 2025**

香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE
AND TECHNOLOGY (GUANGZHOU)

香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

# Background

## What is Image Geo-localization

**Q: Where was this photo taken?**

### ① Classification

Classification-based methods treat geo-localization as a discrete prediction task, assigning each image to a predefined set of geographical regions or cells.

### ② Retrieval

Retrieval-based methods estimate location by comparing the query image to a large geo-tagged reference database, retrieving the closest match in terms of visual features, geographic coordinates, or semantic labels (e.g., city or country names).

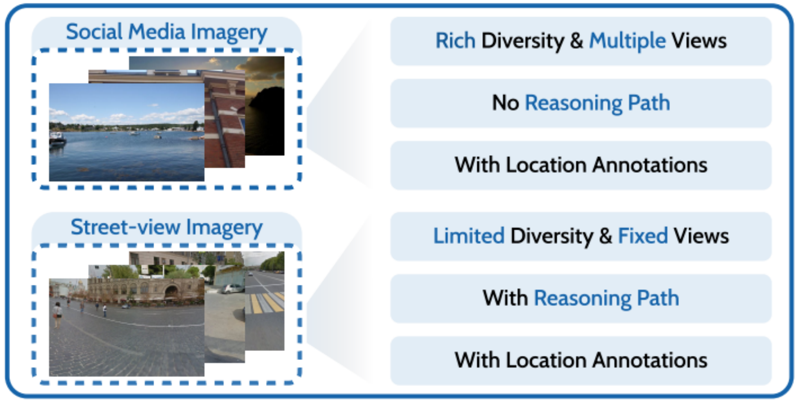
### ③ Generation

The emergence of Large Vision-Language Models (LVLMs) has introduced a new paradigm to tackle image geo-localization. These methods are capable of generating both location predictions and explanations, offering greater interpretability in how decisions are made.

# Motivation

## Limitations in LVLM-based Image Geo-localization

### ① Data



Social Media Imagery

Rich Diversity & Multiple Views

No Reasoning Path

With Location Annotations

Street-view Imagery

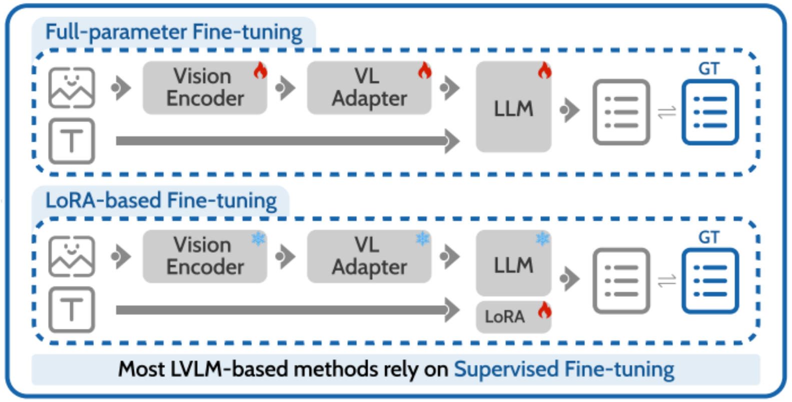
Limited Diversity & Fixed Views

With Reasoning Path

With Location Annotations

- Lack of Reasoning Supervision
  - Existing geo-localization datasets rarely include explicit reasoning — e.g., interpretations of visual cues or justifications for location decisions.
- Over-Reliance on Street-View Imagery
  - Recent efforts focus on street-view data → limited diversity & fixed views.
- Poor Generalization to Real-World Scenes
  - Models trained this way struggle with real-world visual conditions.

### ② Modeling



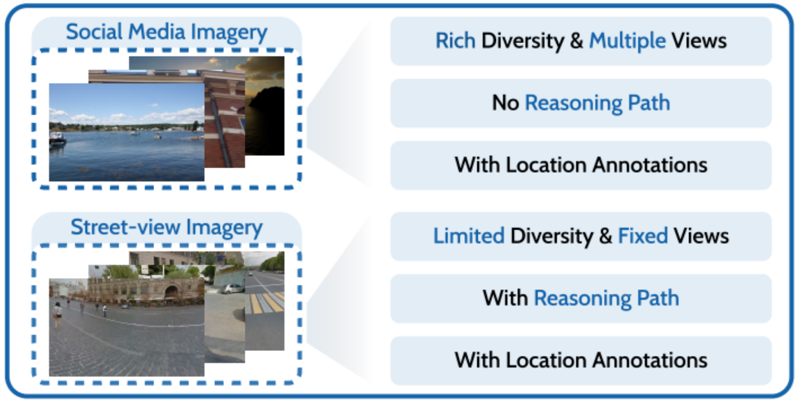
Most LVLM-based methods rely on Supervised Fine-tuning

- All current SFT approaches (full-parameter / LoRA) use instruction-style data → encourage pattern replication

# Methods

## How GLOBE Tackles the Challenges

### ① Data

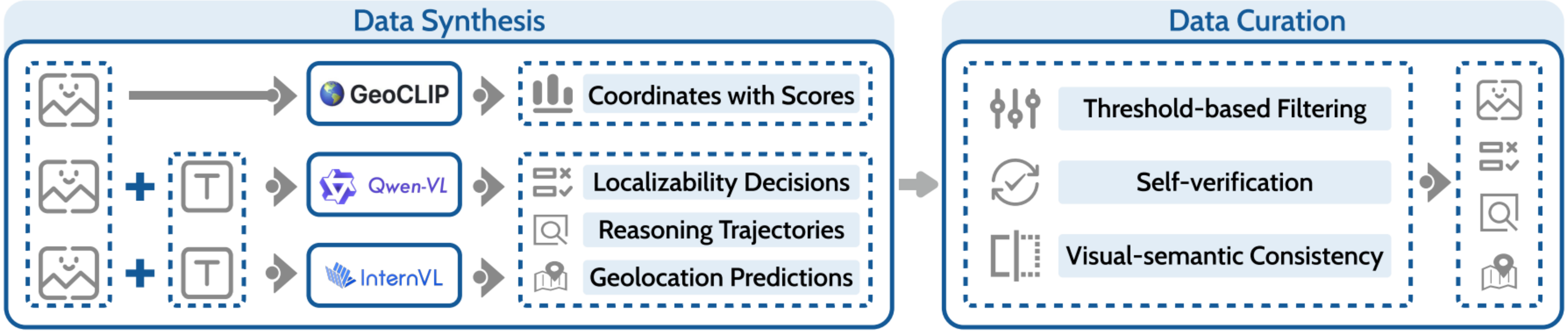


Social Media Imagery
- Rich Diversity & Multiple Views
- No Reasoning Path
- With Location Annotations

Street-view Imagery
- Limited Diversity & Fixed Views
- With Reasoning Path
- With Location Annotations

- How much data is enough
- How to construct trustworthy reasoning trajectories

➢ Multiple Vision-Language Models Knowledge Distillation + Multi-dimensional Verification



**Data Synthesis**

- GeoCLIP → Coordinates with Scores
- Qwen-VL, InternVL → Localizability Decisions, Reasoning Trajectories, Geolocation Predictions

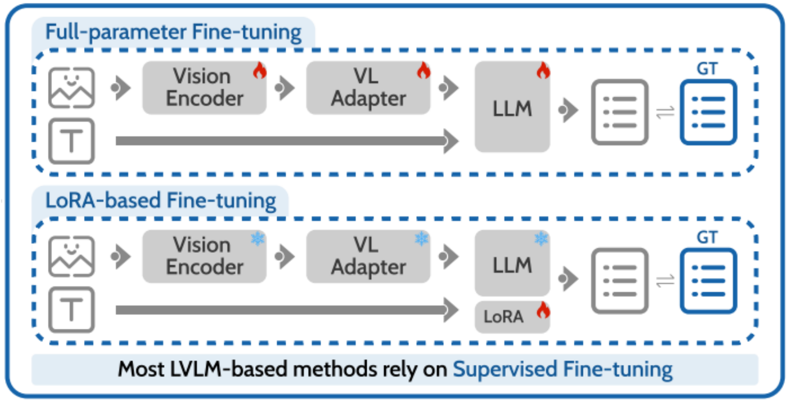
**Data Curation**

- Threshold-based Filtering
- Self-verification
- Visual-semantic Consistency

# Methods

## How GLOBE Tackles the Challenges

### ② Modeling



Most LVLM-based methods rely on Supervised Fine-tuning

- How to enable efficient fine-tuning via curated reasoning data

➢ We develop three task-specific rewards to assess distinct dimensions of reasoning quality

  ➢ **Localizability Reward**

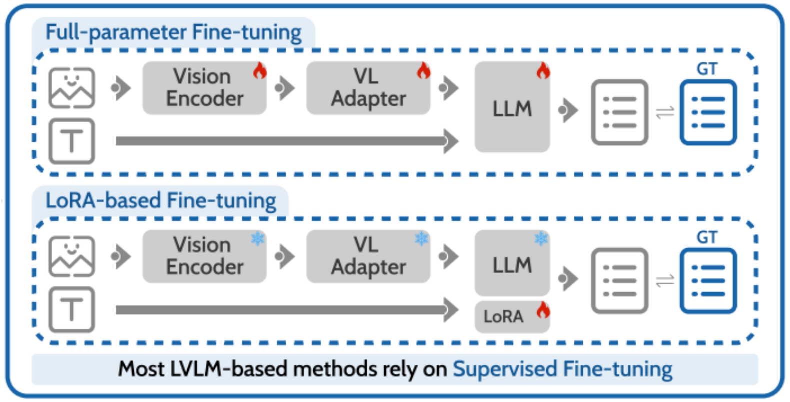
  ➢ **Visual Grounding Consistency Reward**

  ➢ **Geo-localization Accuracy Reward**

# Methods

## How GLOBE Tackles the Challenges

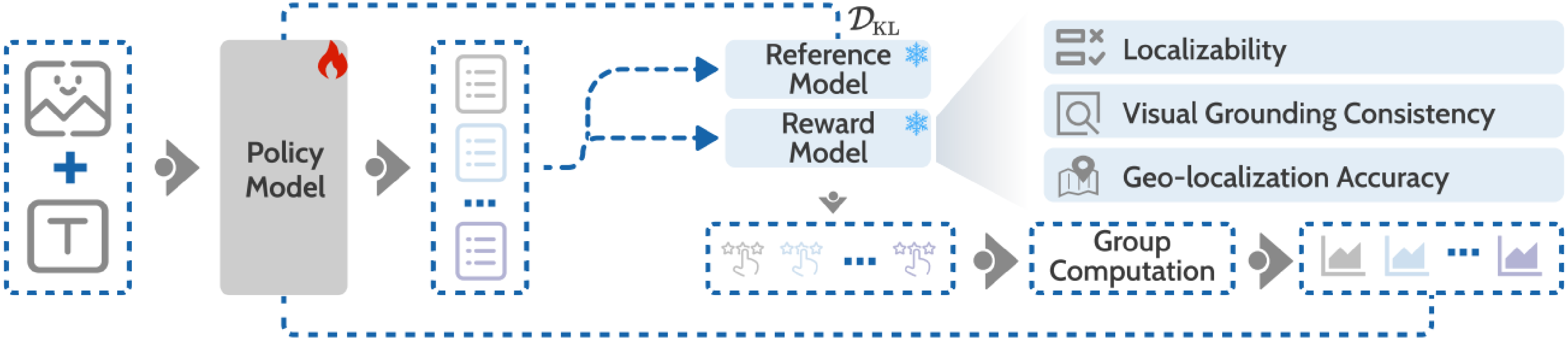
### ② Modeling



Most LVLM-based methods rely on Supervised Fine-tuning

- How to enable efficient fine-tuning via curated reasoning data

➢ Using reward signals, we fine-tune the base model with GRPO (Group Relative Policy Optimization)


Policy Model
$\mathcal{D}_{\mathrm{KL}}$
Reference Model
Reward Model
Localizability
Visual Grounding Consistency
Geo-localization Accuracy
Group Computation

# Experiments

## Experimental Setup

### Datasets

- For Data Curation
  - GeoCLIP
  - Qwen2.5-VL-72B
  - InternVL3-78B

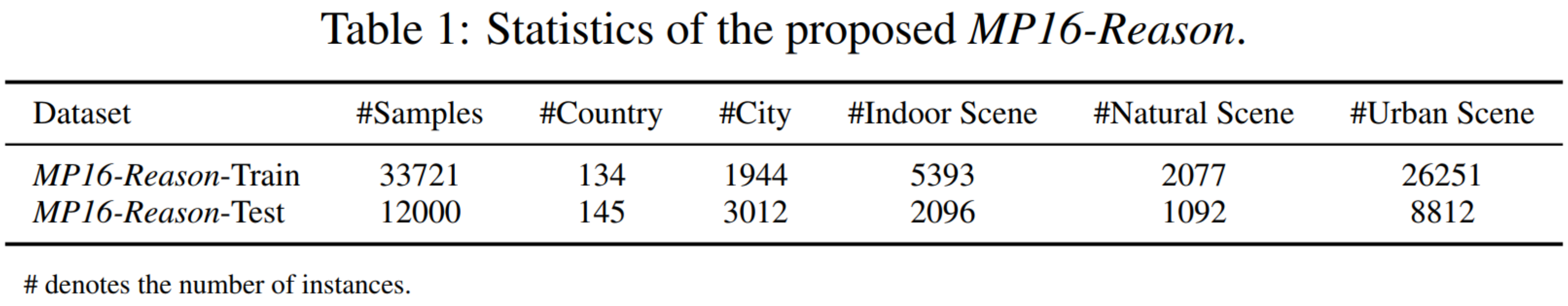


Table 1: Statistics of the proposed *MP16-Reason*.

| Dataset | #Samples | #Country | #City | #Indoor Scene | #Natural Scene | #Urban Scene |
|---|---|---|---|---|---|---|
| *MP16-Reason*-Train | 33721 | 134 | 1944 | 5393 | 2077 | 26251 |
| *MP16-Reason*-Test | 12000 | 145 | 3012 | 2096 | 1092 | 8812 |

# denotes the number of instances.

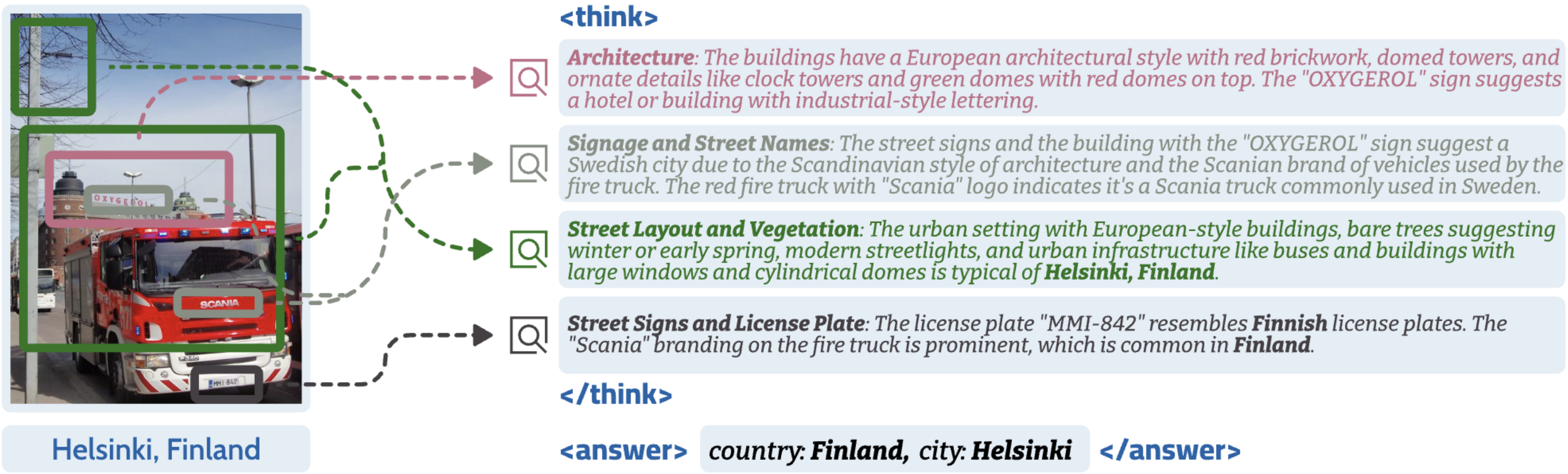
### Evaluation Metrics

We follow previous work and report the percentage of predictions whose geographic distance to the ground-truth coordinate falls within fixed thresholds (1km, 25km, 200km, 750km, and 2500km). Since our model outputs discrete place names (e.g., country or city), we concatenate the predicted city and country into a single string and query Microsoft Azure Maps, which returns the corresponding representative GPS coordinate (e.g., the geographic center of the region) for evaluation.

# Experiments

## Experimental Results

### Qualitative Result



Helsinki, Finland

**<think>**

***Architecture**: The buildings have a European architectural style with red brickwork, domed towers, and ornate details like clock towers and green domes with red domes on top. The "OXYGEROL" sign suggests a hotel or building with industrial-style lettering.*

***Signage and Street Names**: The street signs and the building with the "OXYGEROL" sign suggest a Swedish city due to the Scandinavian style of architecture and the Scanian brand of vehicles used by the fire truck. The red fire truck with "Scania" logo indicates it's a Scania truck commonly used in Sweden.*

***Street Layout and Vegetation**: The urban setting with European-style buildings, bare trees suggesting winter or early spring, modern streetlights, and urban infrastructure like buses and buildings with large windows and cylindrical domes is typical of **Helsinki, Finland**.*

***Street Signs and License Plate**: The license plate "MMI-842" resembles **Finnish** license plates. The "Scania" branding on the fire truck is prominent, which is common in **Finland**.*

**</think>**

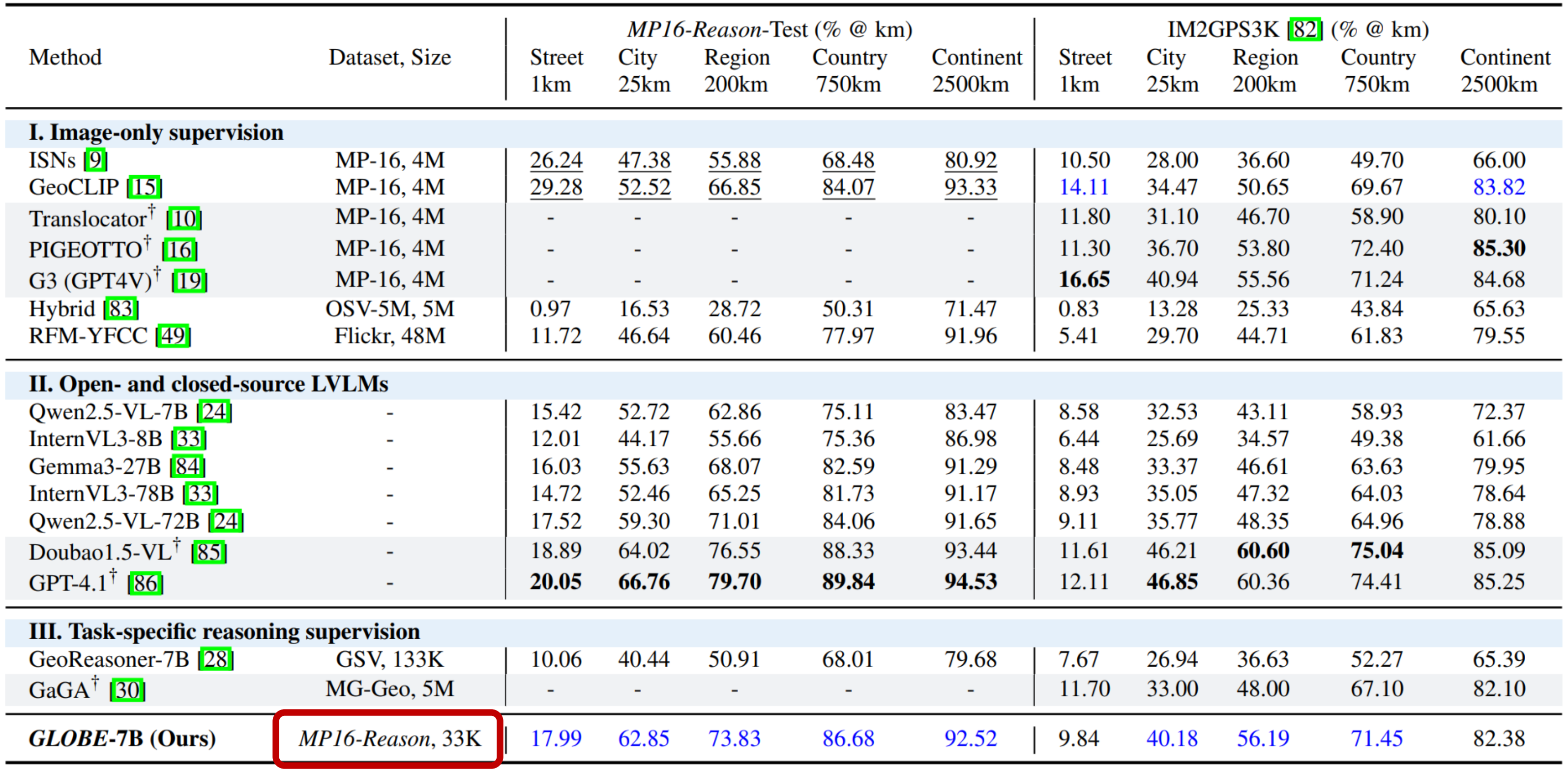
**<answer>** *country: **Finland**, city: **Helsinki*** **</answer>**

# Experiments

## Experimental Results

### Baseline Comparison

Table 2: Geo-localization performance comparison on *MP16-Reason*-Test and IM2GPS3K [82].



| Method | Dataset, Size | *MP16-Reason*-Test (% @ km) | | | | | IM2GPS3K [82] (% @ km) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Street 1km | City 25km | Region 200km | Country 750km | Continent 2500km | Street 1km | City 25km | Region 200km | Country 750km | Continent 2500km |
| **I. Image-only supervision** | | | | | | | | | | | |
| ISNs [9] | MP-16, 4M | 26.24 | 47.38 | 55.88 | 68.48 | 80.92 | 10.50 | 28.00 | 36.60 | 49.70 | 66.00 |
| GeoCLIP [15] | MP-16, 4M | 29.28 | 52.52 | 66.85 | 84.07 | 93.33 | 14.11 | 34.47 | 50.65 | 69.67 | 83.82 |
| Translocator† [10] | MP-16, 4M | - | - | - | - | - | 11.80 | 31.10 | 46.70 | 58.90 | 80.10 |
| PIGEOTTO† [16] | MP-16, 4M | - | - | - | - | - | 11.30 | 36.70 | 53.80 | 72.40 | **85.30** |
| G3 (GPT4V)† [19] | MP-16, 4M | - | - | - | - | - | **16.65** | 40.94 | 55.56 | 71.24 | 84.68 |
| Hybrid [83] | OSV-5M, 5M | 0.97 | 16.53 | 28.72 | 50.31 | 71.47 | 0.83 | 13.28 | 25.33 | 43.84 | 65.63 |
| RFM-YFCC [49] | Flickr, 48M | 11.72 | 46.64 | 60.46 | 77.97 | 91.96 | 5.41 | 29.70 | 44.71 | 61.83 | 79.55 |
| **II. Open- and closed-source LVLMs** | | | | | | | | | | | |
| Qwen2.5-VL-7B [24] | - | 15.42 | 52.72 | 62.86 | 75.11 | 83.47 | 8.58 | 32.53 | 43.11 | 58.93 | 72.37 |
| InternVL3-8B [33] | - | 12.01 | 44.17 | 55.66 | 75.36 | 86.98 | 6.44 | 25.69 | 34.57 | 49.38 | 61.66 |
| Gemma3-27B [84] | - | 16.03 | 55.63 | 68.07 | 82.59 | 91.29 | 8.48 | 33.37 | 46.61 | 63.63 | 79.95 |
| InternVL3-78B [33] | - | 14.72 | 52.46 | 65.25 | 81.73 | 91.17 | 8.93 | 35.05 | 47.32 | 64.03 | 78.64 |
| Qwen2.5-VL-72B [24] | - | 17.52 | 59.30 | 71.01 | 84.06 | 91.65 | 9.11 | 35.77 | 48.35 | 64.96 | 78.88 |
| Doubao1.5-VL† [85] | - | 18.89 | 64.02 | 76.55 | 88.33 | 93.44 | 11.61 | 46.21 | **60.60** | **75.04** | 85.09 |
| GPT-4.1† [86] | - | **20.05** | **66.76** | **79.70** | **89.84** | **94.53** | 12.11 | **46.85** | 60.36 | 74.41 | 85.25 |
| **III. Task-specific reasoning supervision** | | | | | | | | | | | |
| GeoReasoner-7B [28] | GSV, 133K | 10.06 | 40.44 | 50.91 | 68.01 | 79.68 | 7.67 | 26.94 | 36.63 | 52.27 | 65.39 |
| GaGA† [30] | MG-Geo, 5M | - | - | - | - | - | 11.70 | 33.00 | 48.00 | 67.10 | 82.10 |
| ***GLOBE*-7B (Ours)** | *MP16-Reason*, 33K | 17.99 | 62.85 | 73.83 | 86.68 | 92.52 | 9.84 | 40.18 | 56.19 | 71.45 | 82.38 |

† denotes models that are not publicly available. Underlined results indicate test–train overlap. Best open- and closed-source results are in blue and **bold**, respectively.

Fine-tuned on only 33K reasoning-augmented data, our model outperforms mainstream open-source models trained on millions.

# Experiments

## Experimental Results

### Ablation Study - Reward components

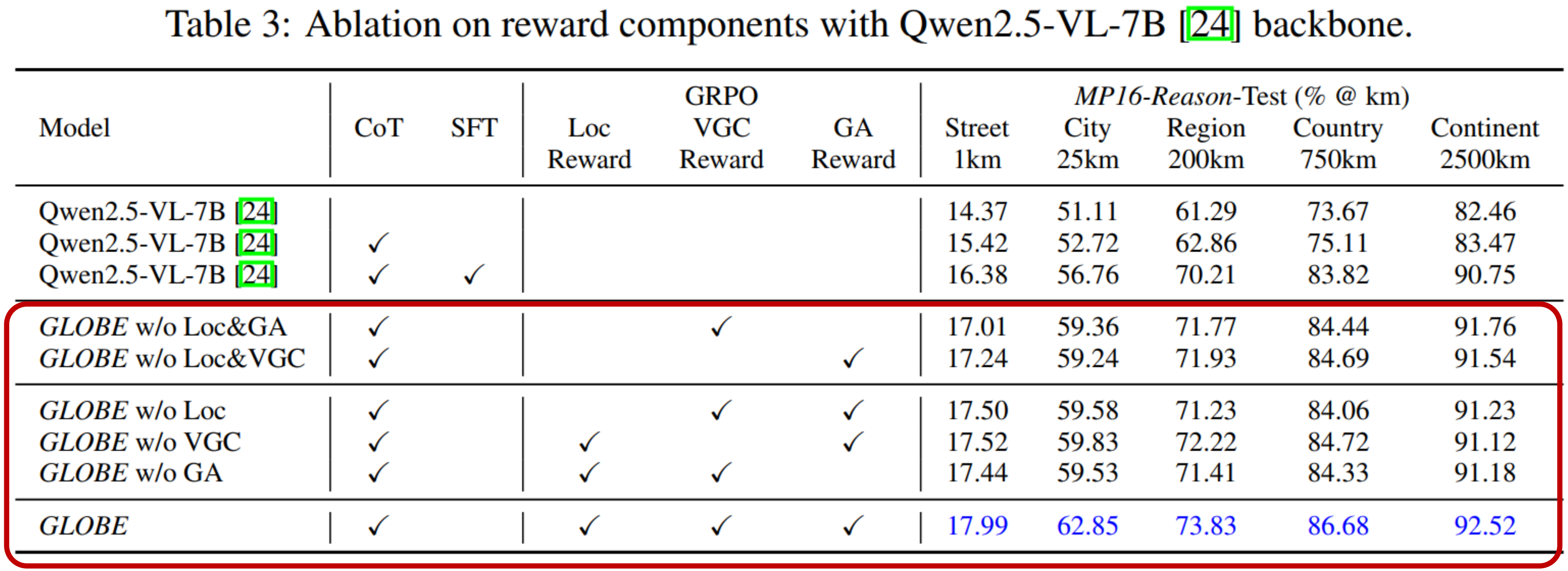


Table 3: Ablation on reward components with Qwen2.5-VL-7B [24] backbone.

| Model | CoT | SFT | GRPO Loc Reward | GRPO VGC Reward | GRPO GA Reward | *MP16-Reason*-Test (% @ km) Street 1km | City 25km | Region 200km | Country 750km | Continent 2500km |
|---|---|---|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-7B [24] | | | | | | 14.37 | 51.11 | 61.29 | 73.67 | 82.46 |
| Qwen2.5-VL-7B [24] | ✓ | | | | | 15.42 | 52.72 | 62.86 | 75.11 | 83.47 |
| Qwen2.5-VL-7B [24] | ✓ | ✓ | | | | 16.38 | 56.76 | 70.21 | 83.82 | 90.75 |
| *GLOBE* w/o Loc&GA | ✓ | | | ✓ | | 17.01 | 59.36 | 71.77 | 84.44 | 91.76 |
| *GLOBE* w/o Loc&VGC | ✓ | | | | ✓ | 17.24 | 59.24 | 71.93 | 84.69 | 91.54 |
| *GLOBE* w/o Loc | ✓ | | | ✓ | ✓ | 17.50 | 59.58 | 71.23 | 84.06 | 91.23 |
| *GLOBE* w/o VGC | ✓ | | ✓ | | ✓ | 17.52 | 59.83 | 72.22 | 84.72 | 91.12 |
| *GLOBE* w/o GA | ✓ | | ✓ | ✓ | | 17.44 | 59.53 | 71.41 | 84.33 | 91.18 |
| *GLOBE* | ✓ | | ✓ | ✓ | ✓ | 17.99 | 62.85 | 73.83 | 86.68 | 92.52 |

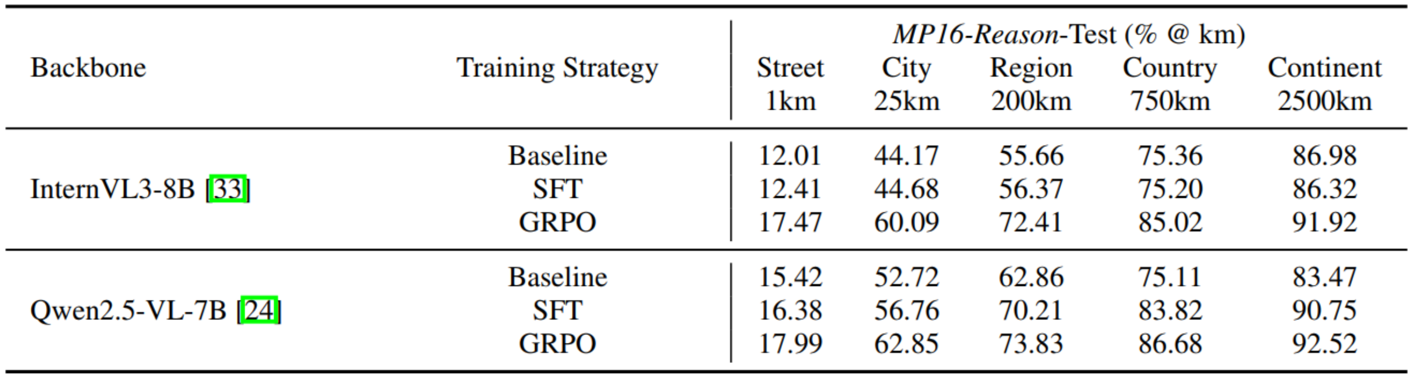
Even with partial reward combinations, GRPO still surpasses SFT, demonstrating the clear advantage of reinforcement learning with reasoning-driven supervision.

# Experiments

## Experimental Results

### Ablation Study - Backbone models

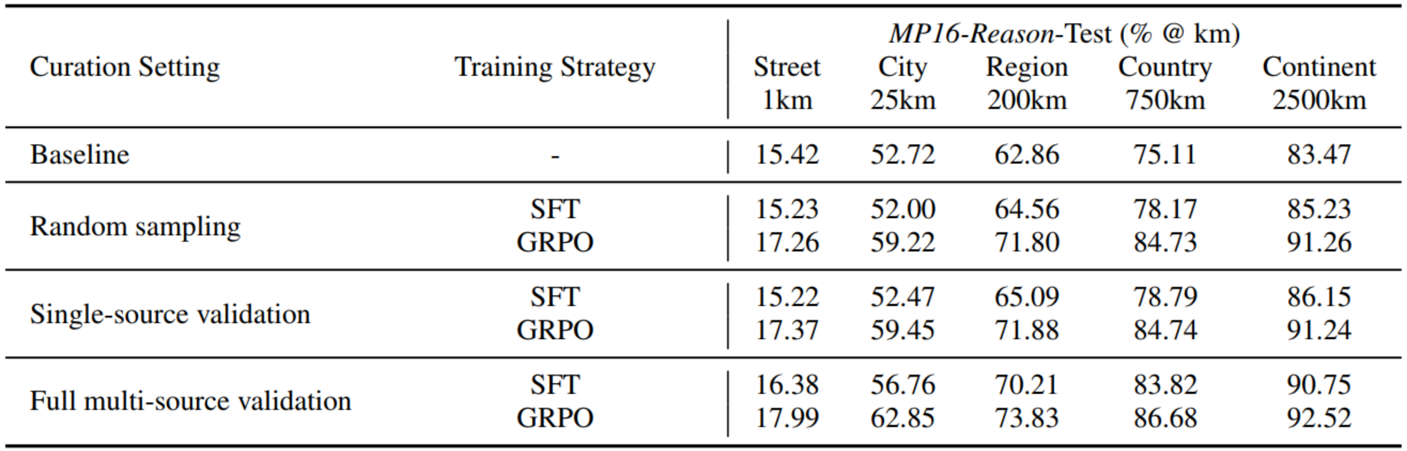
Table 4: Ablation on backbone architectures.



| Backbone | Training Strategy | *MP16-Reason*-Test (% @ km) | | | | |
|---|---|---|---|---|---|---|
| | | Street 1km | City 25km | Region 200km | Country 750km | Continent 2500km |
| InternVL3-8B [33] | Baseline | 12.01 | 44.17 | 55.66 | 75.36 | 86.98 |
| | SFT | 12.41 | 44.68 | 56.37 | 75.20 | 86.32 |
| | GRPO | 17.47 | 60.09 | 72.41 | 85.02 | 91.92 |
| Qwen2.5-VL-7B [24] | Baseline | 15.42 | 52.72 | 62.86 | 75.11 | 83.47 |
| | SFT | 16.38 | 56.76 | 70.21 | 83.82 | 90.75 |
| | GRPO | 17.99 | 62.85 | 73.83 | 86.68 | 92.52 |

### Ablation Study - Distillation datasets

Table 5: Ablation on data curation with Qwen2.5-VL-7B [24] backbone.



| Curation Setting | Training Strategy | *MP16-Reason*-Test (% @ km) | | | | |
|---|---|---|---|---|---|---|
| | | Street 1km | City 25km | Region 200km | Country 750km | Continent 2500km |
| Baseline | - | 15.42 | 52.72 | 62.86 | 75.11 | 83.47 |
| Random sampling | SFT | 15.23 | 52.00 | 64.56 | 78.17 | 85.23 |
| | GRPO | 17.26 | 59.22 | 71.80 | 84.73 | 91.26 |
| Single-source validation | SFT | 15.22 | 52.47 | 65.09 | 78.79 | 86.15 |
| | GRPO | 17.37 | 59.45 | 71.88 | 84.74 | 91.24 |
| Full multi-source validation | SFT | 16.38 | 56.76 | 70.21 | 83.82 | 90.75 |
| | GRPO | 17.99 | 62.85 | 73.83 | 86.68 | 92.52 |

NEURAL INFORMATION
PROCESSING SYSTEMS

# THANKS!

## Recognition through Reasoning: Reinforcing Image Geo-localization with Large Vision-Language Models

---

**Full Paper**

**Code & Datasets**

**Personal Website**

香港科技大學（廣州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY