



GRE Suite

GRE Suite: Geo-localization Inference via Fine-Tuned Vision-Language Models and Enhanced Reasoning Chains

arxiv 2501.00599 Dataset Hugging Face Model Hugging Face Benchmark Hugging Face

Release Process

- Release the checkpoint & code of GRE
- Release the [GRE30K dataset](#) on HuggingFace
- Release the [GREval-Bench](#) on HuggingFace
- Release the Evaluation outputs

🔍 About GRE Suite

GRE Suite is designed to augment VLMs with structured reasoning chains for accurate and interpretable location inference. It consists of three primary components:

- **Dataset (GRE30K)**

GRE30K is a geo-localization reasoning dataset designed to enhance the visual reasoning capability of MLLMs. Specifically, GRE30K consists of **GRE30K-CoT** for cold-start Initialization and **GRE30K-Judge** for reinforcement learning.

GRE30K-CoT



GT: (-31.711°, 148.663°)

<think> 1. Street-level analysis: Two-storey corner hotel with full-length veranda, corrugated-iron roof and brick chimney is a classic late-19th- / early- 20th-century “country pub” style seen throughout rural Australia. Vehicles have right-hand drive and narrow number plates consistent with Australian states. Wooden power poles with cross-arms and exposed phone wires are typical of small-town New South Wales/Queensland. // 2. Textual clues: English signage (“Bistro”) and pub layout suggest an Aussie hotel rather than UK/US. No bilingual or metric/imperial mix points away from North America. // 3. Vegetation/biome: Open grassy verge, sparse street trees that appear to be eucalypts; bright dry light and cloud-free sky indicate an inland, temperate-semi-arid Australian setting (western slopes & plains). // 4. Terrain: Flat town grid with wide asphalt roadway—characteristic of NSW western plains settlements sited on gently undulating country. // 5. Transport context: Mix of utes and sedans parked at an old pub—all common in regional Australian service centres. </think>

<answer> “Latitude”: “-31.28°”, “Longitude”: “149.28°” </answer>

GRE30K-Judge



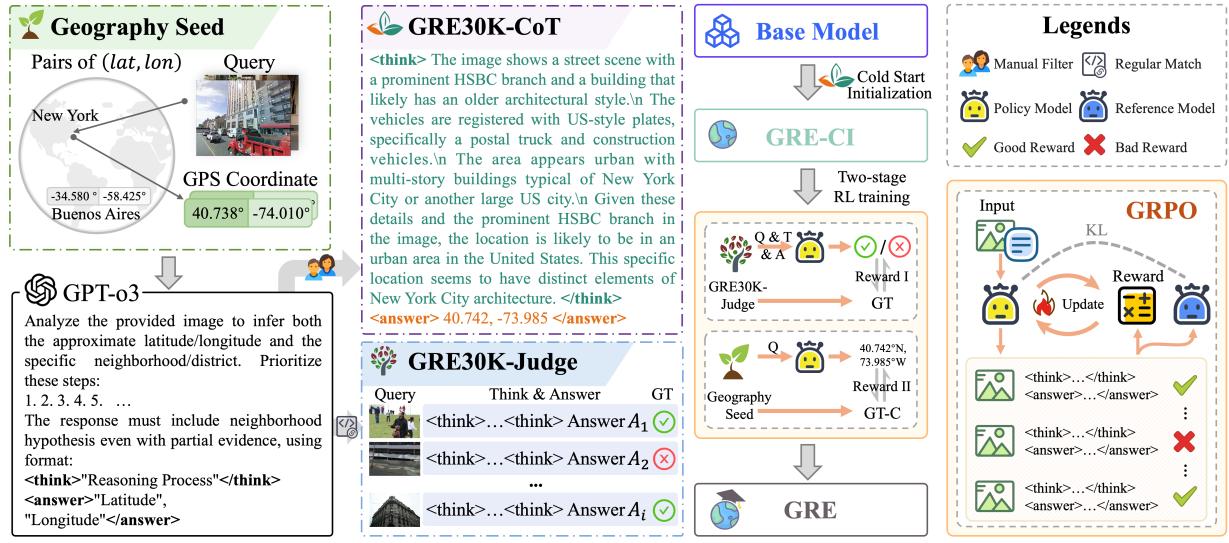
GT Coordinate: (40.724°, -73.901°)

o3 Answer: (33.773°, -118.121°)

Judgment GT: False

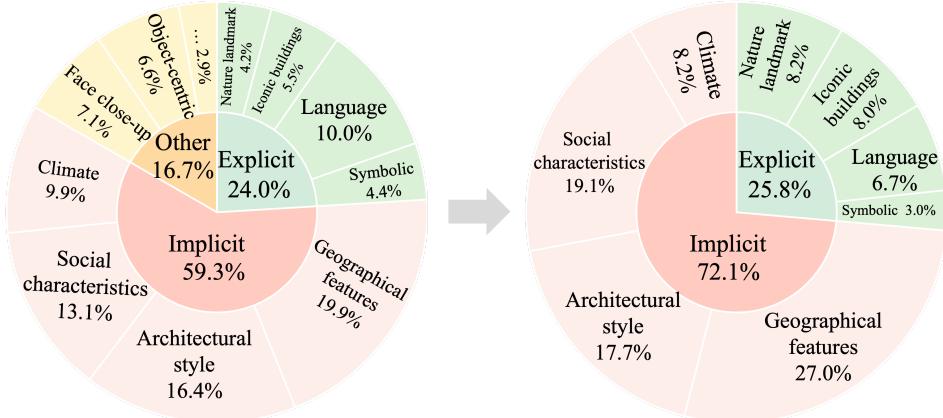
- **Model (GRE)**

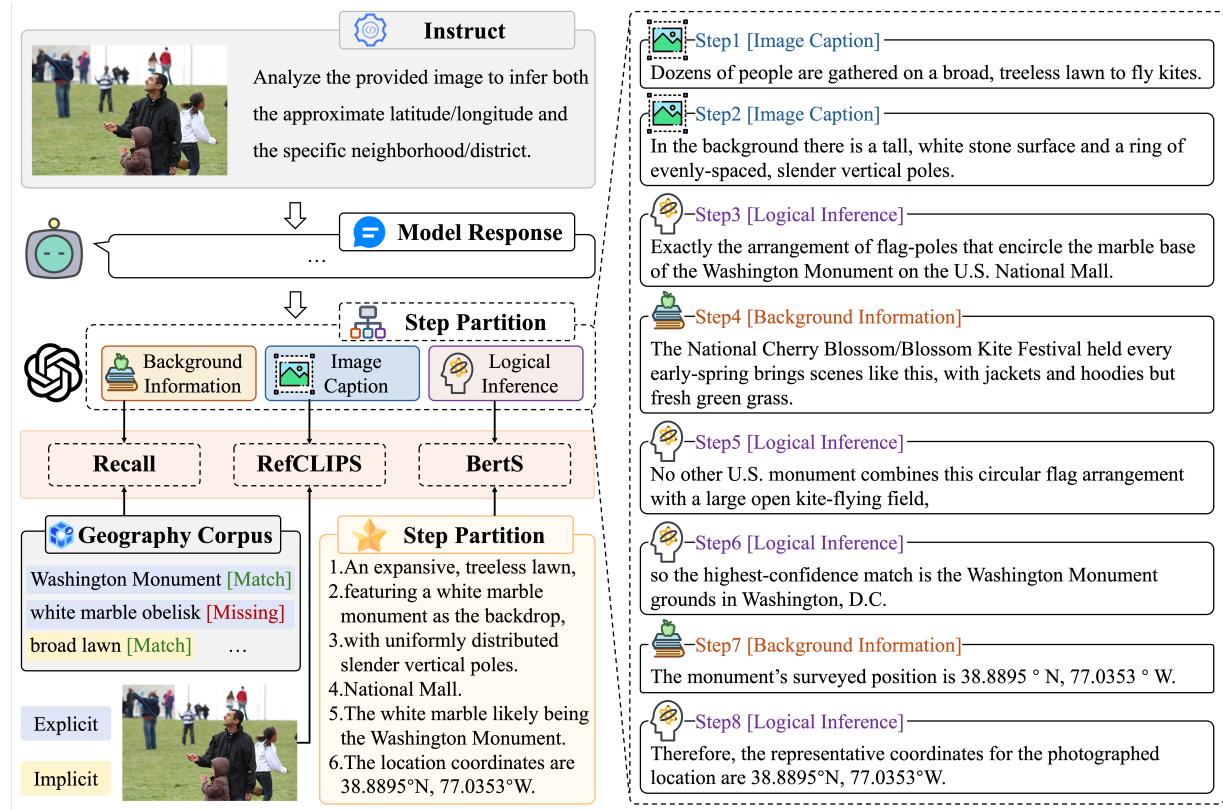
GRE is an effective Reasoning MLLM, which employs a multi-stage reasoning strategy to progressively infer scene attributes, local details, and semantic features, thereby narrowing down potential geographic regions with enhanced precision.



• Benchmark (GREval-Bench)

GREval-Bench is a geographical reasoning benchmark that employs a semi-automated pipeline to curate geographically informative images with explicit and implicit indicators, and provides annotated Chain-of-Thought steps and reference GPS coordinates for comprehensive evaluation of models' geo-localization capabilities.





🛠 Requirements and Installation

Basic Dependencies:

- Python >= 3.8
- Pytorch >= 2.5.0
- CUDA Version >= 11.8
- transformers == 4.40.0
- tokenizers == 0.19.1

```
git clone https://github.com/Thorin215/GRE.git
cd GRE
conda create -n GRE python=3.10
conda activate GRE
bash environment.sh
```

⭐ Getting started

Step1: download all the checkpoints:

The default path of all the checkpoints:

```
└── demo
    └── checkpoints
        ├── Osprey_7b
        │   └── sam_vit_b_01ec64.pth
        └── open_clip_pytorch_model.bin
```

Step2: prepare images in `input` folder.

Step3: inference for image geo-localization and the image name and the corresponding predicted coordinate will be output to the terminal.

```
python infer.py
```

🚀 Main Results

We perform a comparative analysis of GRE against worldwide Geo-Localization benchmarks, Im2GPS3k and GWS15k. In all metrics, our method surpasses the previous **state-of-the-art** model on Im2GPS3k, achieving improvements of +0.5%, +4.2%, +3.0%, +1.7% and +2.5% in the 1km, 25km, 200km, 750km, and 2500km thresholds respectively.

Table 2: We compare the performance of GRE with the state-of-the-art methods on (a) Im2GPS3k [9] and (b) GWS15k [6] datasets. Our method yields consistent gains across datasets and different distance thresholds. † denotes transformer-based models.

(a) Results on the Im2GPS3k [9] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
[L]kNN, $\sigma = 4$ [46]	7.2	19.4	26.9	38.9	55.9
PlaNet [48]	8.5	24.8	34.3	48.4	64.6
CPlaNet [33]	10.2	26.5	34.6	48.6	64.6
ISNs [24]	3.2	9.6	14.3	25.1	43.9
Translocator [†] [29]	7.6	20.3	27.1	40.7	63.3
GeoDecoder [†] [6]	5.7	10.3	21.4	28.9	38.6
GeoCLIP [†] [45]	10.8	31.1	48.7	67.6	83.2
Ours	11.3	35.3	51.7	69.3	85.7

(b) Results on the recent GWS15k [6] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
ISNs [24]	0.05	0.6	4.2	15.5	38.5
Translocator [†] [29]	0.5	1.1	8.0	25.5	48.3
GeoDecoder [†] [6]	0.7	1.5	8.7	26.9	50.5
GeoCLIP [†] [45]	0.6	3.1	16.9	45.7	74.1
Ours	0.9	4.1	18.9	54.8	78.3

We compare our approach on GREval-Bench with the previous generalist models, including InternVL2.5 series, InternVL3 series, Qwen2.5-VL series. We conduct comprehensive evaluations of models, analyzing the above metric across different distance thresholds and scenarios, while also assessing the quality of its reasoning chains.

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km	CoT quality
ISNs	1.76	11.23	16.94	23.08	26.4	-
GeoCLIP	2.45	15.71	34.08	64.85	76.61	-
InternVL2.5-4B	0.05	2.74	5.09	12.08	18.96	31.22
InternVL2.5-8B	0.33	3.44	6.75	14.62	22.64	34.29
InternVL3-2B	0.19	0.75	1.56	3.82	6.18	23.41
InternVL3-8B	1.32	7.50	14.34	25.90	35.38	36.48
Qwen2.5VL-3B	0.19	0.61	2.03	3.40	5.14	37.93
Qwen2.5VL-7B	0.33	4.34	6.84	9.39	10.90	50.36
Qwen2.5VL-32B	5.45	23.12	37.41	54.33	65.00	55.56
Ours	6.14	26.15	44.67	66.56	83.16	59.54

Table 3: Performance comparisons among traditional leading models, open-source MLLMs, and our GRE on GREval-Bench.

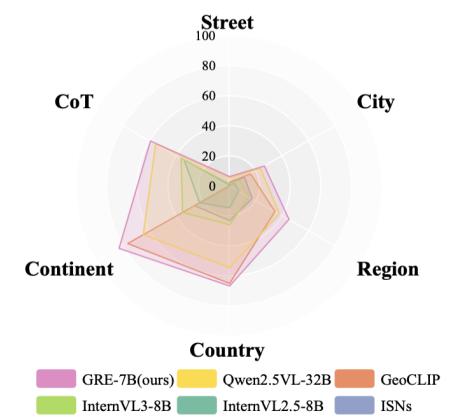


Figure 6: Visual demonstration of the performance of models.

🔑 Training & Evaluation

Training

The training data and data structure can be found in [Dataset preparation](#).

The pretrained projector weights can be found in [VideoLLaMA2.1-7B-16F-Base](#).

The training pipeline of our model is structured into three distinct stages.

- **Stage1: Cold-start Initialization**

- Prepare datasets used for cold-start initialization.
- Run `bash scripts/train/stage-CI.sh`.

- **Stage2: RL stage I**

- Prepare datasets used for stage2.
- Run `bash scripts/train/stage2.sh`.

- **Stage3: RL stage II**

- Prepare datasets used for stage3.
- Run `bash scripts/train/stage3.sh`.

Evaluation

For model evaluation, please refer to [eval](#).

Checkpoints

Model Name	Base Model	# Training Epochs
GRE-7B	Qwen2.5-VL-7B-Instruct	16
GRE-7B-CI	Qwen2.5-VL-7B-Instruct	xx
GRE-7B-stage2	Qwen2.5-VL-7B-Instruct	16

GRE30K

The dataset can be accessed on [huggingface](#).

Data format:

```
[  
  {  
    "video": "videos/xxx.mp4",  
    "conversations": [  
      {  
        "from": "human",  
        "value": "<video>\nWhat is the relationship of <region>  
and <region>?"  
      },  
      {  
        "from": "model",  
        "value": "The relationship between <region> and <region> is <relationship>.  
It is a <type> relationship."  
      }  
    ]  
  }  
]
```

```
        "from": "gpt",
        "value": "...."
    },
    ...
],
"annotation": [
    //object1
    {
        "frame_idx": {
            "segmentation": {
                //rle format or polygon
            }
        }
    },
    "frame_idx": {
        "segmentation": {
            //rle format or polygon
        }
    }
},
//object2
{
    "frame_idx": {
        "segmentation": {
            //rle format or polygon
        }
    }
},
...
]
}
```

🎮 GREval-Bench

GREval-Bench assesses the models in two key areas: localization performance and Chain-of-Thought quality.

- The annotations of the benchmark can be found in 😊 [benchmark](#).
- The usage of GREval-Bench is detailed in [doc](#).

📄 Citation

If you find GRE Suite useful for your research and applications, please cite using this BibTeX: